Supplementary Materials

Unveiling the Human Nasopharyngeal Microbiome Compendium: Systematic Characterization of Community Architecture and Function Through a Comprehensive Meta-Analysis

Kuncheng Song[1], Hayden N Brochu[1], Monica L Bustos[1], Qimin Zhang[1], Crystal R Icenhour[1], Stanley Letovsky[1] and Lakshmanan K Iyer[1]

[1]Labcorp, Burlington, NC 27215, USA

Introduction

The following documents include the supplementary methods and results for many of the analyses that were conducted to support the main publication. If the methods were detailed in the main manuscript, we refrained from including the duplicated text in this document.

Methods

Large language model-assisted background contamination screening

We employed Claude[1] (Anthropic) to systematically evaluate bacterial families for potential contamination signatures. Each SILVA-annotated family underwent independent triplicate assessment using a standardized prompt designed to identify taxa inconsistent with nasopharyngeal ecology (**Figure 2A**). The LLM evaluated each family for likelihood of representing reagent contamination ("kitome")[2,3], environmental sources (water, soil, laboratory)[2], or legitimate nasopharyngeal colonizers. Consensus classification required agreement across all three independent queries, with discordant cases flagged for manual expert review. A microbiome specialist with expertise in respiratory tract ecology performed final adjudication of ambiguous classifications, incorporating published nasopharyngeal microbiome literature and contamination databases. This human-in-the-loop approach ensured that AI-assisted screening augmented rather than replaced expert

taxonomic knowledge, consistent with best practices for AI implementation in microbiome research[93].

Prompt for contamination assessment

"You are a nasopharyngeal microbiome expert tasked with identifying potential contaminants in 16S rRNA sequencing data. Evaluate the provided bacterial families for consistency with genuine nasopharyngeal colonization versus likely contamination from reagents, water sources, laboratory environments, or sample processing artifacts. Consider typical nasopharyngeal ecology, human commensal flora, and known contamination patterns in low-biomass samples.

Return results in CSV format with columns:

1. Bacterial Family (string)
2. Contaminant (boolean: TRUE/FALSE)
3. Source (string: Reagent/Water/Soil/Laboratory/Environmental/Not_applicable)

16S data processing

Raw sequencing data processing was performed using R v4.1.1[4]. Hypervariable region (HVR) targets and primer presence/absence were first identified by constructing a comprehensive list of possible HVR primer pairs and their maximum expected insert sizes (**Tables S1** and **S2**). A kmer hash of each possible HVR configuration was built by in silico amplifying various HVRs from the SILVA v138.2 database[5] using Mash with kmer size 31[6]. When kmer hashing could not distinguish HVRs but PCR primers were present, HVR configuration was inferred directly from the PCR primers (**Table S1**). This process independently verified the HVRs reported in corresponding source publications (**Table 1**). When PCR primers were detected, reads were reoriented to the same strand using an in-house R function utilizing the ShortRead[7] R package.

Quality filtering and trimming were performed using the DADA2 v1.22.0[8] filterAndTrim function. Parameters were optimized using an in-house R function that accounted for read lengths and quality profiles, presence/absence of PCR primers, and maximum expected hypervariable region insert sizes (**Table S2**). Quality-filtered trimmed reads underwent standard DADA2 processing including denoising and merging to generate amplicon sequence variants (ASVs). Singletons and chimeras were removed during ASV filtering.

Taxonomic classification utilized SILVA v138.2 database[5] training data formatted for DADA2 (obtained from Zenodo, DOI: 10.5281/zenodo.14169026). The DADA2 assignTaxonomy function was applied with minBoot=80 using silva_nr99_v138.2_toGenus_trainset.fa.gz, followed by the addSpecies function with default parameters using

silva_v138.2_assignSpecies.fa.gz. ASVs lacking at least family-rank classification were excluded from analysis. Final ASVs were aggregated at the lowest assigned taxonomic rank to generate count matrices for each study.

Beyond the contamination removal procedures described in the main manuscript, we implemented stringent quality control criteria at both sample and study levels. Individual samples required a minimum of 5,000 reads successfully mapped to family-level taxonomy following contaminant family removal to ensure adequate sequencing depth for reliable taxonomic profiling. To maintain dataset consistency for meta-analysis, we further required that studies retain at least 50% of samples passing these quality thresholds for inclusion in the final analysis. This two-tiered filtering strategy eliminated both low-quality samples and studies with systematic technical failures while preserving comparability that powered our study.

NPCST classification comparison between the before and after-background decontamination

Cluster preservation was evaluated using the Adjusted Rand Index (ARI) to quantify agreement between before and after background decontamination cluster assignments, with bootstrap resampling (n = 100) generating 95% confidence intervals. This evaluation used only the 7,790 high-quality samples retained in the final dataset after background decontamination. Internal cluster validity was assessed through silhouette analysis, which measured within-cluster cohesion relative to separation from neighboring clusters. Silhouette coefficients were calculated using Bray-Curtis dissimilarity matrices for k = 6 clusters derived from Ward's hierarchical clustering, enabling direct comparison of clustering quality between before and after background decontamination datasets.

To validate preservation of microbiome community relationships following decontamination, we employed complementary ordination-based approaches using Procrustes analysis and Mantel tests on Bray-Curtis distance matrices. Procrustes analysis optimally rotated and scaled principal coordinate analysis (PCoA) ordinations to maximize alignment between before and after background decontamination datasets, yielding a correlation coefficient and M^2 statistic with 999 permutations. The Mantel test independently evaluated the correlation between pairwise sample distances in both distance matrices using Pearson correlation coefficients with 999 permutations.

Taxonomic resolution comparison between V3-V4 and V4

To determine the appropriate taxonomic rank for meta-analysis, we compared taxonomic resolution between genus and species levels using only studies employing V3-V4 or V4

hypervariable regions, as other regions were rare in our dataset and did not warrant comparison. We performed rarefaction analysis to assess taxonomic saturation at both ranks, calculated the cumulative relative abundance of taxa shared between V3-V4 and V4 regions, and quantified the proportion of sequences that could not be classified at each taxonomic level. Taxa with relative abundance below 0.01% were excluded from all analyses to minimize noise from rare sequences.

Leave-one-study-out (LOSO) NPCST investigation

To assess the reproducibility and stability of identified NPCSTs, we implemented a leave-one-study-out cross-validation approach across all 28 studies. For each iteration, one study was systematically removed, and hierarchical clustering was performed on the remaining 27 studies using Bray-Curtis dissimilarity at the genus level followed by Ward linkage, with cluster assignments tested for k values ranging from 4 to 12. The complete 28-study dataset served as ground truth, and generated clusters were matched to reference NPCSTs using a greedy assignment algorithm based on contingency table analysis, prioritizing clusters with the highest intersection-to-union ratios.

Clustering stability was quantified using three metrics: (1) Adjusted Rand Index (ARI) to measure overall clustering agreement corrected for chance, (2) Mean Jaccard Index to assess average per-cluster similarity between predicted and ground truth assignments, and (3) Overall Accuracy to calculate the proportion of samples correctly assigned to their corresponding ground truth NPCST after optimal cluster matching.

Unsupervised learning based definition of the rare biosphere (ulrb)

To objectively define abundance categories within each CST, we applied the ulrb[9] method as described by Pascoal et al. Following the recommended approach, we employed the default tri-categorization framework (k=3) to classify genera into "Abundant", "Undetermined", and "Rare" categories within each sample. The quality of clustering was evaluated using Silhouette scores, with scores >0.5 indicating reasonable to strong cluster structure across all CSTs. Abundant genera within each CST were characterized by their median relative abundance, detection prevalence across samples, and clustering quality metrics.

Machine learning approach for validating NPCST classification

We developed and validated a comprehensive machine learning framework to classify nasopharyngeal swab samples into six previously defined NPCST categories using relative abundance data from 626 genera across 7,790 samples spanning 28 independent studies. We validated this model using 28 studies and further evaluated its performance on two

external datasets. Each method underwent hyperparameter optimization and evaluation through 100 iterations of 5-fold cross-validation, with performance assessed using accuracy, precision, recall, and F1-score metrics to ensure comprehensive evaluation across all NPCST categories. To ensure robust model generalization across diverse study populations, we implemented a stratified cross-validation strategy that maintained balanced distribution of both target NPCST classifications and source studies (BioProjects) origins within each fold, thereby preventing potential batch effects from influencing model performance and enhancing the generalizability of the final model.

For machine learning model selection, we focused on algorithms widely used in the microbiome field and methodologically distinct approaches across tree-based, regression-based, and kernel-based methods. The selected models included Random Forest (randomForest[10] v4.7-1.2) with hyperparameters including mtry values ranging across different numbers of genera and ntree values of 50, 100, 200, 500, and 1,000 trees; Ridge, LASSO, and elastic net regression (glmnet[11] v4.1-9) with regularization parameter λ optimized through 5-fold cross-validation and alpha fixed at 0 (Ridge), 0.1–0.9 (Elastic Net), and 1 (LASSO); and Support Vector Machine (e1071[12] v1.7-16) with radial basis function kernel, cost parameters ranging from 0.1 to 100, and gamma parameters from 0.001 to 1.0. For each of the iterations and individual 5-fold cross validation, each of these hyperparameter was examined and the best one is recorded.

During evaluation, we removed LASSO regression from analysis because many 5-fold cross-validation sets failed to converge, resulting in over 30% missing data points. We calculated performance metrics using the caret[13] package (v7.0-1), employing balanced accuracy to account for potential class imbalances, where overall accuracy represented the macro-averaged balanced accuracy across all classes and per-class accuracy corresponded to individual class balanced accuracy. Additionally, we tracked Random Forest feature importance through mean decrease in Gini impurity and mean decrease in accuracy.

For consistency analysis, we conducted a detailed evaluation of error patterns from the machine learning models (incorrect predictions on test datasets across 100 iterations) stratified by severity levels. Low-severity errors (<15%, i.e., 15/100 iterations) predominantly represented method-specific weaknesses that could be random. Moderate (15%–50%) to high (>50%) severity errors demonstrated samples that were repeatedly misclassified, indicating that machine learning methods struggled with the same samples and suggesting intrinsic classification challenges rather than methodological limitations.

To enhance deployed prediction models (SVM and Random Forest) reliability, we developed a confidence assessment framework using empirical data from test sets across

100 cross-validation iterations. We applied Youden's J statistic optimization to determine genus-specific probability and relative abundance cutoffs that maximize separation between correct and incorrect predictions for each NPCST classification. High confidence classifications were defined as samples exceeding both the optimized prediction probability and corresponding key genus relative abundance thresholds, while low confidence classifications were assigned to samples falling below both thresholds. We examined the prediction probability range and relative abundance range less than 0.75 as samples exceed this level are almost correctly classified. This approach was applied to NPCSTs I-IV and VI; NPCST V was excluded because its diverse composition precludes reliable confidence assessment without comprehensive characterization of all nasopharyngeal samples.

NPCST classification model deployment

The final SVM and Random Forest models were independently trained on the complete dataset of 626 genera across 7,790 samples from 28 independent studies. We developed customized functions to enable future users to apply these models to new datasets. The deployment pipeline validates genera naming conventions before performing NPCST classifications and generates confidence scores for both SVM and Random Forest predictions. Complete implementation instructions are provided in the Zenodo repository (DOI: 10.5281/zenodo.17068997).

Co-occurrence network

We used UpSet plots from ComplexUpset[14] (v1.3.3) to demonstrate shared association patterns (positive or negative) across NPCST-specific networks. We performed network centrality analysis using the igraph[15] (v2.1.1) R package to calculate closeness, betweenness, and degree centrality measures for genera and edges across individual NPCST-specific networks and the global co-occurrence network.

External validation of the NPCST prediction model

For external validation evaluations, input data underwent the nasopharyngeal-specific background removal protocol followed by data validation and NPCST prediction according to the deployed classification guide available in the Zenodo repository. We established ground-truth classifications for these external validation samples using the same Bray-Curtis dissimilarity followed by Ward linkage methodology, combining the 28 original studies with the 2 external validation studies, then selected the top 6 NPCSTs. We performed ROC calculations using the pROC[16] package (v1.18.5) in R.

Nasopharyngeal microbiome health index (NMHI)

Adapted from Chang et al.'s GMWI2 methodology[17], we developed the Nasopharyngeal Microbiome Health Index (NMHI) using LASSO-penalized logistic regression (glmnet[11] R package v4.1) with balanced class weights to address sample size imbalances. We calculated class weights as ($w_{healthy}$ = 0.5/($n_{healthy}$/n_{total}), $w_{disease}$ = 0.5/($n_{disease}$/n_{total})), ensuring equal class contribution regardless of imbalance. From 5,435 cross-sectional nasopharyngeal samples, we constructed binary presence/absence matrices using a 0.01% relative abundance threshold (present=1, absent=0).

We implemented four binary classification models to four distinct models: (1) healthy controls versus combined diseased samples with all-taxa (All-taxa All-Conditions), (2) healthy controls versus viral infections with all-taxa (All-taxa Viral Infection), (3) healthy controls versus combined diseased samples with genus-only taxa (Genus-only All-Conditions), and (4) healthy controls versus viral infections with genus-only taxa (Genus-only Viral Infection). This design enabled assessment of both taxonomic granularity and disease specificity effects on model performance.

We evaluated seven taxonomic configurations (all taxa combined, phylum, class, order, family, genus, and species), excluding unclassified reads to ensure interpretability of results. Model development and validation proceeded through five stages:

Stage 1: Leave-One-NPCST-Out Cross-Validation for Lambda Selection: We implemented LONO cross-validation to establish stable lambda values, leveraging the biological distinctiveness of NPCSTs and their differential disease susceptibilities. This approach systematically held out each NPCST (I-VI) as a test set while training on the remaining five, ensuring generalization across biologically meaningful community states rather than technical batch effects. Given that NPCSTs explained substantially more variance than study effects (53.19% vs. 13.06%), this strategy provided robust parameter selection. We tested selective lambda values ranging from 0.0001 to 0.03, selecting optimal values based on maximum AUC aggregated across all six held-out NPCST test sets.

Stage 2: Model Performance Evaluation with Prevalence Thresholds: Using optimal lambda values from Stage 1, we performed both LONO and 10-fold cross-validation at five prevalence thresholds (0%, 1%, 5%, 10%, 20%) to assess model robustness. The LONO is run only once for the cross-validation and the 10-fold cross-validation repeated 10 times with reproducible seed numbers (seeds 1-10) to ensure reproducibility while capturing variability. We maintained strict train-test separation within each fold to prevent data leakage, confining lambda selection exclusively to training partitions. The NMHI score was

calculated as the sum of products between coefficients and binary presence values, where positive coefficients indicated health-associated taxa and negative coefficients indicated disease-associated taxa. We assessed model performance using both AUC and balanced accuracy metrics to evaluate discriminatory power and classification performance.

