Volume 3, Number 2, 2026
Pages 1155-1170
Bioscience Insights
ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)
BiosciInsights.com

16S RRNA GENE SEQUENCING FOR IDENTIFICATION AND PHYLOGENETIC ANALYSIS OF ENVIRONMENTAL BACTERIAL ISOLATES

Syed Muhammad Ahmad Shah[1], Kiran Fiaz[2], Sadaf Saleem[3], Hafiz Aneaq Ahmad[4], Abida Mushtaque[5], Dr. Ayesha Imtiaz[*6], Zainab bano[7], Aliza Jahangir[8], Ayesha Tariq[9]

[1]*COMSATS University Islamabad, Abbottabad Campus*
[2]*PHD Zoology, Department of zoology, Quaid-e-Azam University Islamabad*
[3]*MLT student from Sheikha Fatima institute of Nursing & health sciences Lahore affiliated with Punjab University*
[4]*MLT student from Sheikha Fatima institute of Nursing & health sciences Lahore affiliated with Punjab University*
[5]*Centre of Excellence in Molecular biology Punjab university, Lahore*
[*6]*PHD, Institute of Zoology, Bahauddin Zakariya University, Multan*
[7]*Department of Zoology, Kohat University of Science and Technology, Kohat Pakistan*
[8]*Department of Botany, University of Education, Lahore Vehari Campus, Pakistan*
[9]*Department of Microbiology, Cholistan University of Veterinary and Animal Sciences Bahawalpur*

[1]muhammadahmad1998.ma@gmail.com, [2]kiranfiaz9966@gmail.com, [3]sadafsaleemm@gmail.com, [4]aneaqahmad65@gmail.com, [5]abidamushtaqu@yahoo.com, [*6]drayeshaimtiaz@bzu.edu.pk, [7]zainabbano.bangash@gmail.com, [8]alizajahangir030@gmail.com , [9]ayeshatariqbwp@gmail.com

https://doi.org/10.5281/zenodo.23101311

Article Information

Article History

Received:	May 15, 2026
Revised:	June 27, 2026
Accepted:	August 12, 2026
Available Online:	September 30, 2026

Keywords:
16S rRNA gene sequencing, Urban microbiome, Environmental bacteria, Bacterial taxonomy, Phylogenetic analysis, Microbial diversity.

Abstract

Background
Urban environments harbor diverse bacterial communities across water, soil, and air-deposition niches, yet systematic phylogenetic surveys from South Asian megacities remain scarce. Whether niche type predicts taxonomic composition and phylogenetic clustering in Lahore's cultivable aerobic bacteria remains unclear. This study addresses this knowledge gap.

Research Design and Methods
This cross-sectional, culture-based environmental survey enrolled 300 cultivable aerobic bacterial isolates at 15 geographically distributed sites in Lahore, Pakistan, from defined sampling dates. Sites comprised five surface-water, five soil, and five air-deposition locations representing residential, commercial, industrial, peri-urban, and green-space zones. Isolates underwent 16S rRNA gene sequencing (≥400 bp, ≥90% query coverage), MALDI-TOF MS,

and VITEK 2 identification. Primary outcome was phylum-level taxonomic composition by niche. Phylogenetic trees were constructed using maximum-likelihood methods. Alpha-diversity and beta-diversity were assessed at site level.

Results

Of 300 isolates sequenced, 104 were from water, 97 from soil, and 99 from air-deposition. Mean 16S identity was 98.90%±0.80%. Phylum distribution differed significantly by niche (χ^2=156.77, df=6, p<0.001): Pseudomonadota dominated water (76.0%), while Actinomycetota and Bacillota prevailed in soil (41.2%, 38.1%) and air-deposition (46.5%, 40.4%). Eighteen genera were recovered; several were niche-restricted. Alpha-diversity did not differ between niches (p=0.113–0.357). No adverse events applicable.

Conclusions

Among Lahore's urban isolates, niche type was strongly associated with taxonomic composition, with distinct phylum-level distributions across water, soil, and air-deposition surfaces. However, culture-based methods limit generalizability to total community diversity, and site-level sample size (n=15) constrains beta-diversity precision. These findings support niche-specific environmental monitoring. Further metagenomic research is recommended.

INTRODUCTION

Bacteria are the most abundant and phylogenetically diverse organisms on Earth, colonising virtually every environmental compartment from deep ocean sediments to urban surfaces [1]. Estimates suggest that the total number of prokaryotic cells on the planet exceeds 10^30, with the vast majority residing in soil, freshwater, and the overlying atmosphere [2]. Urban environments, despite their comparatively small spatial footprint, represent ecologically complex and heterogeneous niches in which natural and anthropogenic pressures converge to shape microbial community structure in ways that remain incompletely characterised [3]. Water bodies, surface soils, and air-deposition surfaces within cities are subject to fluctuating physicochemical gradients, variable organic loading, and continuous microbial immigration driven by wind dispersal, surface run-off, and human activity, all of which are expected to influence the taxonomic and phylogenetic composition of resident cultivable communities [4].

The 16S ribosomal RNA (rRNA) gene has been the cornerstone of bacterial identification and phylogenetic inference since its utility was first demonstrated by Woese and Fox in the late 1970s [5]. Comprising both highly conserved and hypervariable regions, the gene permits simultaneous determination of phylogenetic affiliation and, with sufficient sequence resolution, genus- and species-level taxonomic assignment [6]. Culture-based surveys that combine selective and non-selective cultivation with Sanger sequencing of 16S rRNA gene amplicons continue to provide tractable and reproducible snapshots of the cultivable aerobic fraction of environmental communities, offering a complementary perspective to culture-independent metagenomic approaches [7]. Phylogenetic tree construction using algorithms such as the neighbour-joining method, maximum parsimony, and maximum likelihood allows investigators to visualise evolutionary relationships among isolates and to test whether community members cluster according to environmental origin rather than phylogenetic chance [8].

South Asia, and Pakistan in particular, presents an underexplored context for urban environmental microbiology. Rapid urbanisation across Pakistani cities has placed severe strain on water infrastructure and waste management systems, increasing the likelihood that pathogenic and opportunistic taxa intermingle with saprophytic environmental bacteria across multiple niches [9]. Lahore, the second-largest city in Pakistan with a population exceeding thirteen million, encompasses a diverse array of urban environmental niches—including river and canal water bodies, peri-urban agricultural soils, and public surfaces subject to aerial deposition—each representing a distinct selective milieu for bacterial communities [9]. Studies from comparable South Asian megacities have documented high loads of both

pathogenic genera such as Escherichia, Klebsiella, and Pseudomonas and environmentally ubiquitous taxa such as Bacillus and Micrococcus in surface waters and soils, yet systematic phylogenetic surveys that simultaneously sample multiple niche types within a single city remain scarce [10]. The Ravi River, which traverses Lahore, receives substantial volumes of untreated municipal and industrial effluent, creating conditions that favour resistant and metabolically versatile taxa whose phylogenetic relationships to globally distributed lineages have not been systematically characterised [11]. Similarly, the soils and open surfaces of a densely populated city are exposed to atmospheric deposition carrying bacterial cells from heterogeneous sources, yet the extent to which air-deposition communities overlap phylogenetically with those in water and soil in this setting is unknown [12].

This gap in the evidence is consequential. Without a baseline characterisation of which phylogenetic lineages inhabit these niches and whether niche type predicts clustering in reconstructed trees, it is impossible to identify niche-specific transmission risks, to detect the emergence of unusual taxa, or to prioritise environmental monitoring targets. Culture-independent surveys, while powerful, are resource-intensive and require bioinformatic infrastructure that remains limited in many Pakistani research settings; a rigorous culture-based survey with 16S rRNA gene sequencing therefore offers a practical and reproducible first step toward filling this knowledge gap [13]. International precedents demonstrate that even modest culture-based surveys, when combined with neighbour-joining or maximum-likelihood phylogenetic reconstruction, can reveal meaningful niche-driven clustering that informs both ecological theory and public health surveillance [14].

The present study was therefore designed to determine the taxonomic composition and phylogenetic structure of cultivable aerobic bacterial communities recovered from urban water, soil, and air-deposition surfaces at geographically distributed sites across Lahore, Pakistan, using 16S rRNA gene sequencing, and to assess whether phylogenetic lineages cluster predictably by environmental niche type as resolved from sequence data alone.

METHODOLOGY

Study Design

A cross-sectional, culture-based environmental survey was conducted to characterize the cultivable aerobic bacterial isolates recovered from three environmental niches in Lahore, Pakistan: surface water, soil, and air-deposition surfaces. The study was designed to compare the taxonomic composition, sequence-based identification, and phylogenetic structure of isolates obtained from these niches.

The sampling unit for environmental comparison was the sampling site, whereas individual bacterial isolates represented subsamples within each site. This distinction was maintained during statistical analysis to avoid treating multiple isolates from the same site as independent observations.

Study Setting and Sampling Frame

The study was conducted at 15 geographically distributed sites in Lahore, Punjab, Pakistan. The sampling frame comprised five surface-water sites, five soil sites, and five air-deposition sites. The sites represented different urban land-use zones, including residential, commercial, industrial, peri-urban, transitional, and green-space areas.

Water, soil, and air-deposition sites were treated as separate niche-specific sampling locations. A water site was not simultaneously classified as a soil or air-deposition site. Overall, 300 cultivable bacterial isolates were included in the final dataset: 104 from water, 97 from soil, and 99 from air-deposition samples. The mean number of isolates recovered per site was approximately 20 across the three niches.

Sample Size

The planned sample size was approximately 100 isolates per environmental niche, giving a total target of 300 isolates. The initial calculation was based on a chi-square goodness-of-fit model using a medium effect size of Cohen's $w = 0.30$, a significance level of alpha = 0.05, and 80% statistical power. This calculation produced a minimum requirement of 88 isolates per group. The target was increased to approximately 100 isolates per niche to allow for sequencing failure, poor-quality amplification, exclusion of duplicate isolates, and incomplete records.

Because isolates collected from the same site are clustered, the effective sample size for community-level comparisons was 15 sites rather than 300 isolates. Consequently, beta-diversity and between-niche analyses were performed at the site level wherever applicable.

Sampling Technique

Purposive multi-site sampling was used to capture environmental variation across the urban landscape of Lahore. Five sites were selected for each environmental niche.

Surface-water samples of approximately 500 mL were collected aseptically in sterile sampling bags from designated water bodies. Soil samples of approximately 50 g were collected from the upper 0–10 cm layer using sterile spatulas and transferred into sterile sampling bags. Air-deposition samples were collected by exposing pre-poured Nutrient Agar settle plates at selected sites for 60 minutes. The air-deposition plates were then covered, labelled, and transported to the laboratory.

For each site, the sampling location, environmental niche, land-use category, date and time of collection, ambient temperature, and relevant physicochemical information were recorded on a standardized Site and Sample Collection Proforma. Samples were transported to the laboratory in an ice-cooled container and processed as soon as possible after collection.

Physicochemical Assessment

Cultivable bacterial density was determined and expressed as $\log_{10}$ colony-forming units (CFU) per millilitre for water samples and per gram for soil samples. The mean cultivable density was 5.45 ± 0.66 $\log_{10}$ CFU per millilitre in water, 6.23 ± 0.38 $\log_{10}$ CFU per gram in soil, and 3.36 ± 0.44 $\log_{10}$ CFU in air-deposition samples.

The pH of water and soil samples was measured in situ using a Hanna HI98130 pH meter with probe. The mean pH was 7.75 ± 0.12 for water and 7.66 ± 0.93 for soil. pH was not measured for air-deposition samples because settle plates do not constitute a liquid or soil matrix with a meaningful pH value. Ambient temperature was recorded at the time of collection for all three niches.

Culture and Isolation of Bacteria

Water and soil samples were homogenized or suspended in sterile phosphate-buffered saline and serially diluted. Appropriate dilutions were spread-plated in duplicate on Nutrient Agar and R2A Agar. Air-deposition plates were incubated directly after exposure.

All plates were incubated aerobically at 37 °C for 24–48 hours. Following incubation, colonies were examined for differences in morphology, pigmentation, size, elevation, and surface characteristics. Morphologically distinct colonies were selected and subcultured to obtain pure isolates. Colony counts were used to estimate cultivable bacterial density.

Each purified isolate was assigned a unique identification code indicating its environmental niche, sampling site, culture medium, and isolate number. Colony morphology and Gram-stain findings were documented in the Laboratory Isolate Tracking and Culture Record Sheet.

Eligibility of Isolates

Isolates were eligible for inclusion if they:

- Produced a discrete bacterial colony on Nutrient Agar or R2A Agar after aerobic incubation at 37 °C for 24–48 hours.
- Originated from a sample collected during the defined sampling period.
- Had complete site and sample documentation.
- Produced a readable 16S rRNA gene sequence of at least 400 base pairs.
- Demonstrated at least 90% sequence query coverage against the selected reference databases.
- Met the predefined phenotypic identification quality criteria, consisting of a VITEK 2 confidence value of at least 60% or a MALDI-TOF MS log-score of at least 1.70.

Isolates were excluded when they produced ambiguous or unreadable sequences after repeat amplification, represented fungal or yeast growth, had incomplete sample documentation, or were associated with compromised transport or storage conditions. Duplicate isolates from the same sample

were also excluded when they had identical colony morphology, the same Gram-stain profile, and at least 99% 16S rRNA sequence identity. In such cases, only one representative isolate was retained.

Phenotypic Identification

Phenotypic identification was performed using the VITEK 2 Compact automated identification system and MALDI-TOF MS. VITEK 2 confidence values and MALDI-TOF MS log-scores were recorded for each isolate and used as corroborating evidence rather than as the sole basis for final taxonomic assignment.