Stage 3: Final Model Training: Using the selected 5% prevalence threshold (based on optimal performance-interpretability trade-off), we trained final models on the complete dataset with optimal lambda values. During this stage, we extracted all non-zero coefficients and intercepts from each model, enabling NMHI score calculation as the sum of the intercept and products of coefficients with presence/absence values for each sample's taxa. To identify key microbial markers, we analyzed taxa with absolute coefficients ≥0.5 and performed comparative abundance analysis across control and disease groups. We visualized abundance distributions using boxplots and assessed statistical significance using Wilcoxon rank-sum tests with FDR correction for multiple testing comparisons.

Stage 4: NMHI Threshold Optimization: Using models trained on the complete dataset, we optimized classification thresholds to distinguish healthy from diseased samples. We evaluated both global (across all NPCSTs) and NPCST-specific thresholds ranging from -5 to 5 (at 0.1 increments), selecting optimal values based on maximum balanced accuracy for each model configuration. We calculated Cohen's d effect sizes to quantify the magnitude of separation between healthy and diseased populations, providing a standardized measure of discriminatory power independent of sample size. We also performed Wilcoxon rank-sum tests with FDR correction for multiple testing to statistically compare disease and control samples for both global and NPCST-specific models.

Stage 5: External Validation: External validation utilized both cross-sectional and longitudinal samples (n=699 total), treating each sampling point as independent given the transient nature of nasopharyngeal microbiome communities during infection. This approach tests the model's ability to distinguish disease states regardless of sampling design, reflecting real-world diagnostic applications where single-timepoint sampling is standard. To ensure valid assessment, we removed samples with ambiguous disease classifications (e.g., pneumococcal carriers, emergency room volunteers) and samples not consistently obtained during symptomatic disease periods. The validation cohort comprised healthy family members with recurrent respiratory tract infections (n=265), varying SARS-CoV-2 severity (n=398), lower respiratory tract infections (n=5), and non-SARS-CoV-2 critically ill patients (n=31). We applied final model coefficients to calculate NMHI scores and generate predictions for these previously unseen samples, evaluating

performance using ROC (Receiver Operating Characteristic) curve analysis and balanced accuracy metrics with both global and NPCST-specific thresholds. The NPCST classification was based on the random forest model we provided in the earlier section. For NPCST-specific validation, we excluded NPCST V from AUC analysis when all samples belonged to a single class (disease), as meaningful discrimination requires representation of both classes.

Results

Validation of signal integrity following background decontamination

Following implementation of our three-stage decontamination pipeline (**Figure 2A**), we evaluated quality control metrics across all 28 studies to validate background removal effectiveness. The sigmoid distribution patterns observed across studies demonstrated that most samples retained robust true signal abundance (>80% cumulative relative abundance) with >5,000 reads after contamination removal, exhibiting minimal background interference (**Figure S1**). Of 8,314 total samples, 7,986 (96.1%) successfully exceeded the 5,000 true-signal read threshold, confirming that our decontamination approach preserves sufficient sequencing depth for downstream analyses. This consistent retention pattern across diverse studies validates our pipeline's capacity to eliminate spurious signals while maintaining biological integrity. The final dataset comprised 7,790 samples after additional quality control for complete disease/health status annotation and exclusion of rare positive and negative control samples. After applying the same blacklist background removal protocol, we revealed only three novel genera (*Tersicoccus*, *Bact-08*, *Eoetvoesia*) absent from our training data, all at minimal abundances all around 0.02%, demonstrating comprehensive capture of the core nasopharyngeal microbiome across diverse populations and confirming model applicability to new cohorts.

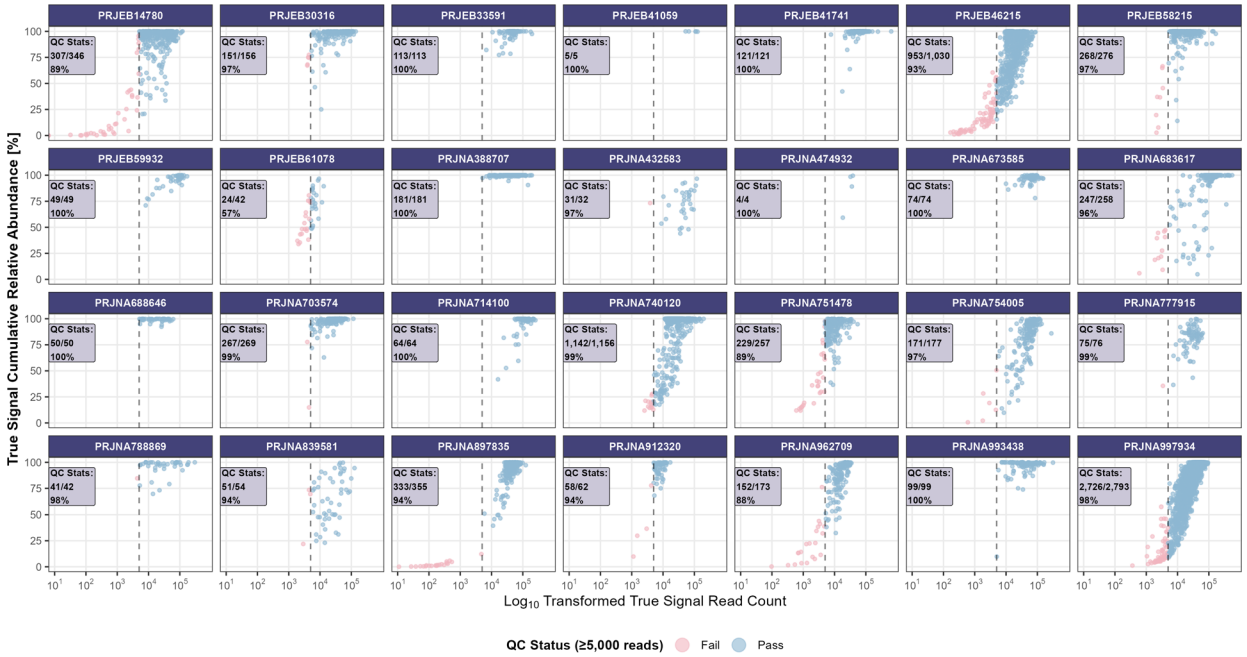

Figure S1. Study-specific quality control assessment. Quality control evaluation of nasopharyngeal microbiome samples across 28 retained studies, displaying the relationship between $\log_{10}$-transformed true signal read counts and cumulative relative abundance following background removal. Each panel presents study-specific retention statistics (passed/total samples and percentage) based on the 5,000-read threshold criterion, with blue points representing QC-passed samples and pink points indicating failed samples.

Taxonomic resolution comparison between V3-V4 and V4

Rarefaction analysis revealed that the majority of samples reached diversity plateaus before 5,000 reads for both genus and species levels, confirming the adequacy of our quality control parameters (**Figure S2A-B**). Although selected samples with higher microbial richness required over 10,000 reads for complete saturation, genus-level rarefaction curves consistently plateaued earlier than species-level curves across all samples. Evaluation of cumulative relative abundance of shared taxa between V3-V4 and V4 regions identified 485 shared genera and 887 shared species (**Figure S2C**). Notably, nearly all V4-identified taxa were present within the V3-V4 dataset, while V3-V4 contained additional unique taxa, and expected pattern given the inclusion of the V3 region. Genus-level classification demonstrated significantly better consistency between regions, with virtually all taxonomic assignments shared between V3-V4 and V4, enabling harmonized analyses independent of hypervariable region selection.

Analysis of unassigned sequences revealed substantial differences in classification success between taxonomic levels, regardless of hypervariable region (**Figure S2D**). Species-level classification failed for a median of 75.9% (V3-V4) and 84.7% (V4) of reads, with high variability across studies (SD = 25.0% and 18.1%, respectively). Nearly all studies contained samples with >50% species-level assignment failure, creating severe resolution imbalances that would compromise downstream analyses. In contrast, genus-level classification maintained robust performance with median unassigned proportions of only 0.4% (V3-V4) and 0.6% (V4). These findings demonstrate that genus-level classification provides the taxonomic resolution necessary for reliable meta-analysis across heterogeneous nasopharyngeal microbiome studies from 16S data.

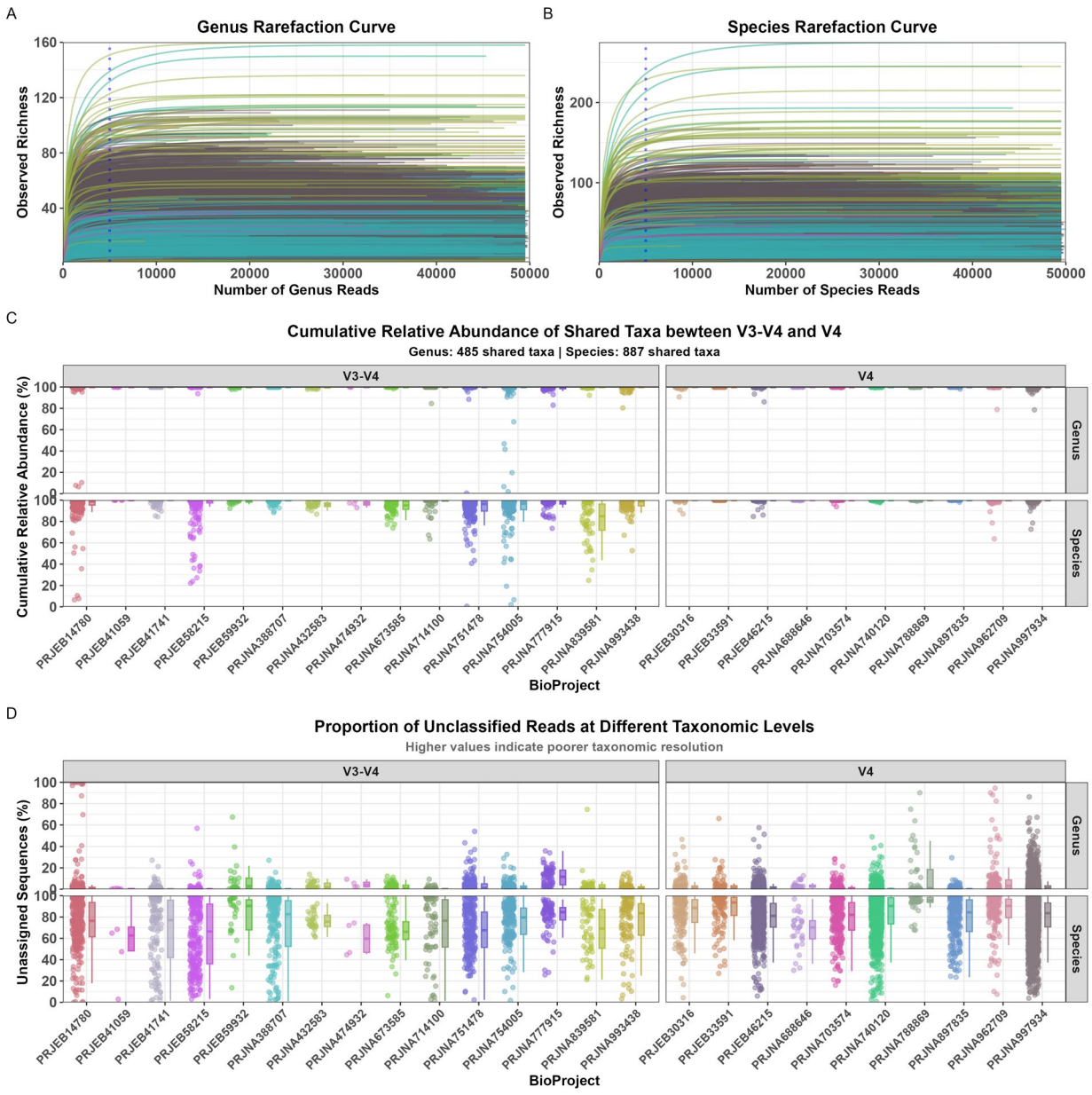

Figure S2. Taxonomic resolution comparison between genus and species levels for V3-V4 and V4 16S rRNA gene regions in nasopharyngeal microbiome meta-analysis. A-B. Rarefaction curves demonstrate that genus-level diversity approaches saturation while species-level richness continues to increase without plateauing, with a vertical blue dotted line indicating the 5,000 read threshold. **C**. Cumulative relative abundance of shared taxonomic features between V3-V4 and V4 regions reveals that genus-level classification captures substantially higher proportions of the microbial community (485 shared genera) compared to species-level classification (887 shared species), stratified by hypervariable region and BioProject. **D**. Proportion of unclassified sequences shows consistently higher assignment failure rates at species level compared to genus level across both V3-V4 and V4

regions, indicating poor species-level resolution reliability. Colors throughout all panels represent distinct BioProjects included in the meta-analysis.

Comparison of microbial community structure before and after background decontamination

Comprehensive validation analyses confirmed that our decontamination pipeline preserved biological signal integrity while enhancing data quality. Cluster assignments demonstrated high stability, with 6,327 of 7,790 samples (81.2%) maintaining their original nasopharyngeal community state type classification and moderate-to-strong agreement between before and after background decontamination datasets (ARI = 0.605, 95% CI: 0.592–0.617). Clustering quality improved substantially by 18.5%, increasing from 0.245 to 0.291, with cluster and group transitions visualized in **Figure S3A–B**. Procrustes analysis revealed near-perfect preservation of sample relationships (correlation = 0.973, $M^2 = 0.05$, $p < 0.001$), corroborated by Mantel test results showing exceptionally strong correlation between distance matrices ($r = 0.971$, $p < 0.001$). These convergent lines of evidence demonstrate that removing 1,810 contaminating genera (reducing the dataset from 2,436 to 626 genera) enhanced detection of genuine nasopharyngeal microbiome patterns without distorting underlying biological relationships. Principal coordinate analysis further validated this preservation, with the first two dimensions each exhibiting approximately 2% increased variance explained after background decontamination, indicating improved sample separation (**Figure S3C–D**). The reduction in study-associated variance from 13.97% to 13.06% (R^2 from PERMANOVA) represents meaningful mitigation of batch effects. Sparsity decreased from 97.89% to 95.53% following decontamination, as removing 1,810 predominantly sparse contaminating genera eliminated approximately 14 million data points (mostly zeros) while resulting in a denser matrix with 2.36 percentage points fewer zeros, thereby improving data quality for downstream analyses.

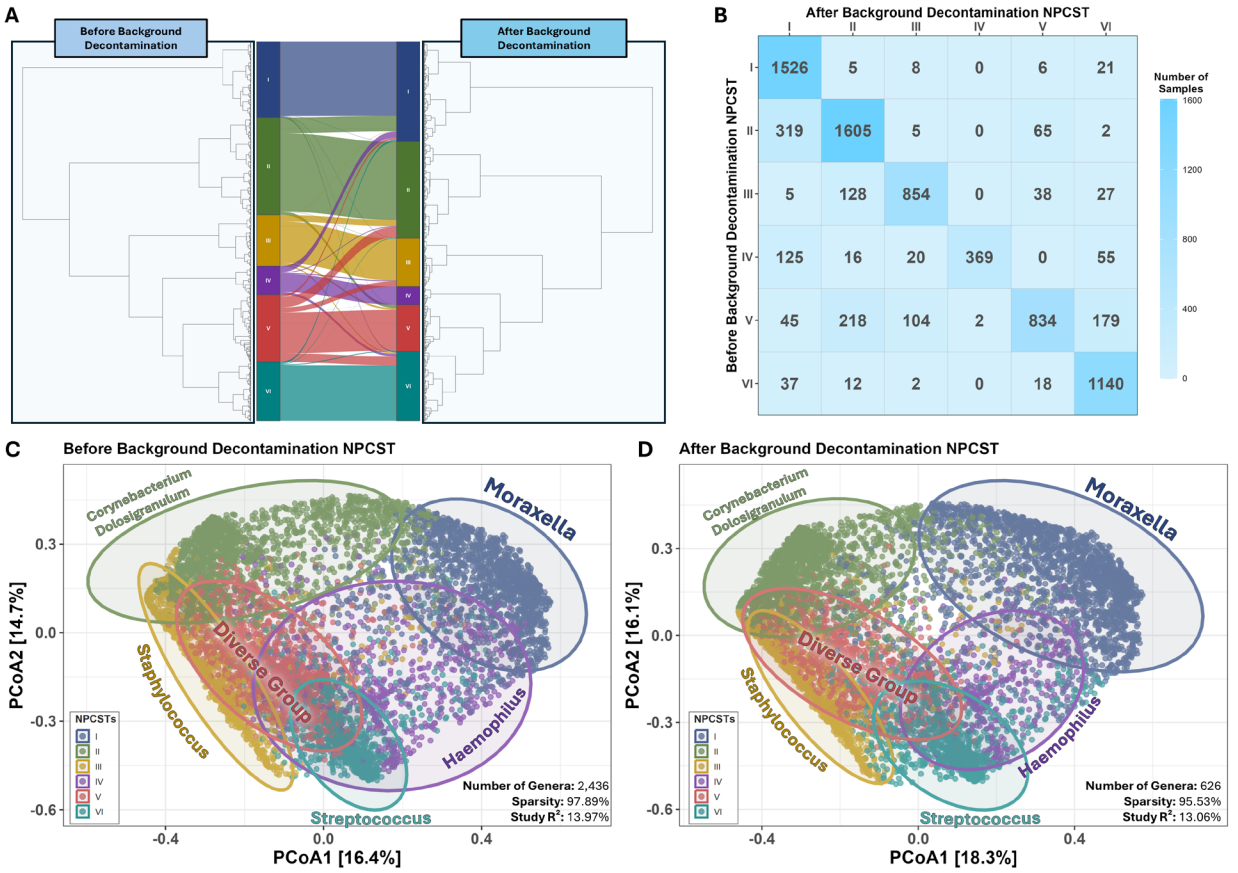

Figure S3. Comparison of microbial NPCST structure before and after background decontamination. A. Hierarchical clustering dendrograms displaying six NPCSTs identified in before and after background decontamination datasets, with an alluvial plot illustrating sample transitions between corresponding clusters. **B.** Confusion matrix quantifying the redistribution of samples across NPCSTs from before background decontamination (x-axis) to after background decontamination (y-axis) stages. **C&D.** Principal coordinate analysis (PCoA) ordination of the two PCoA dimensions before and after decontamination datasets, respectively. Ecological metrics including the number of retained genera, matrix sparsity (genera × samples), and variance explained (R^2) from PERMANOVA analysis are displayed in the lower right corner of each panel.

NPCST-specific cumulative relative abundance of top 14 families

We examined the NPCST-specific median cumulative abundance patterns (**Figure S4**). The top six families elevated most samples in NPCSTs I-IV to achieve 80-90% cumulative relative abundance. In contrast, the more diverse NPCSTs V and VI demonstrated slower abundance progression. Collectively, the top 14 families explained a median of at least 90% of total relative abundance across all samples and NPCSTs, confirming that these 14

families (comprising 161 genera and 513 ASVs) represent the core nasopharyngeal bacterial composition.

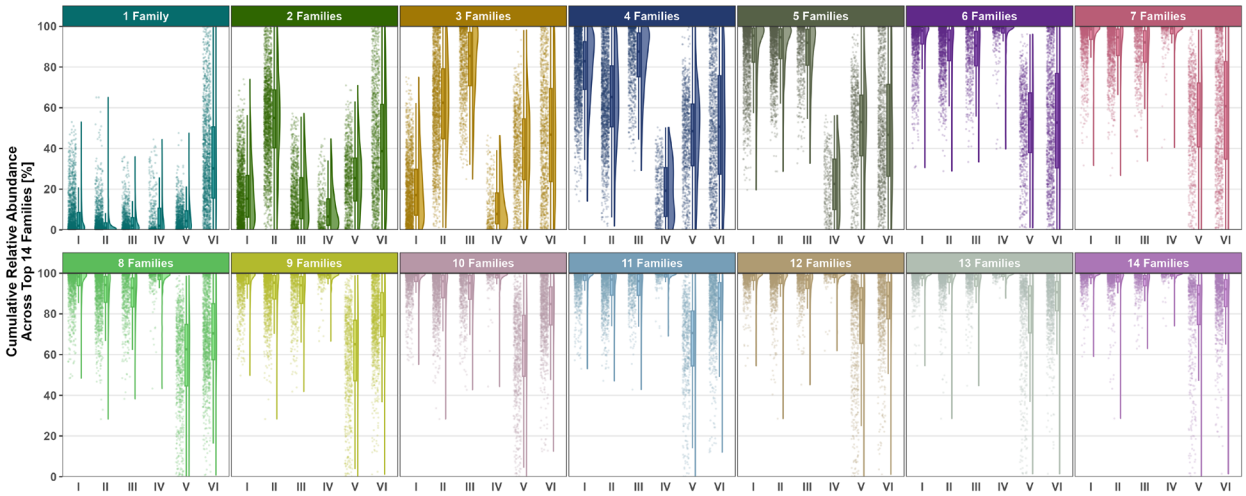

Figure S4. NPCST-specific cumulative relative abundance of top 14 families. Cumulative relative abundance distributions across the top 14 most prevalent families after background decontamination, with each panel representing sequential accumulation of relative abundance from the highest-ranked family through progressively lower-ranked families (left to right), demonstrating sample-level variability within each cumulative stage across six NPCSTs.

LOSO validations

Leave-one-study-out cross-validation analysis identified 6 NPCSTs as the optimal clustering solution, demonstrating robust stability across all evaluated metrics (**Figure S5**). At k=6, the Adjusted Rand Index achieved 0.711 and the mean Jaccard Index reached 0.75, indicating strong clustering agreement and high per-cluster similarity that substantially exceeded random performance. Overall median accuracy attained 0.867, demonstrating that 87% of samples were correctly assigned to their corresponding ground truth NPCSTs during cross-validation. These converging stability metrics collectively support the selection of 6 NPCSTs as the most reproducible and biologically meaningful clustering

structure in the nasopharyngeal microbiome data.

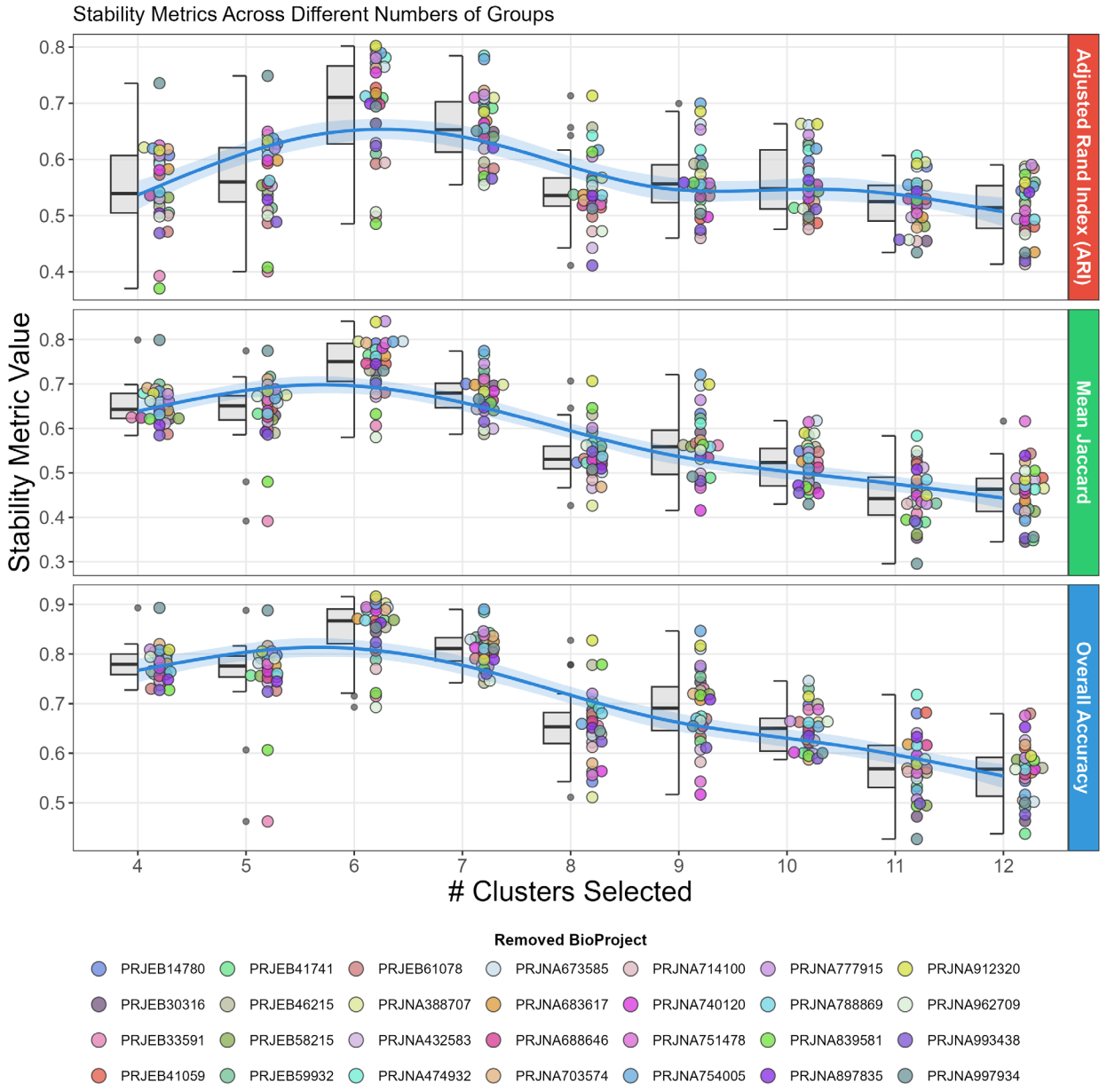

Figure S5. Stability metrics across different cluster numbers evaluated through leave-one-study-out (LOSO) cross-validation. Half boxplots (left, gray) show distribution quartiles for each cluster number. Individual data points (right, colored circles) represent stability measurements from the LOSO CV with different removed BioProjects. Blue trend lines show generalized additive model (GAM)-smoothed curves with 95% confidence intervals (shaded regions). Top panel: Adjusted Rand Index (ARI) measures clustering agreement corrected for chance. Middle panel: Mean Jaccard Index quantifies average cluster overlap. Bottom panel: Overall Accuracy represents proportion of correctly assigned samples.

Ulrb results

The ulrb analysis revealed distinct abundance patterns across all NPCSTs, with detailed results presented in **Figure S6-11**. The two diverse NPCSTs (V and VI) exhibited substantially higher numbers of abundant genera (94 and 65 genera, respectively) compared to the remaining NPCSTs (4-27 abundant genera each). This pattern corresponded well with the increased overall genera diversity observed in these two NPCSTs, confirming their classification as compositionally heterogeneous community types. One key distinguishing pattern among NPCSTs was the higher relative abundance of Streptococcus in specific community types. Silhouette score distributions consistently indicated strong clustering quality for genera abundance classifications across all NPCSTs, validating the robustness of the abundance categorizations.