A MALDI-TOF MS log-score of at least 2.00 was interpreted as supporting species-level identification, a score of 1.70–1.99 as supporting genus-level identification, and a score below 1.70 as unreliable. The overall mean MALDI-TOF score was 2.01 ± 0.24. Species-level scores were obtained for 152 isolates, genus-level scores for 112 isolates, and unreliable scores for 36 isolates.

DNA Extraction and 16S rRNA Gene Sequencing

Genomic DNA was extracted from pure bacterial cultures using a validated bacterial DNA extraction procedure. The bacterial 16S rRNA gene was amplified by polymerase chain reaction using universal bacterial primers. Amplified products were assessed for successful amplification and suitable product quality before purification and Sanger sequencing.

Sequence chromatograms were inspected for quality, and low-quality terminal regions were removed. Only sequences containing at least 400 readable base pairs and at least 90% query coverage were retained for analysis. Isolates that failed to meet the quality criteria were re-amplified and resequenced where possible. Those that continued to produce inadequate sequences were excluded from the final dataset.

The final dataset consisted of 300 isolates. The mean 16S rRNA sequence identity was 98.90 ± 0.80%, with comparable values for water, soil, and air-deposition isolates. The mean query coverage was 97.6 ± 2.1%.

Taxonomic Assignment

The processed 16S rRNA sequences were compared with reference sequences in NCBI GenBank and the SILVA database. Taxonomic assignments were based on the highest-quality matching sequence, sequence identity, query coverage, alignment quality, and phylogenetic context.

A sequence identity of at least 95% was used to support genus-level assignment, whereas a threshold of at least 97% was used in the predefined protocol for species-level assignment. However, because 16S rRNA sequencing has limited discriminatory power for some closely related bacterial species, species names were reported as closest-match assignments rather than confirmed species identifications when appropriate. This limitation is particularly relevant to members of the *Bacillus cereus* group, *Enterobacter*, and several *Pseudomonas* lineages.

The recovered isolates belonged to four major phyla: Pseudomonadota, Bacillota, Actinomycetota, and Bacteroidota. At the genus level, 18 genera were identified. The most frequently recovered genera included *Bacillus*, *Micrococcus*, *Acinetobacter*, *Pseudomonas*, *Aeromonas*, *Streptomyces*, and *Lysinibacillus*.

Phylogenetic Analysis

Representative and high-quality 16S rRNA sequences were aligned with reference sequences using a multiple-sequence alignment procedure. The alignment was inspected and trimmed to remove poorly aligned or ambiguous regions. Phylogenetic relationships were reconstructed using the maximum-likelihood method in MEGA version 11.

The GTR+G substitution model was specified for maximum-likelihood analysis. Branch support was evaluated using bootstrap resampling, and the number of bootstrap replicates was recorded in the final analysis. Bootstrap support was interpreted as support for the internal node subtending the corresponding genus-level clade rather than as an independent characteristic of an individual isolate.

Phylogenetic trees were examined to determine whether isolates clustered according to environmental niche and whether closely related lineages were distributed across water, soil, and air-deposition environments.

Community-Level Diversity Analysis

A site-by-genus abundance matrix was constructed using the counts of isolates assigned to each genus at each sampling site. Alpha-diversity measures included genus richness, Shannon diversity, and Pielou's evenness. These indices were calculated at the site level rather than treating individual isolates as independent sites. The mean genus richness was 8.6 for water, 8.0 for soil, and 7.2 for air-deposition sites. Mean Shannon diversity values were 1.993, 1.923, and 1.752, respectively, while mean Pielou's evenness values were 0.930, 0.925, and 0.889. Beta-diversity was assessed using UniFrac distances in QIIME 2. Analyses were performed on the 15 site-level communities, consisting of five water, five soil, and five air-deposition sites. Where applicable, both weighted and unweighted UniFrac distances were considered to evaluate differences in community composition and phylogenetic relatedness between niches.

Statistical Analysis

Data were entered, cleaned, and analysed using SPSS software. Continuous variables were summarized using means and standard deviations when approximately normally distributed and medians with interquartile ranges when distributional assumptions were not met. Categorical variables were presented as frequencies and percentages. Differences in site-level physicochemical characteristics, isolate counts, richness, Shannon diversity, evenness, sequence identity, query coverage, MALDI-TOF scores, VITEK 2 confidence values, and bootstrap support among the three environmental niches were evaluated using the Kruskal–Wallis test. The reported p values were interpreted cautiously because only five sites were available per niche.

Associations between broad taxonomic categories and environmental niche were assessed using Pearson's chi-square test. Because isolates from the same site are not statistically independent, isolate-level chi-square results were treated as descriptive or exploratory. Inferential comparisons of community structure were based primarily on site-level analyses. Differences in beta-diversity were evaluated using PERMANOVA on UniFrac distance matrices. Environmental niche was used as the primary explanatory variable, and site was treated as the independent community unit. Dispersion among groups was assessed where appropriate to distinguish differences in centroid location from differences in within-group variability. Statistical tests were two-sided, and a p value below 0.05 was considered statistically significant.

Quality Assurance

Aseptic procedures were used during sample collection, culture, subculture, DNA extraction, and molecular processing. Culture media, instruments, and reagents were checked according to laboratory quality-control procedures. Duplicate plating was performed for water and soil dilutions, and pure cultures were confirmed before molecular testing. Sequence quality was monitored using readable sequence length, query coverage, chromatogram quality, and database alignment quality. Phenotypic identification results were compared with 16S rRNA sequencing results. Discrepant or low-confidence identifications were reviewed and reported at the highest taxonomic level supported by the available evidence.

Ethical and Biosafety Considerations

The study involved environmental samples and did not involve human participants, human-derived specimens, or animal experimentation. Institutional permission for field sampling and laboratory work was obtained according to applicable local requirements. All environmental bacterial isolates were handled using appropriate microbiological biosafety procedures, and potentially pathogenic or opportunistic isolates were processed by trained laboratory personnel.

RESULTS

Study Population and Flow

Between the defined sampling dates, 20 potential sampling sites across Lahore were initially assessed for eligibility. Of these, five sites were excluded because the original site-level data assigned multiple, mutually incompatible land-use categories and niche types to single geographic locations, which is inconsistent with

the sampling protocol. After resolution of these inconsistencies, 15 niche-specific sites were retained: five surface-water sites, five soil sites, and five air-deposition sites (Table 1).

From these 15 sites, 300 cultivable aerobic bacterial isolates were obtained and sequenced: 104 from water, 97 from soil, and 99 from air-deposition surfaces. The initial sampling and screening yielded more than the minimum required number of isolates; after application of inclusion and exclusion criteria (readable 16S rRNA sequence ≥400 bp, query coverage ≥90%, and acceptable phenotypic identification scores), the final analytic sample comprised exactly 300 isolates, meeting the pre-specified target of approximately 100 isolates per niche. No isolates were lost to follow-up, as this was a cross-sectional environmental survey without longitudinal follow-up.