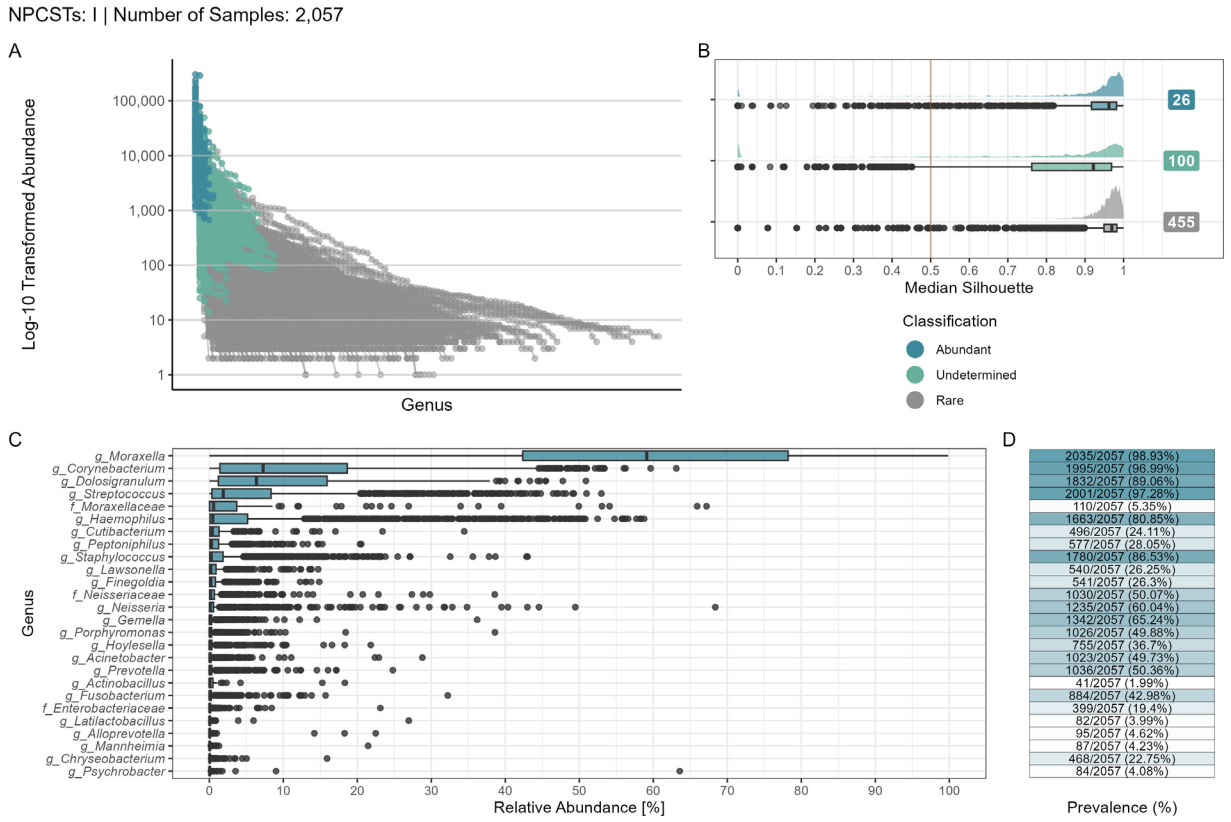

Figure S6. Ulrb diagnostic plots for NPCST I (n = 2,057 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are

displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

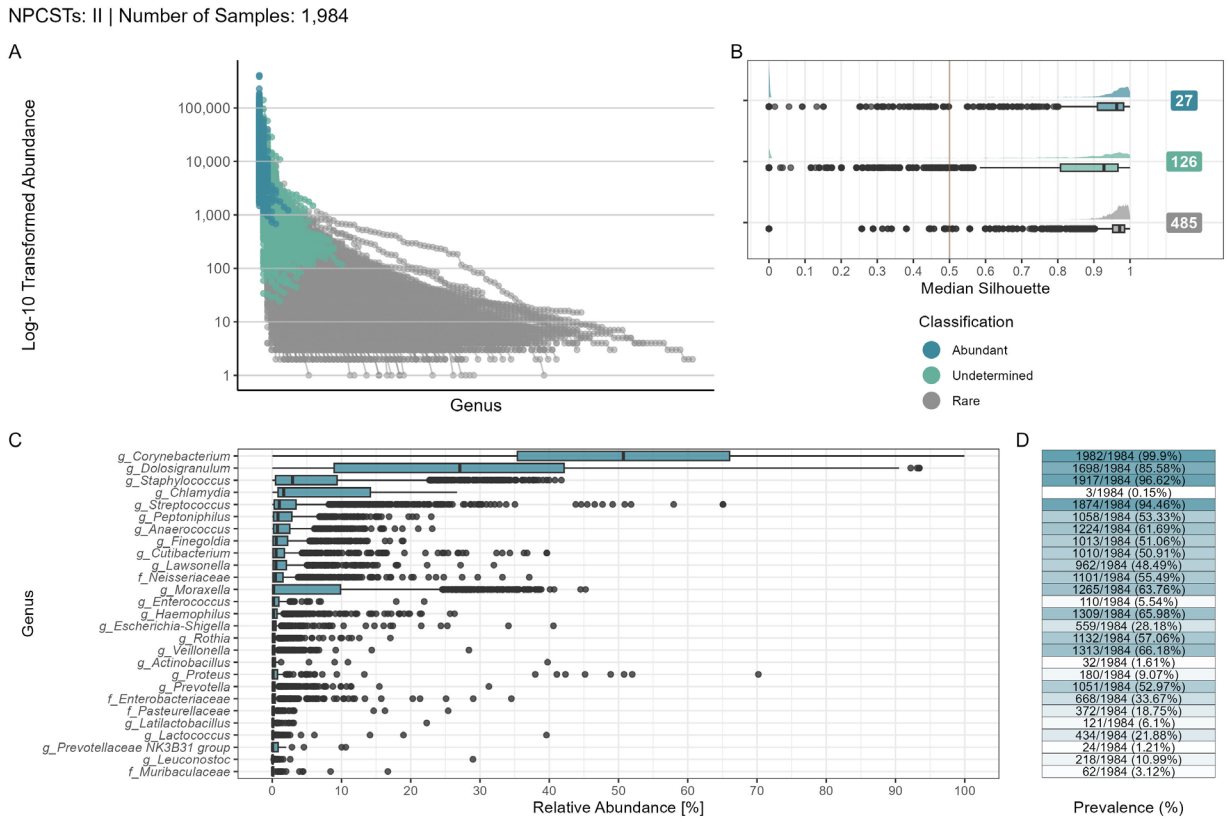

Figure S7. Ulrb diagnostic plots for NPCST II (n = 1,984 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

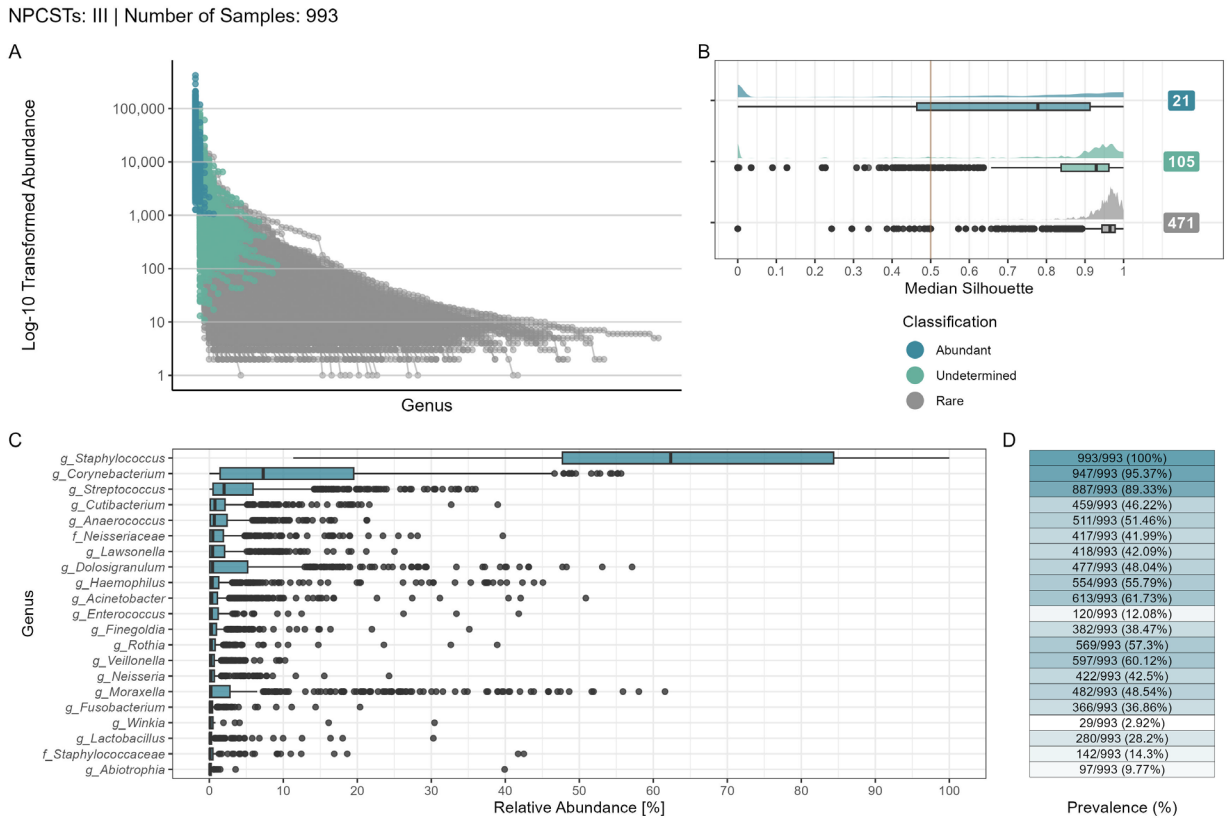

Figure S8. Ulrb diagnostic plots for NPCST III (n = 993 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

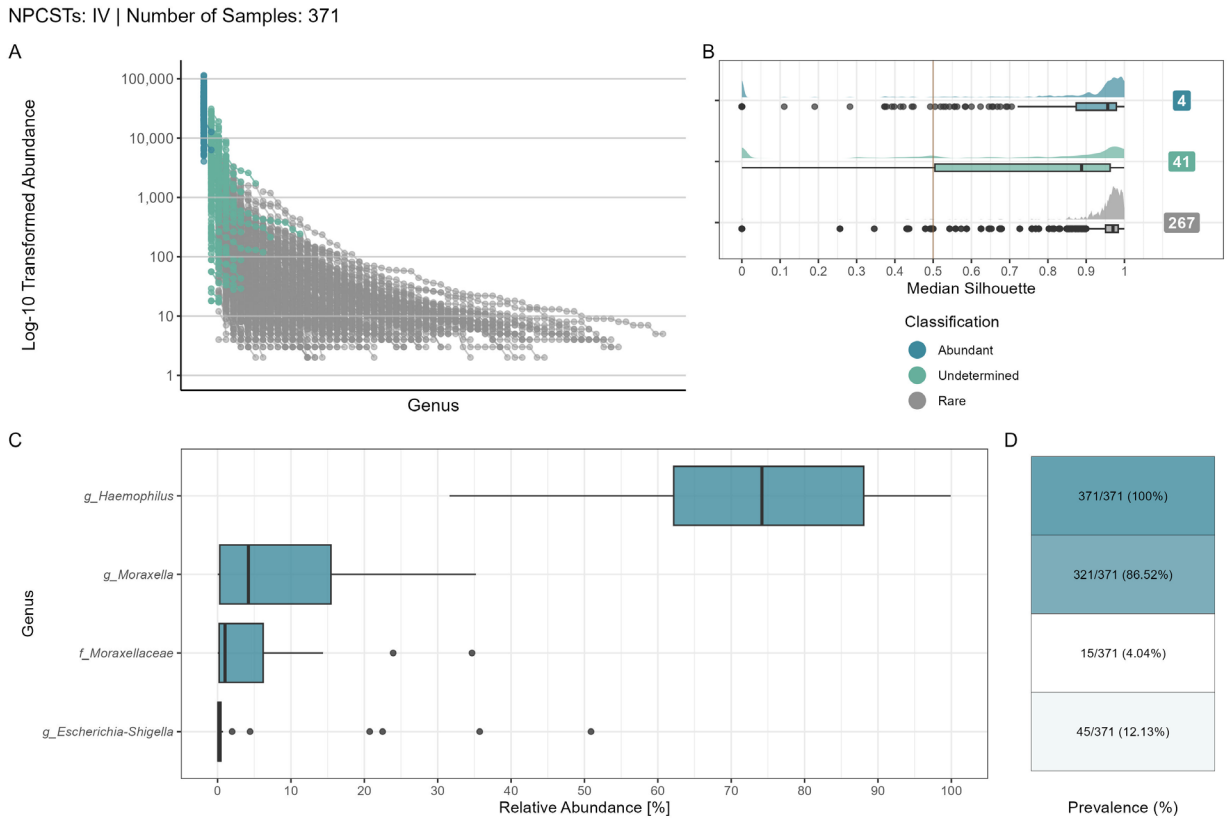

Figure S9. Ulrb diagnostic plots for NPCST IV (n = 371 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