Baseline Characteristics of Sampling Sites

Site-level characteristics are presented in Table 2. The mean number of isolates sequenced per site was 20.8 ± 1.3 (median 21.0, IQR 20.0–22.0) for water, 19.4 ± 1.3 (median 20.0, IQR 18.0–20.0) for soil, and 19.8 ± 1.3 (median 20.0, IQR 19.0–21.0) for air-deposition sites ($p = 0.262$). Cultivable bacterial density differed significantly between niches, with mean $\log_{10}$ CFU/mL or CFU/g values of 5.45 ± 0.66 (median 5.39, IQR 5.11–5.57) for water, 6.23 ± 0.38 (median 6.40, IQR 5.91–6.52) for soil, and 3.36 ± 0.44 (median 3.23, IQR 3.03–3.65) for air-deposition ($p = 0.004$).

In-situ pH was measured for water and soil only. Mean pH was 7.75 ± 0.12 (median 7.82, IQR 7.68–7.82) for water and 7.66 ± 0.93 (median 7.36, IQR 7.15–8.31) for soil ($p = 0.600$). Ambient temperature at collection was 32.6 ± 1.9 °C (median 33.3, IQR 33.3–33.3) for water, 29.4 ± 2.6 °C (median 28.6, IQR 28.0–28.7) for soil, and 29.7 ± 3.7 °C (median 29.3, IQR 28.0–30.6) for air-deposition sites ($p = 0.254$).

Alpha-diversity metrics at the site level showed no statistically significant differences between niches. Mean genus richness per site was 8.6 ± 1.1 (median 9.0, IQR 8.0–9.0) for water, 8.0 ± 0.7 (median 8.0, IQR 8.0–8.0) for soil, and 7.2 ± 0.8 (median 7.0, IQR 7.0–8.0) for air-deposition ($p = 0.113$). Mean Shannon diversity (H') was 1.993 ± 0.122 (median 2.048, IQR 1.917–2.051) for water, 1.923 ± 0.169 (median 1.943, IQR 1.923–1.963) for soil, and 1.752 ± 0.200 (median 1.646, IQR 1.593–1.937) for air-deposition ($p = 0.174$). Mean Pielou evenness (J') was 0.930 ± 0.006 (median 0.932, IQR 0.924–0.933) for water, 0.925 ± 0.044 (median 0.935, IQR 0.925–0.944) for soil, and 0.889 ± 0.059 (median 0.885, IQR 0.846–0.932) for air-deposition ($p = 0.357$).

Primary Outcome: Taxonomic Composition at Phylum and Class Level

The primary outcome was the taxonomic composition of the 300 cultivable aerobic isolates at phylum and class level, overall and by environmental niche (Table 3). Four phyla were identified: Pseudomonadota (107 isolates, 35.7%), Actinomycetota (90 isolates, 30.0%), Bacillota (82 isolates, 27.3%), and Bacteroidota (21 isolates, 7.0%). The overall association between phylum distribution and environmental niche was statistically significant (Pearson $\chi^2 = 156.77$, $df = 6$, $p < 0.001$).

Pseudomonadota predominated in the water niche (79 of 104 isolates, 76.0%) but were less common in soil (15 of 97, 15.5%) and air-deposition (13 of 99, 13.1%) ($p < 0.001$). Actinomycetota were rare in water (4 of 104, 3.8%) but dominant in soil (40 of 97, 41.2%) and air-deposition (46 of 99, 46.5%) ($p < 0.001$). Bacillota showed a similar pattern, with low representation in water (5 of 104, 4.8%) and higher representation in soil (37 of 97, 38.1%) and air-deposition (40 of 99, 40.4%) ($p < 0.001$). Bacteroidota were more common in water (16 of 104, 15.4%) than in soil (5 of 97, 5.2%) and were absent from air-deposition (0 of 99) ($p < 0.001$). All phylum-level comparisons were based on $n = 300$ isolates.

Secondary Outcomes: Genus-Level Composition

Genus-level composition by environmental niche is presented in Table 4. A total of 18 genera were recovered across all niches. The most frequently isolated genera overall were *Bacillus* (43 isolates, 14.3%), *Micrococcus* (39 isolates, 13.0%), *Acinetobacter* (30 isolates, 10.0%), *Pseudomonas* (27 isolates, 9.0%), and *Streptomyces* (17 isolates, 5.7%). Several genera were effectively restricted to specific niches. *Aeromonas* (17 isolates), *Escherichia* (8 isolates), *Klebsiella* (13 isolates), *Enterobacter* (7 isolates), and *Chryseobacterium* (7 isolates) were recovered almost exclusively from water. *Streptomyces* (17 isolates), *Arthrobacter* (15 isolates), and *Paenibacillus* (8 isolates) were recovered almost exclusively from soil. *Kocuria* (9 isolates), *Staphylococcus*

(13 isolates), and *Corynebacterium* (10 isolates) were recovered almost exclusively from air-deposition surfaces. Formal χ^2 tests were not performed at the genus level because most genera had expected counts below 5 in at least one niche; counts and percentages are therefore reported descriptively (n = 300).

Species-Level Assignments and Sequence Identity

Closest-match species-level assignments from 16S rRNA gene sequencing are presented in Table 5. A total of 28 distinct closest-match taxa were identified. The most frequently assigned taxa included *Bacillus licheniformis* (20 isolates, mean 16S identity 98.74%), *Micrococcus yunnanensis* (21 isolates, mean identity 98.85%), *Bacillus cereus* (16 isolates, mean identity 99.07%), *Acinetobacter baumannii* (18 isolates, mean identity 99.07%), and *Lysinibacillus fusiformis* (18 isolates, mean identity 98.54%).

Several taxa were predominantly or exclusively associated with a single niche. *Aeromonas caviae* (11 isolates), *Aeromonas hydrophila* (6 isolates), *Escherichia coli* (8 isolates), *Klebsiella pneumoniae* (13 isolates), *Enterobacter cloacae* (7 isolates), and *Chryseobacterium indologenes* (7 isolates) were found only in water. *Arthrobacter globiformis* (15 isolates), *Paenibacillus polymyxa* (8 isolates), *Streptomyces albidoflavus* (11 isolates), and *Streptomyces griseus* (6 isolates) were found only in soil. *Corynebacterium glutamicum* (10 isolates), *Kocuria rhizophila* (5 isolates), *Kocuria palustris* (4 isolates), *Staphylococcus epidermidis* (8 isolates), and *Staphylococcus hominis* (5 isolates) were found only in air-deposition. Mean 16S sequence identity across all species-level assignments ranged from 98.51% to 99.50%.

Sequence Quality, Phenotypic Identification, and Phylogenetic Support

Sequence quality, phenotypic identification scores, and phylogenetic support measures are presented in Table 6. The overall mean 16S rRNA BLAST identity was 98.90 ± 0.80% (n = 300), with no significant difference between niches (water: 98.97 ± 0.76%, soil: 98.92 ± 0.81%, air-deposition: 98.81 ± 0.82%; p = 0.271). Mean query coverage was 97.6 ± 2.1% overall, with similar values across niches (p = 0.982).

Phenotypic identification by MALDI-TOF MS yielded an overall mean log-score of 2.01 ± 0.24 (n = 300), with 152 isolates (50.7%) achieving species-level scores (≥2.00), 112 isolates (37.3%) achieving genus-level scores (1.70–1.99), and 36 isolates (12.0%) classified as unreliable (<1.70). Mean MALDI-TOF scores did not differ significantly between niches (water: 2.05 ± 0.24, soil: 1.99 ± 0.23, air-deposition: 1.98 ± 0.25; p = 0.073).

VITEK 2 identification confidence values averaged 83.4 ± 10.1% overall, with no significant difference between niches (p = 0.449). Maximum-likelihood bootstrap support for genus-level clades averaged 83.4 ± 10.0% overall, with a small but statistically significant difference between niches (water: 82.7 ± 10.1%, soil: 85.8 ± 9.5%, air-deposition: 81.8 ± 10.2%; p = 0.012).

Taxonomic resolution based on 16S identity showed that 297 of 300 isolates (99.0%) met the pre-specified threshold for species-level assignment (≥97% identity), while 3 isolates (1.0%) were assignable only to genus level (≥95% identity). All analyses in this section were based on n = 300 isolates with complete sequence and phenotypic data; no missing data were observed for these variables.

Community-Level Beta-Diversity and Phylogenetic Clustering

[This subsection would present the results of UniFrac beta-diversity analyses and PERMANOVA testing of niche-based clustering, referencing the relevant tables and figures. As these specific numerical results were not provided in the uploaded tables, they should be inserted here in the same format as above, with exact p-values, effect sizes, and sample sizes (n = 15 sites) once available.]

Subgroup and Sensitivity Analyses

[Pre-specified subgroup analyses (e.g., by land-use zone within each niche) and sensitivity analyses (e.g., excluding isolates with borderline sequence quality or phenotypic scores) should be reported here, with exact p-values, confidence intervals, and sample sizes. These analyses were not detailed in the uploaded tables and should be added according to the pre-specified statistical analysis plan.]

Discussion

Key Findings

This cross-sectional, culture-based survey characterised 300 cultivable aerobic isolates from 15 niche-specific sites across Lahore — five water, five soil and five air-deposition — spanning five urban land-use zones (Table 1). Site-level characteristics differed mainly in cultivable density, highest in soil (6.23 ± 0.38 $\log_{10}$ CFU/g) and lowest in air-deposition (3.36 ± 0.44 $\log_{10}$ CFU; $p = 0.004$), while isolate yield, pH and temperature were comparable and alpha-diversity did not differ (richness $p = 0.113$; Shannon $p = 0.174$; evenness $p = 0.357$) (Table 2). The primary outcome was strongly niche-dependent: *Pseudomonadota* (formerly *Proteobacteria*) comprised 76.0% of water isolates, while *Actinomycetota* (*Actinobacteria*) and *Bacillota* (*Firmicutes*) together comprised 79.3% of soil and 86.9% of air-deposition isolates ($\chi^2 = 156.77$, $df = 6$, $p < 0.001$) (Table 3) [15, 16]. Eighteen genera and 28 closest-match taxa were recovered, many effectively niche-restricted (Tables 4 and 5). Sequence and phenotypic quality were high and niche-independent apart from a small bootstrap-support difference ($p = 0.012$) (Table 6). Notably, equivalent alpha-diversity accompanied marked compositional divergence: niche reorganised membership without altering site-level richness or evenness.

Possible Mechanisms and Interpretation

The contrast is consistent with established niche physiology rather than artefact. Nutrient-loaded water columns favour fast-growing, metabolically versatile Gram-negative *Pseudomonadota*, and comparable dominance is documented along urban river gradients [17, 18]. Soil selects for filamentous *Actinomycetota* and endospore-forming *Bacillota*, whose sporulation, desiccation tolerance and capacity to degrade complex polymers confer persistence under fluctuating moisture and nutrient supply [19, 20]. The air-deposition profile, dominated by *Micrococcus*, *Kocuria*, *Staphylococcus* and *Corynebacterium*, is characteristic of desiccation- and ultraviolet-tolerant, often human-associated *Actinomycetota* and *Bacillota* that survive aerial transport and settle onto exposed surfaces [12, 21]. The absence of *Bacteroidota* from air-deposition surfaces is congruent with the low desiccation tolerance of that lineage [21]. Alternative explanations cannot be excluded: incubation at 37 °C on Nutrient and R2A agar imposes a shared selective filter that may amplify apparent niche separation, and land use is confounded with niche in this design [22]. These mechanisms are plausible and literature-supported rather than demonstrated here.

Discussion of Objectives Concerning Existing Literature

The primary objective was met and contributes a South Asian megacity baseline where systematic multi-niche phylogenetic surveys have been scarce. Our water-niche findings extend earlier Pakistani reports of enteric and opportunistic genera in Lahore's surface waters by adding sequence-based phylogenetic placement rather than phenotypic identification alone. Regionally, 16S surveys in Pakistan have addressed mangrove sediments and mine walls [23, 24] and airborne profiling in Sri Lanka reported comparable staphylococcal and micrococcal predominance [25], but none sampled water, soil and air within one urban frame.

The second objective, whether lineages cluster by niche in reconstructed trees, was only partially addressed. Bootstrap support for genus-level clades was high (83.4 ± 10.0%) (Table 6) and genus- and species-level distributions were strongly niche-partitioned (Tables 4 and 5), yet the pre-specified UniFrac and PERMANOVA analyses of the 15 site-level communities remain unreported. Phylogenetic clustering by niche therefore remains an inference from taxonomic composition rather than a quantified result, and should be presented as such [26].

Comparison with Other Studies

Our phylum-level pattern aligns with the MetaSUB atlas of 4,728 urban mass-transit samples, in which city-specific signatures were dominated by surface-associated *Actinomycetota* and *Bacillota* [27]; that survey was culture-independent and sampled built surfaces, so concordance at phylum level is the most that can be claimed. A paired comparison of air and water in a temperate urban centre likewise found minimal taxonomic overlap between compartments, consistent with our near-complete segregation of *Pseudomonadota* and *Bacteroidota* from air-deposition surfaces [12].

Conversely, culture-independent studies of urban river water report higher richness and larger *Bacteroidota* and *Actinomycetota* fractions than we observed [17, 18], a discrepancy more plausibly attributable to our cultivable aerobic restriction than to a genuine ecological difference [28]. Sediment work along a pollution gradient in Kenyan urban streams identified pollution rather than niche as the principal driver of structure [29]; our design cannot test this, since physicochemical characterisation was limited to pH, temperature and cultivable density (Table 2). These comparators sampled more community units than the 15 sites available here, so their estimates of between-niche separation are more precise.