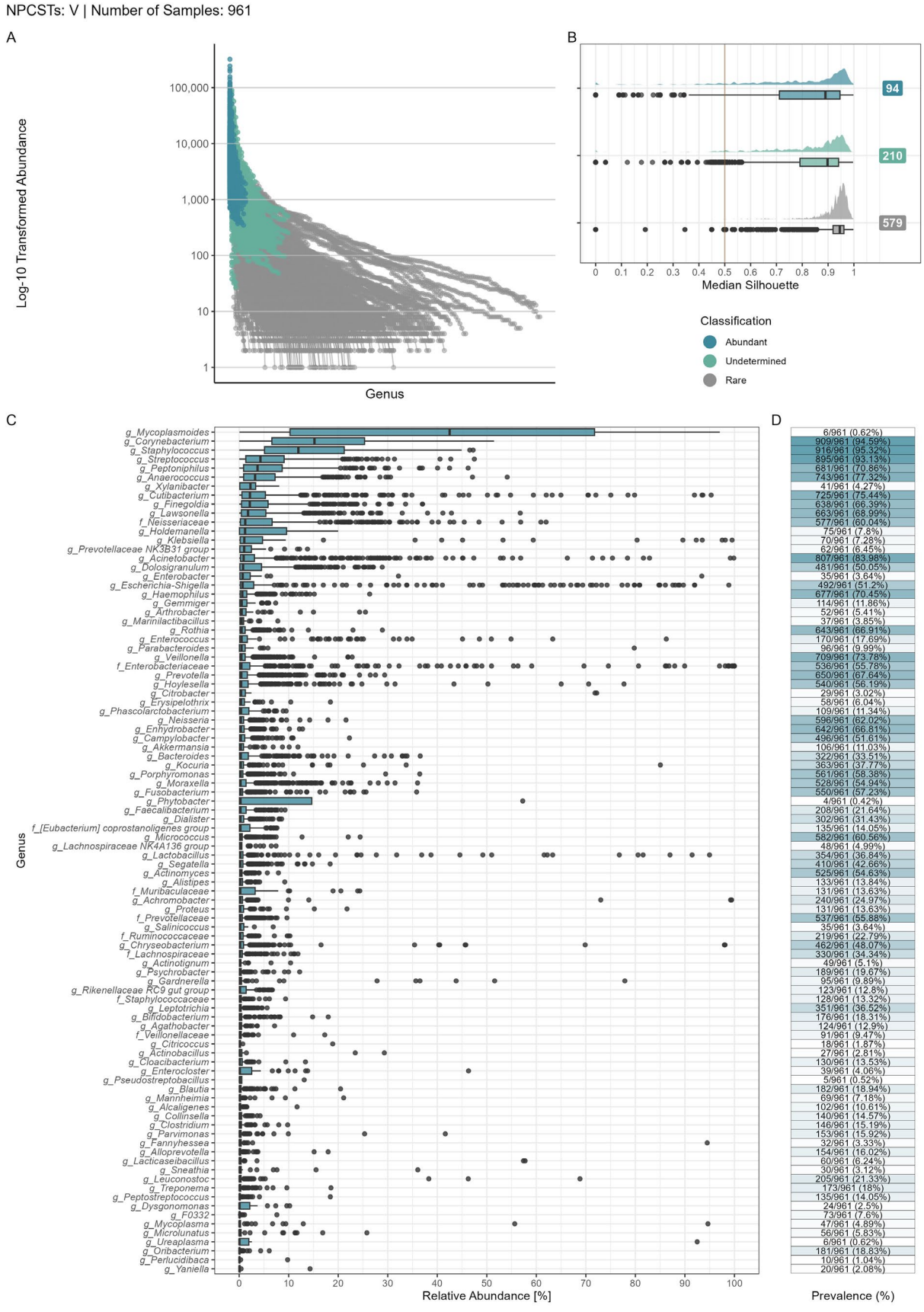
NPCSTs: V | Number of Samples: 961
A
Log-10 Transformed Abundance
100,000
10,000
1,000
100
10
1
Genus
B
94
210
579
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Median Silhouette
Classification
Abundant
Undetermined
Rare
C
g_Mycoplasmoides
g_Corynebacterium
g_Staphylococcus
g_Streptococcus
g_Peptoniphilus
g_Anaerococcus
g_Xylanibacter
g_Cutibacterium
g_Finegoldia
g_Lawsonella
f_Neisseriaceae
g_Holdemanella
g_Klebsiella
g_Prevotellaceae NK3B31 group
g_Acinetobacter
g_Dolosigranulum
g_Enterobacter
g_Escherichia-Shigella
g_Haemophilus
g_Gemmiger
g_Arthrobacter
g_Marinilactibacillus
g_Rothia
g_Enterococcus
g_Parabacteroides
g_Veillonella
f_Enterobacteriaceae
g_Prevotella
g_Hoylesella
g_Citrobacter
g_Erysipelothrix
g_Phascolarctobacterium
g_Neisseria
g_Enhydrobacter
g_Campylobacter
g_Akkermansia
g_Bacteroides
g_Kocuria
g_Porphyromonas
g_Moraxella
g_Fusobacterium
g_Phytobacter
g_Faecalibacterium
g_Dialister
f_[Eubacterium] coprostanoligenes group
g_Micrococcus
g_Lachnospiraceae NK4A136 group
g_Lactobacillus
g_Segatella
g_Actinomyces
g_Alistipes
f_Muribaculaceae
g_Achromobacter
g_Proteus
f_Prevotellaceae
g_Salinicoccus
f_Ruminococcaceae
g_Chryseobacterium
f_Lachnospiraceae
g_Actinotignum
g_Psychrobacter
g_Gardnerella
g_Rikenellaceae RC9 gut group
f_Staphylococcaceae
g_Leptotrichia
g_Bifidobacterium
g_Agathobacter
f_Veillonellaceae
g_Citricoccus
g_Actinobacillus
g_Cloacibacterium
g_Enterocloster
g_Pseudostreptobacillus
g_Blautia
g_Mannheimia
g_Alcaligenes
g_Collinsella
g_Clostridium
g_Parvimonas
g_Fannyhessea
g_Alloprevotella
g_Lacticaseibacillus
g_Sneathia
g_Leuconostoc
g_Treponema
g_Peptostreptococcus
g_Dysgonomonas
g_F0332
g_Mycoplasma
g_Microlunatus
g_Ureaplasma
g_Oribacterium
g_Perlucidibaca
g_Yaniella
Genus
0 10 20 30 40 50 60 70 80 90 100
Relative Abundance [%]
D
6/961 (0.62%)
909/961 (94.59%)
916/961 (95.32%)
895/961 (93.13%)
681/961 (70.86%)
743/961 (77.32%)
41/961 (4.27%)
725/961 (75.44%)
638/961 (66.39%)
663/961 (68.99%)
577/961 (60.04%)
75/961 (7.8%)
70/961 (7.28%)
62/961 (6.45%)
807/961 (83.98%)
481/961 (50.05%)
35/961 (3.64%)
492/961 (51.2%)
677/961 (70.45%)
114/961 (11.86%)
52/961 (5.41%)
37/961 (3.85%)
643/961 (66.91%)
170/961 (17.69%)
96/961 (9.99%)
709/961 (73.78%)
536/961 (55.78%)
650/961 (67.64%)
540/961 (56.19%)
29/961 (3.02%)
58/961 (6.04%)
109/961 (11.34%)
596/961 (62.02%)
642/961 (66.81%)
496/961 (51.61%)
106/961 (11.03%)
322/961 (33.51%)
363/961 (37.77%)
561/961 (58.38%)
528/961 (54.94%)
550/961 (57.23%)
4/961 (0.42%)
208/961 (21.64%)
302/961 (31.43%)
135/961 (14.05%)
582/961 (60.56%)
48/961 (4.99%)
354/961 (36.84%)
410/961 (42.66%)
525/961 (54.63%)
133/961 (13.84%)
131/961 (13.63%)
240/961 (24.97%)
131/961 (13.63%)
537/961 (55.88%)
35/961 (3.64%)
219/961 (22.79%)
462/961 (48.07%)
330/961 (34.34%)
49/961 (5.1%)
189/961 (19.67%)
95/961 (9.89%)
123/961 (12.8%)
128/961 (13.32%)
351/961 (36.52%)
176/961 (18.31%)
124/961 (12.9%)
91/961 (9.47%)
18/961 (1.87%)
27/961 (2.81%)
130/961 (13.53%)
39/961 (4.06%)
5/961 (0.52%)
182/961 (18.94%)
69/961 (7.18%)
102/961 (10.61%)
140/961 (14.57%)
146/961 (15.19%)
153/961 (15.92%)
32/961 (3.33%)
154/961 (16.02%)
60/961 (6.24%)
30/961 (3.12%)
205/961 (21.33%)
173/961 (18%)
135/961 (14.05%)
24/961 (2.5%)
73/961 (7.6%)
47/961 (4.89%)
56/961 (5.83%)
6/961 (0.62%)
181/961 (18.83%)
10/961 (1.04%)
20/961 (2.08%)
Prevalence (%)

Figure S10. Ulrb diagnostic plots for NPCST V (n = 961 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

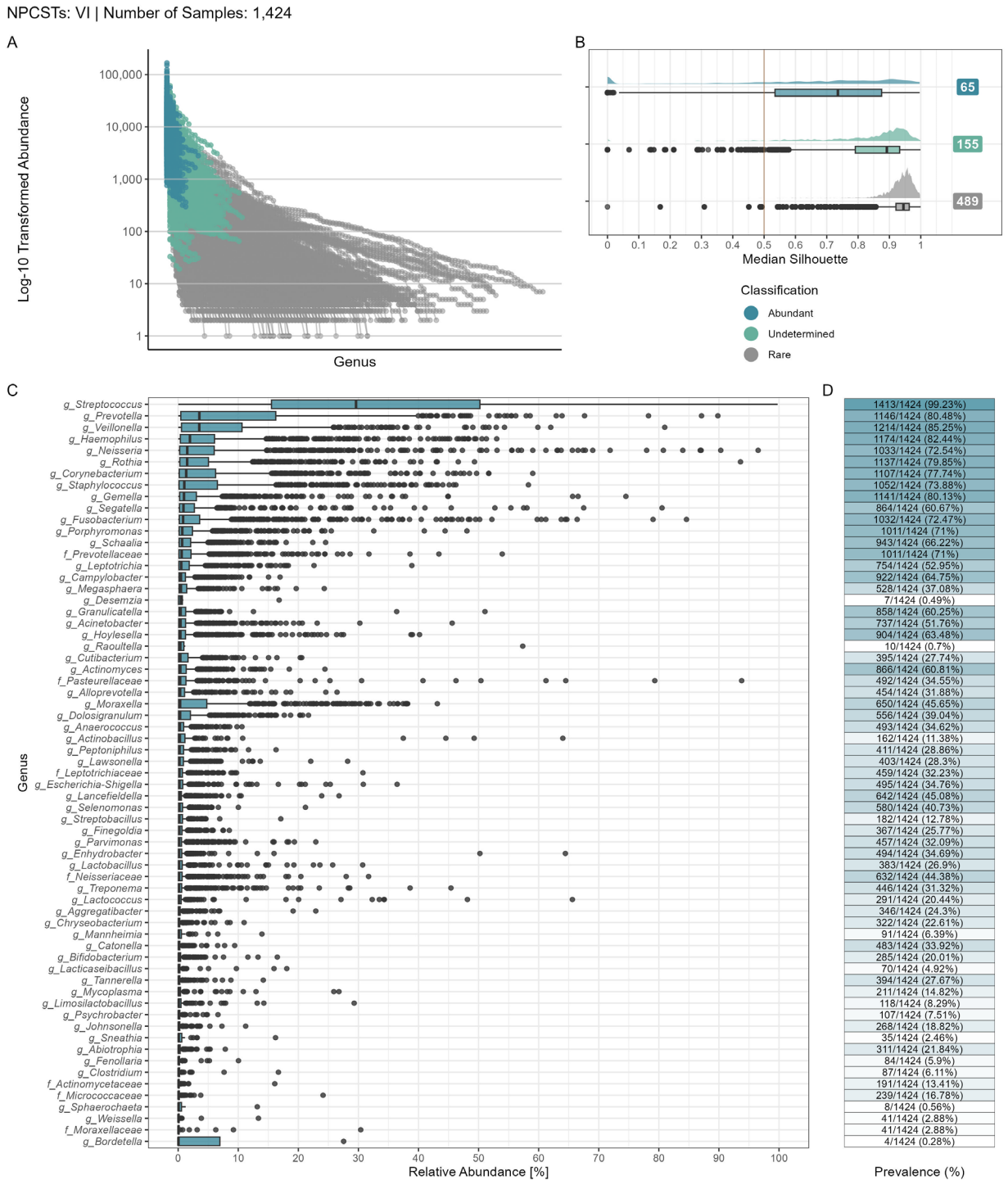

Figure S11. Ulrb diagnostic plots for NPCST VI (n = 1,424 samples). A. Log_{10}-transformed relative abundance distribution across all genera, stratified by ulrb classifications ("Abundant", "Undetermined", and "Rare"). **B.** Silhouette score distributions shown as density plots and boxplots for each classification. The quality threshold (Silhouette score = 0.5) is indicated by the red dashed line. Numbers of unique genera per classification are

displayed adjacent to each plot. **C.** Genus-level relative abundance for genera classified as "Abundant" by ulrb. **D.** Detection prevalence of abundant genera across samples, expressed as percentages. Classifications are color-coded: Abundant (blue), Undetermined (teal), and Rare (gray). Prevalence is represented on a gradient scale from 0-100% (white to blue).

NPCST machine learning results

Model evaluation

Across 100 iterations of 5-fold cross-validation, machine learning models demonstrated strong performance for NPCST prediction. SVM achieved the highest performance (0.972), followed by Random Forest (0.966), Elastic Net (0.936), and Ridge (0.922) (**Table S4, Figure S12-13**). Given that Elastic Net and Ridge regression models showed significantly lower performance compared to SVM and Random Forest, we excluded these methods from detailed evaluation and focused on the two superior-performing algorithms. When examining per-NPCST performance, NPCST V (the diverse NPCST) demonstrated significantly reduced performance metrics compared to other NPCSTs. SVM achieved superior performance on NPCST V with 0.936 accuracy on the testing set, while Random Forest achieved 0.922.

For hyperparameter optimization, we evaluated 500 runs across 100 iterations of 5-fold cross-validation to identify optimal parameters. Random Forest hyperparameter optimization revealed mtry=209 (representing one-third of the 626 genera features) and ntree=50 as the most prevalent and optimal selection. Specifically, ntree=500 was optimal in 345/500 runs (69%), followed by ntree=100 in 152/500 runs (30.4%) and ntree=150 in 3/500 runs (0.6%), while mtry=209 was utilized across all runs. Importantly, our evaluation showed negligible performance differences between 50 and 100 trees, supporting selection of ntree=50 for computational efficiency. SVM optimization consistently identified cost=100 and gamma=1 as optimal parameters across all 500 runs (100% consistency).

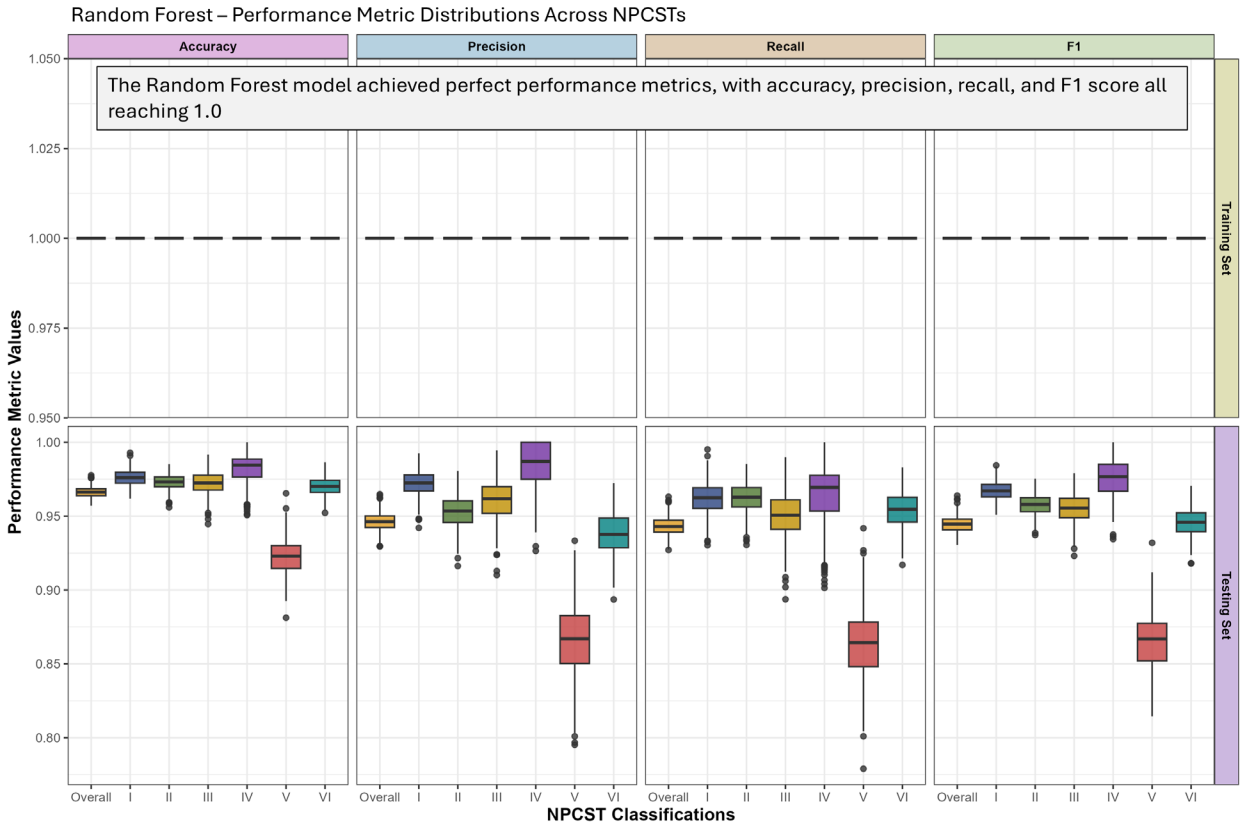

Figure S12. Random forest performance metric distributions across NPCST classification types for training and testing datasets. Boxplots display the distribution of balanced accuracy, precision, recall, and F1-score values across 100 iterations of 5-fold cross-validation for each NPCST class (Overall represents macro-averaged performance across all classes, i.e., simple arithmetic mean, while I-VI represent individual NPCST). Each panel compares training set performance (yellow strips) against testing set performance (purple strips).