Strengths and Limitations

The design is cross-sectional and observational, so niche is associated with, not shown to determine, composition, and single-timepoint sampling cannot address seasonality.

Only the cultivable aerobic fraction recovered on two media at 37 °C is represented; obligate anaerobes, oligotrophs, psychrotrophs and the uncultivable majority are absent by design, and richness from roughly 20 isolates per site is downward-biased [28]. Species-level assignments are closest-match BLAST results, and 16S rRNA does not resolve species within the *Bacillus cereus* group, *Enterobacter*, or many *Pseudomonas* lineages, so these names are provisional (Table 5) [30, 31]. Isolates from the same site are not independent, so the isolate-level χ^2 tests underlying Tables 3 and 4 overstate significance; with five sites per niche the Kruskal–Wallis comparisons in Table 2 have very limited power, and non-significant alpha-diversity results are not evidence of equivalence [31]. Land use is confounded with niche, and sequences were not deposited with citable accessions. Strengths include a pre-specified sample-size calculation, protocol-defined eligibility thresholds (≥400 bp, ≥90% query coverage), triangulation of 16S sequencing with MALDI-TOF MS and VITEK 2 [32], explicit separation of site- from isolate-level inference, and uniformly high sequence quality across niches (Table 6).

Clinical and Research Implications

These findings support niche-specific rather than uniform environmental monitoring.

Recovery of *Escherichia coli*, *Klebsiella pneumoniae*, *Aeromonas* species, *Enterobacter cloacae* and *Acinetobacter baumannii* almost exclusively from water (Tables 4 and 5) identifies water as the priority compartment for opportunistic-pathogen surveillance [34,35].

Because no resistance phenotypes or genotypes were measured, no inference about resistance burden or human risk is warranted; integrated One Health surveillance combining culture with resistance profiling is the next step [33]. Research priorities are specific: complete the pre-specified UniFrac and PERMANOVA analyses with more sites, since community-level power scales with sites rather than isolates [26]; add shotgun metagenomics or full-length rRNA operon sequencing to resolve taxa beyond partial 16S [34]; deposit all sequences in GenBank with citable accessions; and sample across seasons with fuller physicochemical characterisation. Clinical readiness is not implied: these are ecological and surveillance findings, not a basis for patient-care change.

Conclusion:

Across 300 cultivable aerobic isolates from 15 urban sites in Lahore, environmental niche was strongly associated with phylum-level taxonomic composition, while site-level alpha-diversity was comparable. Species assignments remain provisional and niche-based phylogenetic clustering was not quantified. Because the findings describe only the cultivable aerobic fraction and measured no resistance, they support niche-specific monitoring, with surface water the priority compartment, rather than inference about resistance burden or clinical risk. Metagenomic confirmation across more sites is needed.

REFERENCES

1. Whitman, W.B., D.C. Coleman, and W.J.J.P.o.t.n.a.o.s. Wiebe, *Prokaryotes: the unseen majority.* 1998. **95**(12): p. 6578-6583.
2. Bar-On, Y.M., R. Phillips, and R.J.P.o.t.N.A.o.S. Milo, *The biomass distribution on Earth.* 2018. **115**(25): p. 6506-6511.

3. Laforest-Lapointe, I., C. Messier, and S.W.J.M. Kembel, *Tree leaf bacterial community structure and diversity differ along a gradient of urban intensity.* 2017. **2**(6): p. 10.1128/msystems. 00087-17.
4. Kembel, S.W., et al., *Architectural design influences the diversity and structure of the built environment microbiome.* 2012. **6**(8): p. 1469-1479.
5. Woese, C.R. and G.E.J.P.o.t.N.A.o.S. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* 1977. **74**(11): p. 5088-5090.
6. Janda, J.M. and S.L.J.J.o.c.m. Abbott, *16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls.* 2007. **45**(9): p. 2761-2764.
7. Amann, R.I., W. Ludwig, and K.-H.J.M.r. Schleifer, *Phylogenetic identification and in situ detection of individual microbial cells without cultivation.* 1995. **59**(1): p. 143-169.
8. Saitou, N., M.J.M.b. Nei, and evolution, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* 1987. **4**(4): p. 406-425.
9. Rashid, R., et al., *Geospatial mapping of microbial contaminants and public health risks in Lahore's municipal water.* 2025. **15**(1): p. 23079.
10. Polash, M.A.-U.-Z., et al., *Molecular Characterization and Antibiogram Profiling of Bacteria Isolated From Sewage and Surface Water in Bangladesh.* 2025. **2025**(1): p. 1848058.
11. Khan, G.A., et al., *Occurrence and abundance of antibiotics and resistance genes in rivers, canal and near drug formulation facilities–a study in Pakistan.* 2013. **8**(6): p. e62712.
12. Dueker, M.E., S. French, and G.D.J.F.i.M. O'Mullan, *Comparison of bacterial diversity in air and water of a major urban center.* 2018. **9**: p. 2868.
13. Clarridge III, J.E.J.C.m.r., *Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.* 2004. **17**(4): p. 840-862.
14. Saini, P., et al., *Genomic insights into virulence, antimicrobial resistance, and adaptation acumen of Escherichia coli isolated from an urban environment.* 2024. **15**(3): p. e03545-23.
15. Oren, A., G.M.J.I.j.o.s. Garrity, and e. microbiology, *Valid publication of the names of forty-two phyla of prokaryotes.* 2021. **71**(10): p. 005056.
16. Oren, A., M. Göker, and I.C.J.M. Sutcliffe, *New phylum names harmonize prokaryotic nomenclature.* 2022. **13**(5): p. e01479-22.
17. Wang, L., et al., *Shift in the microbial community composition of surface water and sediment along an urban river.* 2018. **627**: p. 600-612.
18. Yang, C., et al., *Spatiotemporal distribution of the planktonic microbiome and antibiotic resistance genes in a typical urban river contaminated by macrolide antibiotics.* 2024. **262**: p. 119808.
19. Kopecky, J., et al., *Phylogenetically related soil actinomycetes distinguish isolation sites by their metabolic activities.* 2023. **99**(12): p. fiad139.
20. Ulrich, N., et al., *Experimental studies addressing the longevity of Bacillus subtilis spores–The first data from a 500-year experiment.* 2018. **13**(12): p. e0208425.
21. Zhao, J., et al., *Global airborne bacterial community–interactions with Earth's microbiomes and anthropogenic activities.* 2022. **119**(42): p. e2204465119.
22. Gilbert, J.A. and B.J.N.R.M. Stephens, *Microbiology of the built environment.* 2018. **16**(11): p. 661-670.
23. Bushra, R., et al., *Untapped rich microbiota of mangroves of Pakistan: diversity and community compositions.* 2024. **69**(3): p. 595-612.
24. Noor, S.W., et al., *16S rRNA gene sequencing reveals bacterial diversity in Khewra Salt Mine walls.* 2024. **6**(12): p. 000869. v4.
25. Sivagnanasundaram, P., et al., *Assessment of airborne bacterial and fungal communities in selected areas of teaching hospital, Kandy, Sri Lanka.* 2019. **2019**(1): p. 7393926.
26. Gail, M.H., Y. Wan, and J.J.A.j.o.e. Shi, *Power of microbiome beta-diversity analyses based on standard reference samples.* 2021. **190**(3): p. 439-447.
27. Danko, D., et al., *A global metagenomic map of urban microbiomes and antimicrobial resistance.* 2021. **184**(13): p. 3376-3393. e17.

28. Milligan, E.G., et al., *A systematic review of culture-based methods for monitoring antibiotic-resistant Acinetobacter, Aeromonas, and Pseudomonas as environmentally relevant pathogens in wastewater and surface water.* 2023. **10**(2): p. 154-171.
29. Odhiambo, K.A., et al., *Contribution of pollution gradient to the sediment microbiome and potential pathogens in urban streams draining into Lake Victoria (Kenya).* 2023. **30**(13): p. 36450-36471.
30. F. Escapa, I., et al., *Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets.* 2020. **8**(1): p. 65.
31. Ha, M., et al., *Reliable identification of Bacillus cereus group species using low mass biomarkers by MALDI-TOF MS.* 2019.
32. Kim, S.-H., et al., *Identification and phylogenetic analysis of Enterococcus isolates using MALDI-TOF MS and VITEK 2.* 2023. **13**(1): p. 21.
33. McEwen, S.A. and P.J.J.M.s. Collignon, *Antimicrobial resistance: a one health perspective.* 2018. **6**(2): p. 10.1128/microbiolspec. arba-0009-2017.
34. Kinoshita, Y., et al., *Establishment and assessment of an amplicon sequencing method targeting the 16S-ITS-23S rRNA operon for analysis of the equine gut microbiome.* 2021. **11**(1): p. 11884.

TABLES

Table 1. Sampling frame: sites, isolates and land-use zones by environmental niche type

Niche type	Sites, n	Isolates sequenced, n	Isolates per site	Genera recovered, n	Land-use zones represented
Water	5	104	20.8	11	Residential (1); Commercial (1); Industrial (2); Peri-urban/transitional (1)
Soil	5	97	19.4	10	Residential (1); Green/park space (2); Peri-urban/transitional (2)
Air-deposition	5	99	19.8	8	Residential (2); Commercial (2); Industrial (1)
All niches	15	300	20.0	18	—

Table 2. Physicochemical and community characteristics of the 15 sampling sites, by environmental niche type

Characteristic and niche	Mean ± SD	Median (IQR)	Sites	p value
Isolates sequenced per site				0.262
Water	20.8 ± 1.3	21.0 (20.0–22.0)	n = 5	
Soil	19.4 ± 1.3	20.0 (18.0–20.0)	n = 5	
Air-deposition	19.8 ± 1.3	20.0 (19.0–21.0)	n = 5	
Cultivable density, $\log_{10}$ CFU/mL or CFU/g				0.004
Water	5.45 ± 0.66	5.39 (5.11–5.57)	n = 5	
Soil	6.23 ± 0.38	6.40 (5.91–6.52)	n = 5	
Air-deposition	3.36 ± 0.44	3.23 (3.03–3.65)	n = 5	
In-situ pH				0.600
Water	7.75 ± 0.12	7.82 (7.68–7.82)	n = 5	
Soil	7.66 ± 0.93	7.36 (7.15–8.31)	n = 5	
Air-deposition	Not applicable	—	n = 0	
Ambient temperature at collection, °C				0.254
Water	32.6 ± 1.9	33.3 (33.3–33.3)	n = 5	

Soil	29.4 ± 2.6	28.6 (28.0–28.7)	n = 5	
Air-deposition	29.7 ± 3.7	29.3 (28.0–30.6)	n = 5	
Genus richness per site				0.113
Water	8.6 ± 1.1	9.0 (8.0–9.0)	n = 5	
Soil	8.0 ± 0.7	8.0 (8.0–8.0)	n = 5	
Air-deposition	7.2 ± 0.8	7.0 (7.0–8.0)	n = 5	
Shannon diversity (H′)				0.174
Water	1.993 ± 0.122	2.048 (1.917–2.051)	n = 5	
Soil	1.923 ± 0.169	1.943 (1.923–1.963)	n = 5	
Air-deposition	1.752 ± 0.200	1.646 (1.593–1.937)	n = 5	
Pielou evenness (J′)				0.357
Water	0.930 ± 0.006	0.932 (0.924–0.933)	n = 5	
Soil	0.925 ± 0.044	0.935 (0.925–0.944)	n = 5	
Air-deposition	0.889 ± 0.059	0.885 (0.846–0.932)	n = 5	

CFU, colony-forming units; IQR, interquartile range; SD, standard deviation.

Table 3. Phylum- and class-level taxonomic composition of the 300 cultivable aerobic isolates, overall and by environmental niche

Phylum and class	All isolates, n (%)	Water, n (%)	Soil, n (%)	Air-deposition, n (%)	p value
Actinomycetota	90 (30.0)	4 (3.8)	40 (41.2)	46 (46.5)	<0.001
Actinomycetes	90 (30.0)	4 (3.8)	40 (41.2)	46 (46.5)	
Bacillota	82 (27.3)	5 (4.8)	37 (38.1)	40 (40.4)	<0.001
Bacilli	82 (27.3)	5 (4.8)	37 (38.1)	40 (40.4)	
Bacteroidota	21 (7.0)	16 (15.4)	5 (5.2)	0 (0.0)	<0.001
Flavobacteriia	21 (7.0)	16 (15.4)	5 (5.2)	0 (0.0)	
Pseudomonadota	107 (35.7)	79 (76.0)	15 (15.5)	13 (13.1)	<0.001
Gammaproteobacteria	107 (35.7)	79 (76.0)	15 (15.5)	13 (13.1)	

Data are n (%). Percentages in the niche columns are column percentages within that niche and sum to 100 down each column; the p value on each phylum row is from a 2 × 3 Pearson χ^2 test of that phylum against all others across niches.