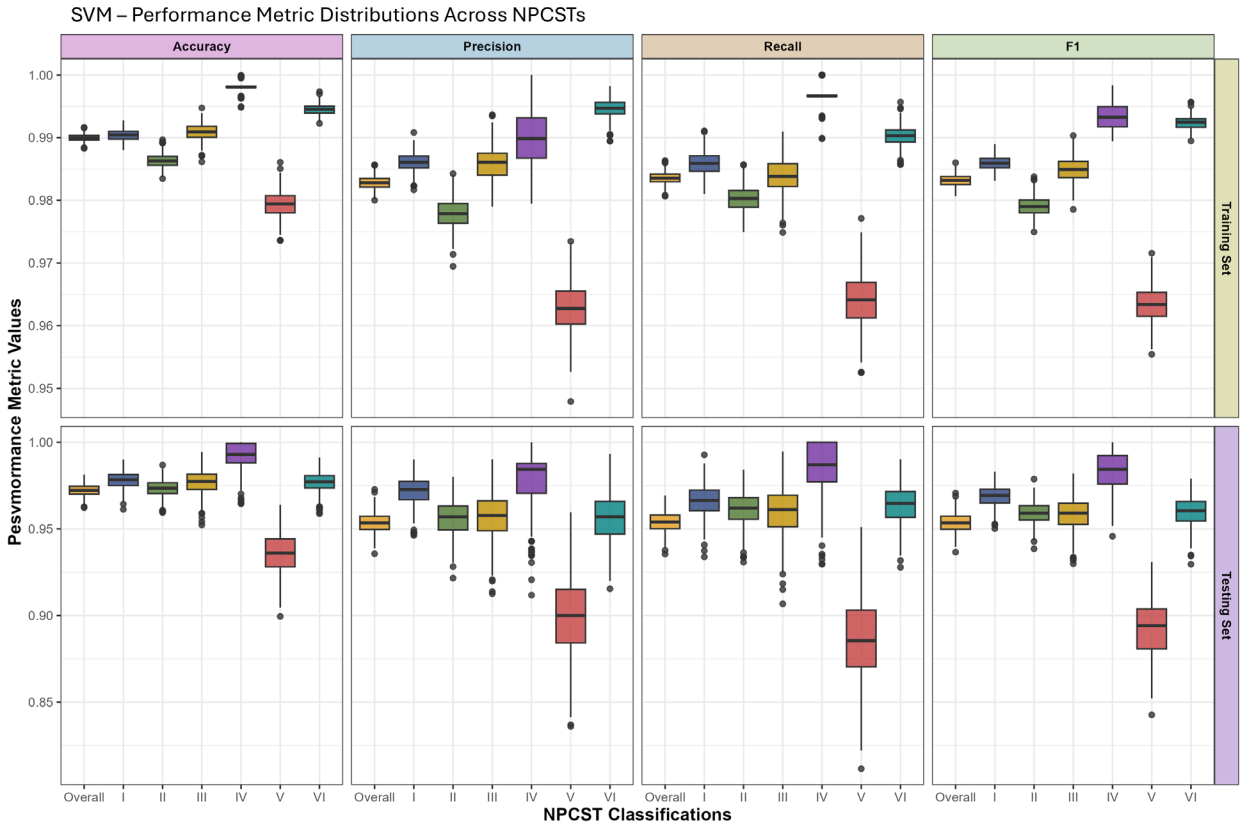

Figure S13. SVM performance metrics distribution across NPCST classification types for training and testing datasets. Boxplots display the distribution of balanced accuracy, precision, recall, and F1-score values across 100 iterations of 5-fold cross-validation for each NPCST class (Overall represents macro-averaged performance across all classes, i.e., simple arithmetic mean, while I-VI represent individual NPCST). Each panel compares training set performance (yellow strips) against testing set performance (purple strips).

Random forest feature importance

Feature importance analysis using Random Forest mean decrease in accuracy and Gini impurity metrics identified the most predictive genera for NPCST classification (**Table S6**). The six NPCST-defining genera ranked among the top six important features based on both accuracy and Gini rankings, validating our previous analytical findings and confirming their biological relevance for NPCST classification.

SVM and random Forest incorrect prediction analysis

We evaluated the frequency and severity of incorrect predictions by SVM and Random Forest models relative to the six key NPCST-defining genera. Analysis of relative abundance differences across these genera for each NPCST revealed significant statistical differences between correct and incorrect predictions for both models (**Figure S14-15**). Wilcoxon tests

confirmed that the dominant NPCST-defining genera were the primary drivers of misclassification, with incorrect predictions consistently occurring when these key genera fell below characteristic abundance thresholds. This finding directly motivated our development of the confidence evaluation system described in the following section, which leverages these genus-specific thresholds to flag potentially ambiguous classifications.

Next, we examined the consistency of misclassified samples between SVM and Random Forest models. This analysis focused on the 3.34% and 2.8% misclassified samples across cross-validation runs for SVM and Random Forest, respectively. Frequency analysis of misclassified samples demonstrated that SVM and Random Forest collectively misclassified 34-53% of incorrectly predicted samples, with the remainder distributed as method-specific errors (**Figure S16**). Furthermore, severity stratification analysis revealed distinct error patterns: samples with low error rates (<50%) showed only 8-9% consensus misclassification between methods, while high-error samples (>50%) exhibited 42% consensus misclassification. These results indicated that each method exhibited distinct types of method-dependent misclassifications.

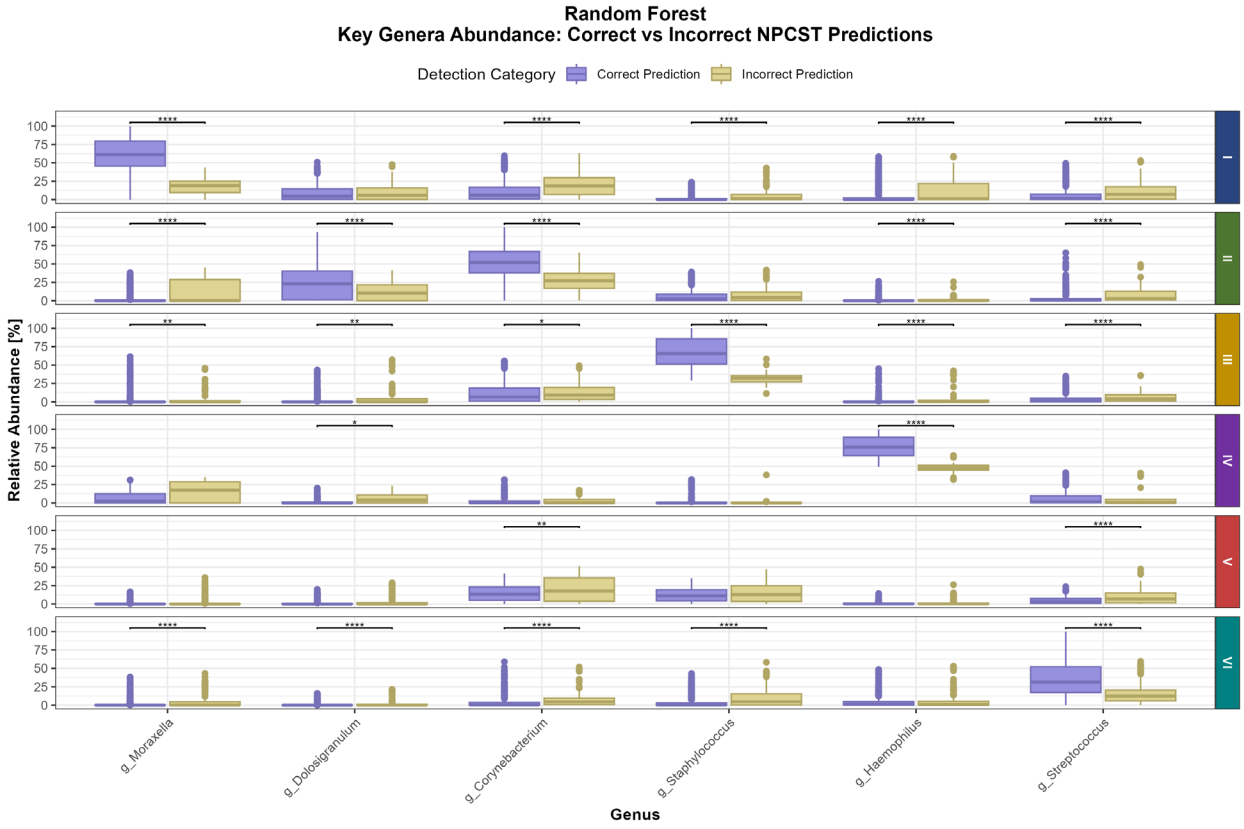

Figure S14. Relative abundance of key bacterial genera in correct versus incorrect Random Forest NPCST Testing Set predictions. Boxplots show the relative abundance

[%] distributions for six genera grouped by random forest accuracy and stratified by NPCST. Statistical significance between groups for each genus/NPCST was assessed using Wilcoxon rank-sum tests with FDR correction, with significant differences indicated above boxplots. Statistical significance levels are denoted as * p ≤ 0.05, **p ≤ 0.01, *** p ≤ 0.001, and **** p ≤ 0.0001 for adjusted p-values (p.adj).

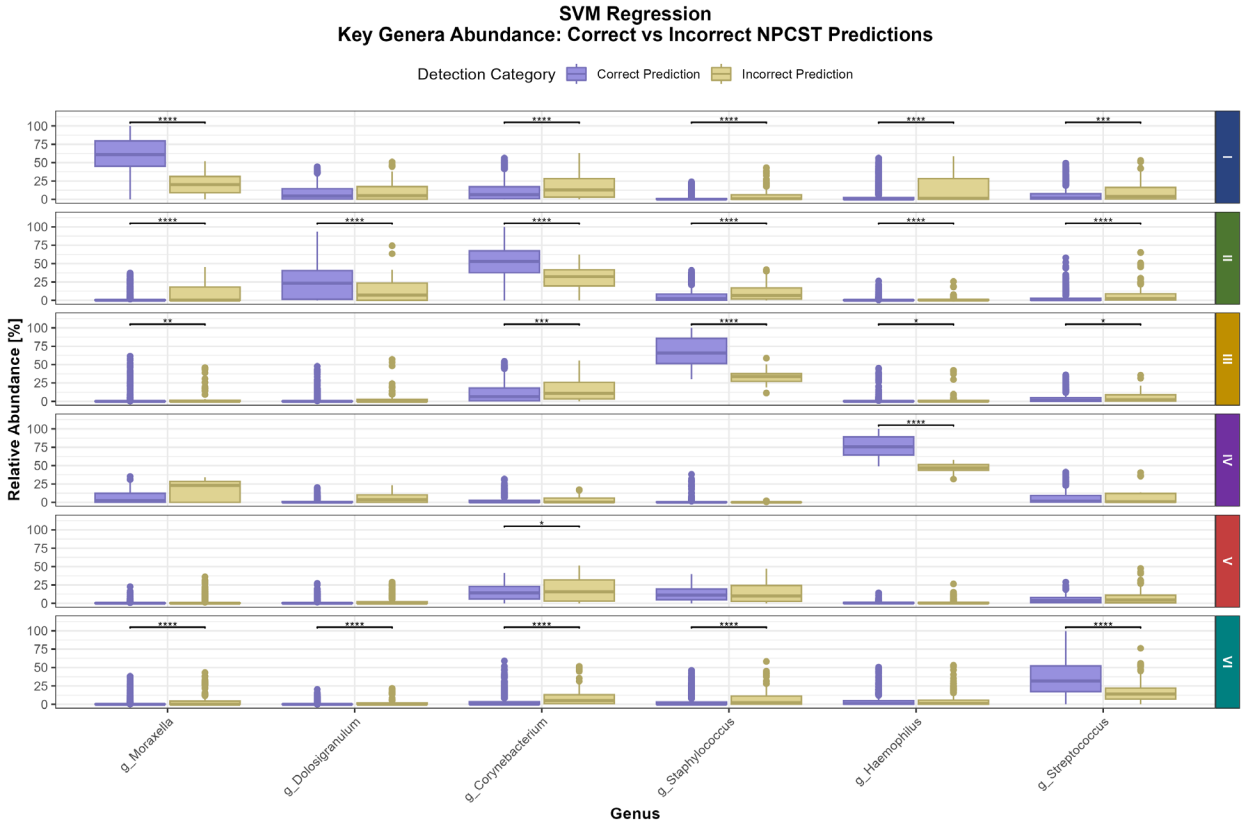

Figure S15. Relative abundance of key bacterial genera in correct versus incorrect SVM NPCST Testing Set predictions. Boxplots show the distribution of relative abundance [%] for six genera grouped by random forest accuracy and stratified by NPCST. Statistical significance between groups was assessed for each genus/NPCST using Wilcoxon rank-sum tests with FDR correction, with significant differences indicated above boxplots. Statistical significance levels are denoted as * p ≤ 0.05, **p ≤ 0.01, *** p ≤ 0.001, and **** p ≤ 0.0001 for adjusted p-values (p.adj).