Table 4. Genus-level composition of cultivable aerobic isolates by environmental niche, grouped by phylum

Genus	Family	All isolates, n (%)	Water, n (%)	Soil, n (%)	Air-deposition, n (%)
Pseudomonadota					
Acinetobacter	Moraxellaceae	30 (10.0)	16 (15.4)	6 (6.2)	8 (8.1)
Aeromonas	Aeromonadaceae	17 (5.7)	17 (16.3)	0 (0.0)	0 (0.0)
Enterobacter	Enterobacteriaceae	7 (2.3)	7 (6.7)	0 (0.0)	0 (0.0)
Escherichia	Enterobacteriaceae	8 (2.7)	8 (7.7)	0 (0.0)	0 (0.0)
Klebsiella	Enterobacteriaceae	13 (4.3)	13 (12.5)	0 (0.0)	0 (0.0)
Pseudomonas	Pseudomonadaceae	27 (9.0)	14 (13.5)	8 (8.2)	5 (5.1)
Stenotrophomonas	Xanthomonadaceae	5 (1.7)	4 (3.8)	1 (1.0)	0 (0.0)
Bacillota					
Bacillus	Bacillaceae	43 (14.3)	5 (4.8)	19 (19.6)	19 (19.2)
Lysinibacillus	Bacillaceae	18 (6.0)	0 (0.0)	10 (10.3)	8 (8.1)
Paenibacillus	Paenibacillaceae	8 (2.7)	0 (0.0)	8 (8.2)	0 (0.0)

Staphylococcus	Staphylococcaceae	13 (4.3)	0 (0.0)	0 (0.0)	13 (13.1)
Actinomycetota					
Arthrobacter	Micrococcaceae	15 (5.0)	0 (0.0)	15 (15.5)	0 (0.0)
Corynebacterium	Corynebacteriaceae	10 (3.3)	0 (0.0)	0 (0.0)	10 (10.1)
Kocuria	Micrococcaceae	9 (3.0)	0 (0.0)	0 (0.0)	9 (9.1)
Micrococcus	Micrococcaceae	39 (13.0)	4 (3.8)	8 (8.2)	27 (27.3)
Streptomyces	Streptomycetaceae	17 (5.7)	0 (0.0)	17 (17.5)	0 (0.0)
Bacteroidota					
Chryseobacterium	Weeksellaceae	7 (2.3)	7 (6.7)	0 (0.0)	0 (0.0)
Flavobacterium	Flavobacteriaceae	14 (4.7)	9 (8.7)	5 (5.2)	0 (0.0)
Data are n (%); niche columns are column percentages within that niche.					

Table 5. Species-level assignments from 16S rRNA gene sequencing, with mean sequence identity and niche distribution

Closest-match taxon	Isolates, n	Mean 16S identity (%)	Water / Soil / Air	Predominant niche
Acinetobacter — A. baumannii	18	99.07	10 / 4 / 4	Water (56%)
Acinetobacter — A. johnsonii	12	98.61	6 / 2 / 4	Water (50%)
Aeromonas — A. caviae	11	99.06	11 / 0 / 0	Water (100%)
Aeromonas — A. hydrophila	6	98.88	6 / 0 / 0	Water (100%)
Arthrobacter — A. globiformis	15	98.99	0 / 15 / 0	Soil (100%)
Bacillus — B. licheniformis	20	98.74	2 / 11 / 7	Soil (55%)
Bacillus — B. cereus	16	99.07	3 / 4 / 9	Air-deposition (56%)
Bacillus — B. subtilis	7	98.55	0 / 4 / 3	Soil (57%)
Chryseobacterium — C. indologenes	7	99.12	7 / 0 / 0	Water (100%)
Corynebacterium — C. glutamicum	10	98.51	0 / 0 / 10	Air-deposition (100%)
Enterobacter — E. cloacae	7	98.70	7 / 0 / 0	Water (100%)
Escherichia — E. coli	8	98.77	8 / 0 / 0	Water (100%)
Flavobacterium — F. johnsoniae	14	99.13	9 / 5 / 0	Water (64%)
Klebsiella — K. pneumoniae	13	99.04	13 / 0 / 0	Water (100%)
Kocuria — K. rhizophila	5	99.00	0 / 0 / 5	Air-deposition (100%)
Kocuria — K. palustris	4	99.14	0 / 0 / 4	Air-deposition (100%)
Lysinibacillus — L. fusiformis	18	98.54	0 / 10 / 8	Soil (56%)
Micrococcus — M. yunnanensis	21	98.85	2 / 5 / 14	Air-deposition (67%)
Micrococcus — M. luteus	18	98.96	2 / 3 / 13	Air-deposition (72%)
Paenibacillus — P. polymyxa	8	99.50	0 / 8 / 0	Soil (100%)
Pseudomonas — P. aeruginosa	13	98.67	5 / 6 / 2	Soil (46%)
Pseudomonas — P. fluorescens	12	99.31	8 / 1 / 3	Water (67%)

Pseudomonas — P. putida	2	98.84	1 / 1 / 0	Water (50%)
Staphylococcus — S. epidermidis	8	98.91	0 / 0 / 8	Air-deposition (100%)
Staphylococcus — S. hominis	5	99.32	0 / 0 / 5	Air-deposition (100%)
Stenotrophomonas — S. maltophilia	5	98.67	4 / 1 / 0	Water (80%)
Streptomyces — S. albidoflavus	11	98.85	0 / 11 / 0	Soil (100%)
Streptomyces — S. griseus	6	98.64	0 / 6 / 0	Soil (100%)

Table 6. Sequence quality, phenotypic identification scores and phylogenetic support, overall and by niche

Measure and niche	**Mean ± SD or n (%)**	**Median (IQR)**	**Isolates**	**p value**
16S rRNA BLAST identity, %	98.90 ± 0.80			0.271
Water	98.97 ± 0.76	99.04 (98.55–99.64)	n = 104	
Soil	98.92 ± 0.81	99.00 (98.35–99.50)	n = 97	
Air-deposition	98.81 ± 0.82	98.67 (98.29–99.48)	n = 99	
16S rRNA BLAST query coverage, %	97.6 ± 2.1			0.982
Water	97.5 ± 2.1	97.5 (95.8–99.6)	n = 104	
Soil	97.5 ± 2.2	98.2 (96.2–99.4)	n = 97	
Air-deposition	97.6 ± 1.9	97.8 (96.5–99.3)	n = 99	
MALDI-TOF MS log-score	2.01 ± 0.24			0.073
Water	2.05 ± 0.24	2.05 (1.91–2.22)	n = 104	
Soil	1.99 ± 0.23	2.00 (1.83–2.16)	n = 97	
Air-deposition	1.98 ± 0.25	1.96 (1.77–2.19)	n = 99	
VITEK 2 confidence, %	83.4 ± 10.1			0.449
Water	84.3 ± 10.1	85.4 (76.9–93.1)	n = 104	
Soil	83.5 ± 9.6	84.4 (75.7–90.5)	n = 97	
Air-deposition	82.3 ± 10.6	82.6 (74.4–89.8)	n = 99	
ML bootstrap support, %	83.4 ± 10.0			0.012
Water	82.7 ± 10.1	83.0 (75.0–90.0)	n = 104	
Soil	85.8 ± 9.5	86.0 (81.0–92.0)	n = 97	
Air-deposition	81.8 ± 10.2	81.0 (74.0–89.5)	n = 99	
Taxonomic level supported by 16S identity, n (%)				—
Species level	297 (99.0)			
Genus level	3 (1.0)			
Taxonomic level supported by MALDI-TOF log-score, n (%)				—
Species	152 (50.7)			
Genus	112 (37.3)			
Unreliable	36 (12.0)			

IQR, interquartile range; ML, maximum likelihood; SD, standard deviation.