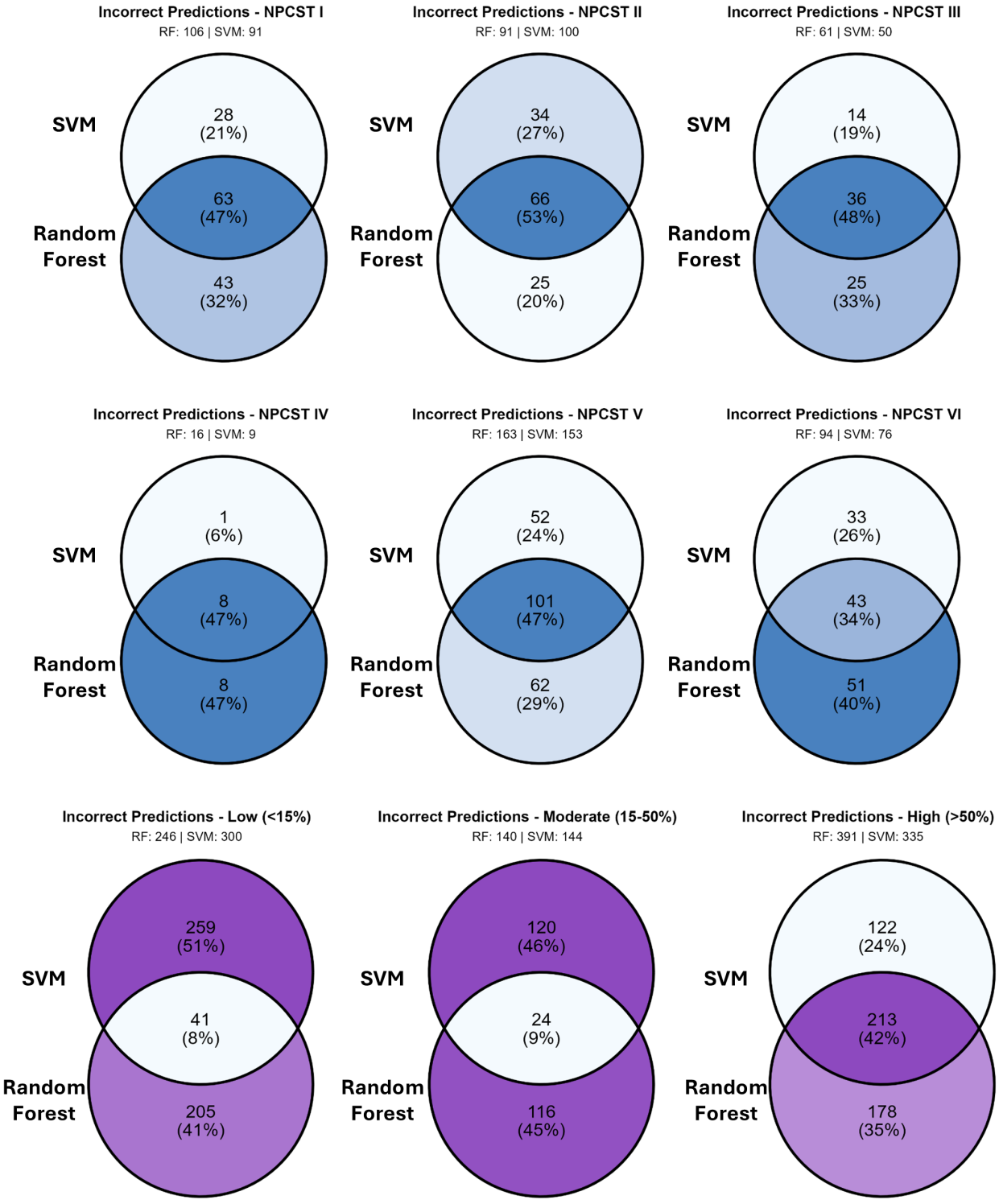

Figure S16. Incorrect prediction pattern analysis for NPCST. Venn diagrams illustrate overlapping incorrect predictions between Random Forest and Support Vector Machine (SVM) models across different analytical dimensions. The upper two rows (blue theme)

display method-specific and consensus prediction errors for each of the six NPCSTs, with overlapping regions indicating samples consistently misclassified by both methods. The bottom row (purple theme) stratifies prediction errors by severity levels: low (<15% error rate), moderate (15-50% error rate), and high (>50% error rate) incorrect predictions across 100 cross-validation iterations.

SVM and random forest confidence evaluation

The clear relative abundance separation patterns motivated us to construct a confidence-based evaluation that provides additional information and improves the usability of our final deployed models. For both SVM and Random Forest models, the confidence evaluation uses both predicted probability and relative abundance thresholds to classify samples into low and high confidence groups. First, we identified samples with low prediction probability and low relative abundance for each non-NPCST V group, which represented <10% of samples where the machine learning models achieved only median accuracies of 56% and 67% for SVM and Random Forest, respectively. As illustrated in **Figure S17 & 18 A-F**, there was an unmistakably strong pattern between relative abundance and prediction probability. We established optimal probability and relative abundance thresholds for NPCST-specific genera to classify predictions as high or low confidence. For example, for NPCST I and its dominant genus *Moraxella*, predictions with SVM probability <0.46 and *Moraxella* relative abundance <52% were considered low confidence.

Through Youden's J statistic-guided selection of optimal prediction probability and relative abundance thresholds, we significantly improved prediction accuracy of above-threshold samples to >95% and >97% for SVM and Random Forest, respectively (**Table S7**). The low confidence group achieved 6.2-26.3% and 9.4-19.3% accuracy for SVM and Random Forest models, respectively. These results highlight the effectiveness of the confidence evaluation in identifying and flagging low-confidence predictions.

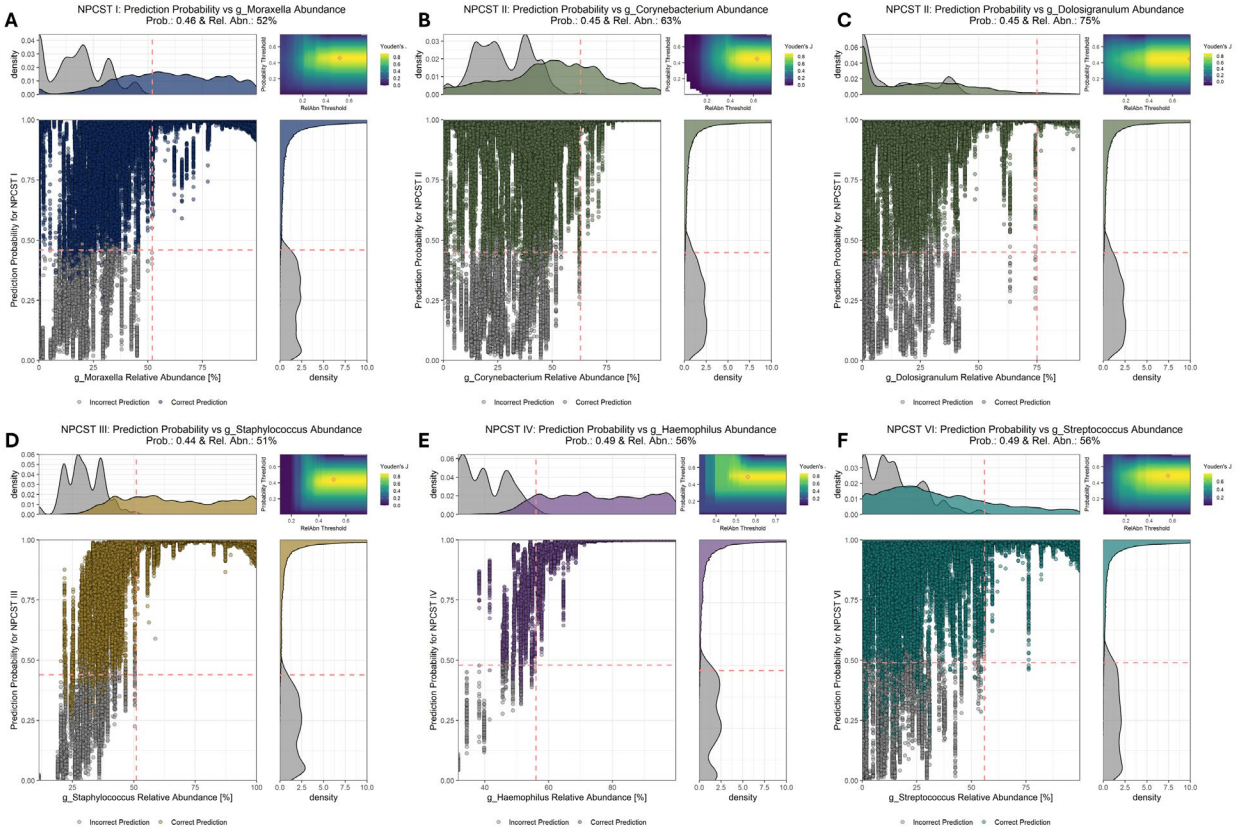

Figure S17. Evaluation of optimal SVM prediction probability and dominant genera relative abundance thresholds for NPCST classification using Youden's J statistic optimization. **A-F** display density plots of relative abundance (x-axis) versus SVM prediction probability (y-axis) for NPCST groups I, II (two dominated genera were evaluated separately), III, IV, and VI, respectively, using data from testing sets across 100 iterations of 5-fold cross-validation. Colored dots are correct predictions, whereas gray-colored dots represent incorrect predictions from SVM. Each panel's top-right heatmap illustrates Youden's J statistics calculated at 0.01 intervals for relative abundance and predicted probability combinations within the 0-0.75 range. The red circle indicates the optimal threshold combination with the highest Youden's J statistic; when multiple combinations yielded identical Youden's J values, the combination with the lowest relative abundance and probability thresholds was selected. This optimal combination is visualized as a red dot on the scatter plot, with corresponding threshold lines shown on the density plots to illustrate classification across all correctly and incorrectly predicted samples. This analysis was restricted to samples with both predicted probability and relative abundance < 0.75 to focus on challenging predictions where confidence classification is most critical. The right density plot displays only 0 to 10 density range for better visualization.

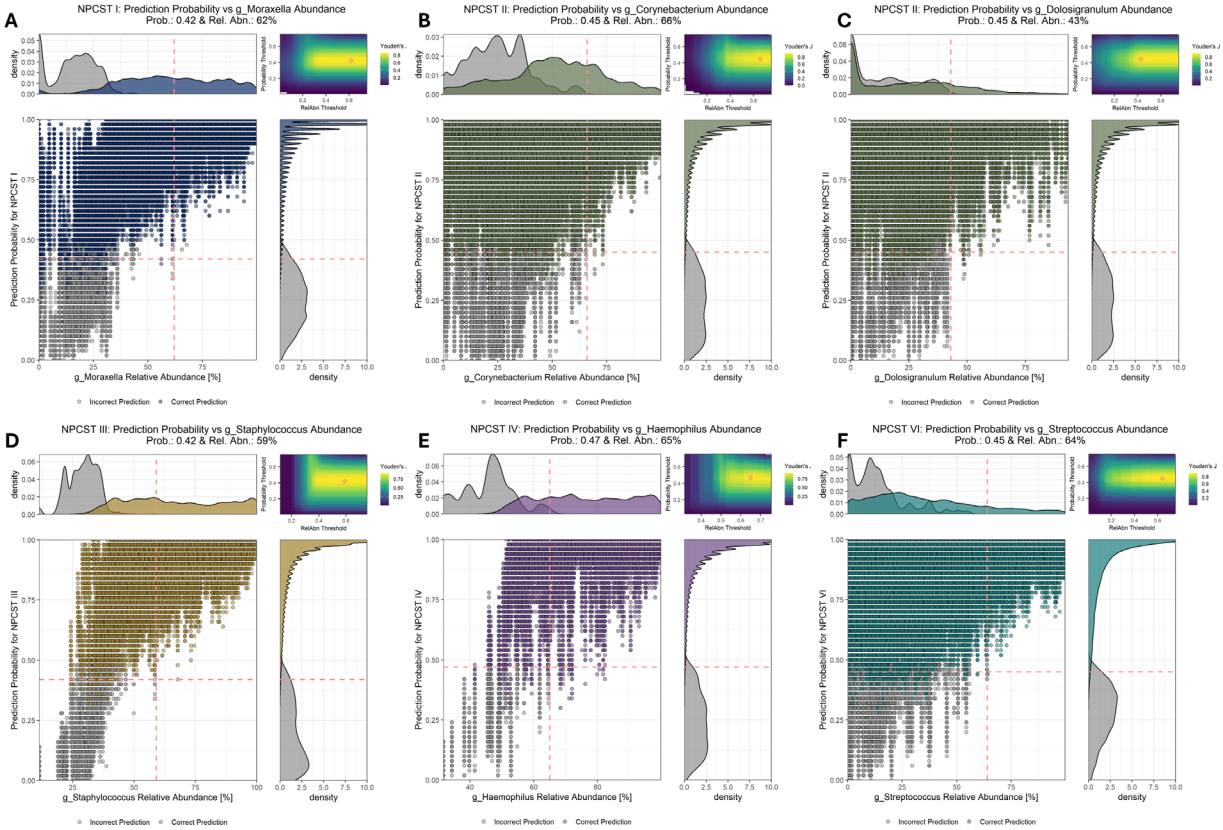

Figure S18. Evaluation of optimal Random Forest prediction probability and dominant genera relative abundance thresholds for NPCST classification using Youden's J statistic optimization. **A-F** display density plots of relative abundance (x-axis) versus SVM prediction probability (y-axis) for NPCST groups I, II, III, IV, and VI, respectively, using data from testing sets across 100 iterations of 5-fold cross-validation. Colored dots are correct predictions, whereas gray-colored dots represent incorrect predictions from SVM. Each panel's top-right heatmap illustrates Youden's J statistics calculated at 0.01 intervals for relative abundance and predicted probability combinations within the 0-0.75 range. The red circle indicates the optimal threshold combination with the highest Youden's J statistic; when multiple combinations yielded identical Youden's J values, the combination with the lowest relative abundance and probability thresholds was selected. This optimal combination is visualized as a red dot on the scatter plot, with corresponding threshold lines shown on the density plots to illustrate classification across all correctly and incorrectly predicted samples. This analysis was restricted to samples with both predicted probability and relative abundance < 0.75 to focus on challenging predictions where confidence classification is most critical. The right density plot displays only 0 to 10 density range for better visualization.

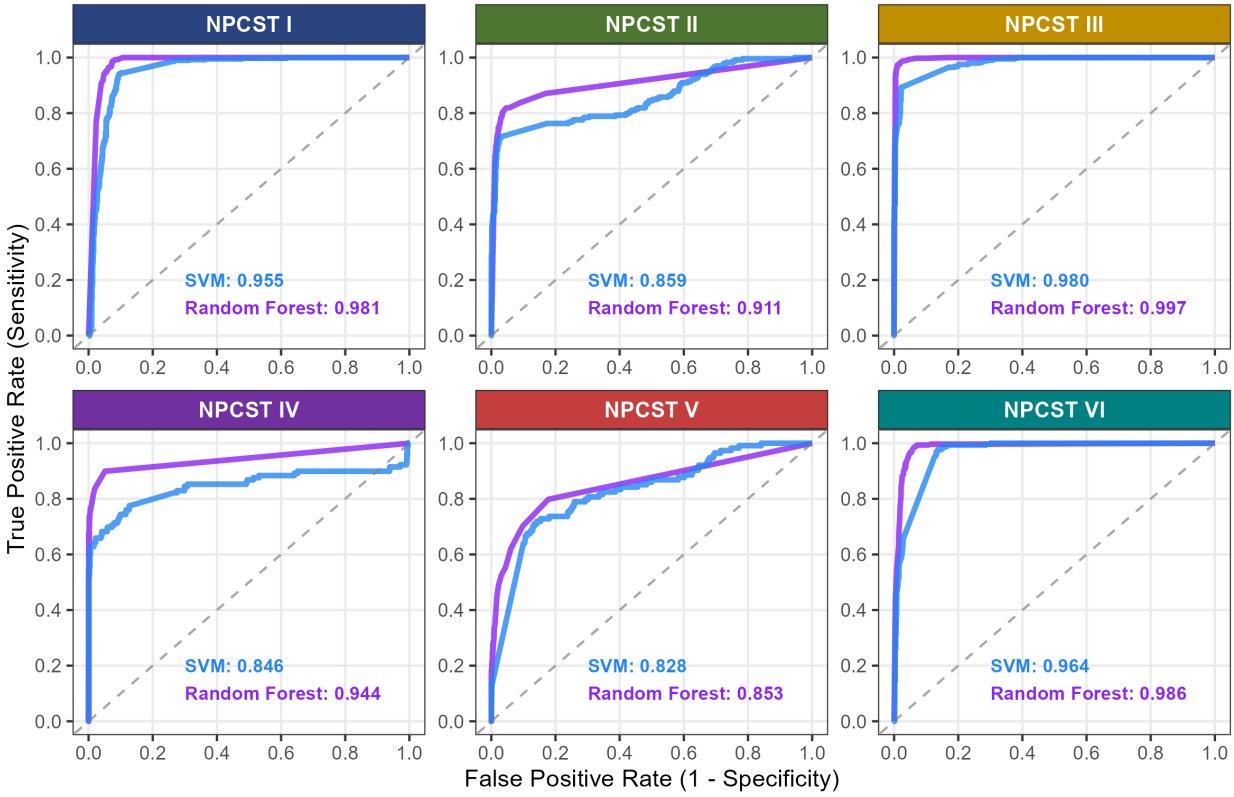

Figure S19. ROC curve analysis of external validation NPCST predictions. Ground truth assignments derived from hierarchical clustering analysis combining external validation studies with the original 28-study dataset (Bray-Curtis dissimilarity, Ward linkage; see **Methods** and **Supplementary File Methods**). Model performance is shown via ROC curves and AUC values for both SVM (blue) and Random Forest (purple) across all six NPCSTs.

Community-state-specific patterns in demographics, disease risk, and microbial diversity

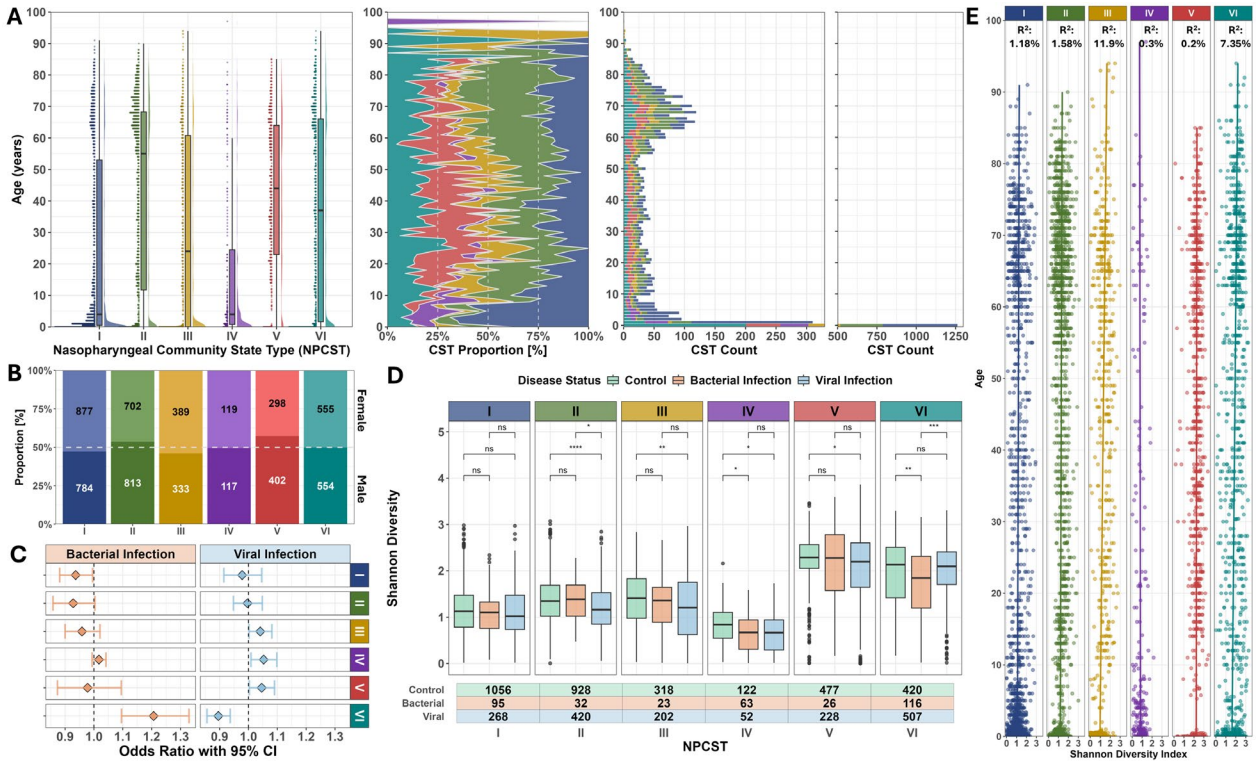

Figure S20. Host demographic and ecological characteristics of nasopharyngeal community-state types (NPCSTs). A. Age distribution across six NPCSTs displayed as raincloud plots combining density distributions (right), boxplots (center), and individual data points (left). The middle panel shows the proportion of each NPCST across age groups using 1-year increments. The two rightmost panels display the sample count for each NPCST at 1-year age increments. **B.** Sex distribution within each NPCST showing proportions of female (light shading) and male (dark shading) participants, with sample counts indicated within bars. Dashed white line indicates 50% proportion. **C.** Forest plot displaying odds ratios (OR) with 95% confidence intervals for bacterial and viral infections compared to controls across NPCSTs, derived from multinomial logistic regression stratified by BioProject with OR=1 indicated by vertical black dashed lines. ORs >1.0 indicate increased infection risk, while ORs <1.0 indicate decreased risk relative to controls. **D** Shannon diversity distributions comparing control, bacterial infection, and viral infection groups within each NPCST. Sample sizes are shown below each group with significance levels from pairwise Wilcoxon rank-sum tests with FDR correction indicated above with brackets (*p < 0.05, **p < 0.01, ***p < 0.001, ns = not significant). **E.** Relationship between Shannon diversity and age within each NPCST, with linear regression R^2 values displayed above each plot. Each point represents an individual

sample colored by NPCST membership. In all panels, NPCSTs are color-coded as follows: I (blue), II (green), III (yellow), IV (purple), V (red), and VI (teal).

Co-occurrence network

Figure S21A-B presents centrality rankings (closeness, betweenness, and degree) for the top-performing genera. These 44 top-ranked genera (of 72 total analyzed) exhibited two distinct functional patterns: 17 multi-hub genera (including *g_Acinetobacter*, *g_Anaerococcus*, and *g_Peptoniphilus*) consistently ranked highly across all three network metrics in both NPCST-specific and global networks, indicating their role as keystone anchors that maintain community structure regardless of compositional shifts. The remaining 27 specialized genera showed prominence in only 1-2 metrics within specific NPCST contexts. For instance, *g_Enhydrobacter*, a gram-negative bacterium, ranked first for degree centrality in NPCST VI and first for closeness centrality in NPCSTs I and II, indicating its function as a context-specific influencer rather than a universal community hub. This specialization pattern provides new avenues for investigating these understudied genera. Importantly, *Moraxella* from NPCST I consistently showed low network centrality rankings despite its numerical dominance, demonstrating that competitive dominance can paradoxically result in network isolation and limited co-existence capacity. **Figure S21C** displays the co-occurrence networks for NPCSTs V and VI, which exhibit more complex association patterns consistent with their higher microbial diversity.

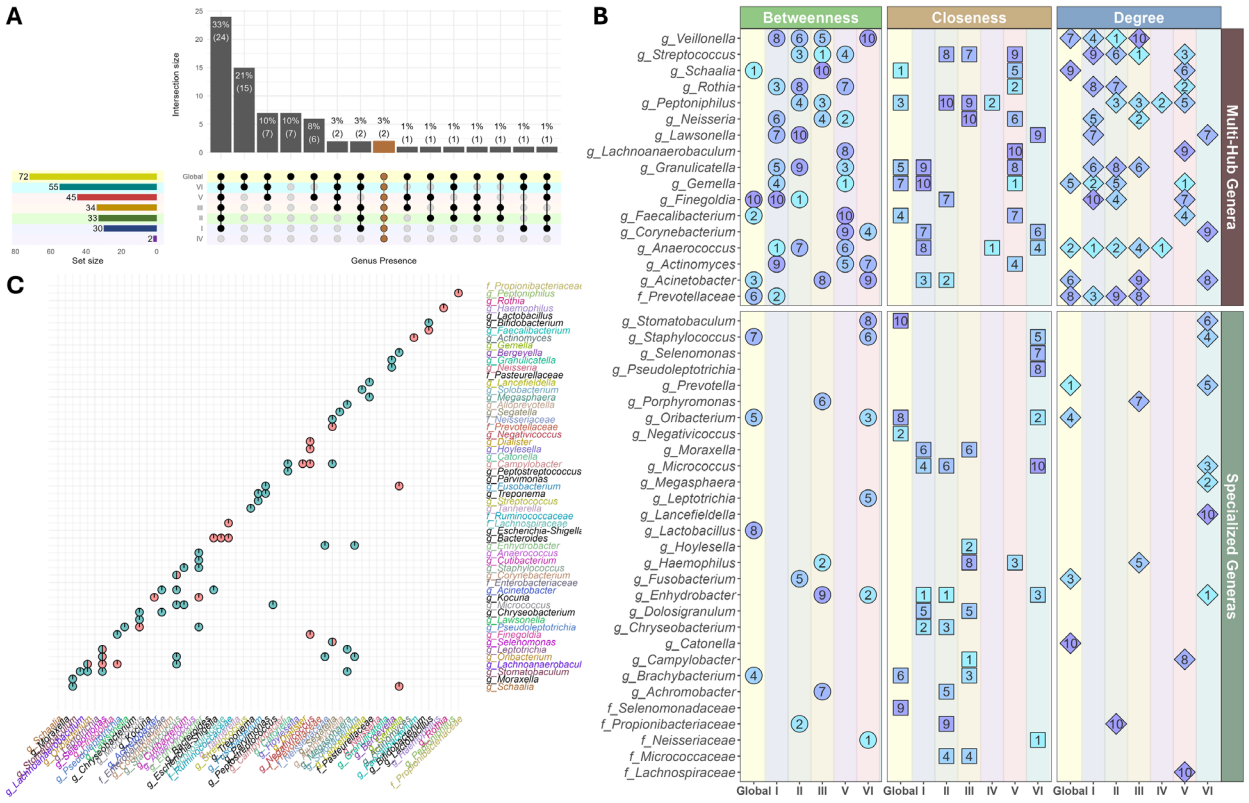

Figure S21. Network Centrality Analysis of Structurally Important Genera. A. UpSet plot showing the distribution of genera across global (orange dots and bar on upper marginal distribution) and NPCST-specific co-occurrence networks. Each column represents a unique combination of networks sharing specific genera, with the number and percentage of genera shown above each bar. The left panel indicates the total number of genera in each network (set size). The brown column highlights the, genera shared across all networks. Connected dots in the matrix indicate which networks contribute to each intersection. **B.** Centrality rankings (closeness, betweenness, and degree) for 44 top-ranked genera across the global and six NPCST-specific co-occurrence networks. Data point color indicates centrality ranking position (ranging from light blue=1 to purple=10), and shape indicates centrality type (circle=betweenness, square=closeness, and diamond=degree). Genera are stratified into two functional categories: multi-hub genera (n=17), representing taxa with consistently high centrality across multiple networks, and specialized genera (n=27), which exhibit prominent centrality within specific NPCST contexts but limited rankings across networks. **C.** Associations across networks, focusing on associations involving NPCSTs V and VI. Each pie chart represents microbial associations unique to NPCST V (red), unique to NPCST VI (teal), or shared between both NPCSTs (side by side color).

Nasopharyngeal microbiome health index

Stage 1. LONO cross-validation results: optimal lambda selection across taxonomic levels and prevalence thresholds

Leave-One-NPCST-Out (LONO) cross-validation successfully identified optimal regularization parameters across taxonomic ranks, with all-taxa and genus-level models consistently achieving superior performance across both disease classifications (Table S14). The all-taxa models achieved the highest AUC values (0.886 for all conditions, 0.890 for viral infections at 0% prevalence), followed by genus-level models (0.847 and 0.855, respectively), demonstrating robust discriminatory power. Species-level models ranked third with AUC values of 0.798-0.813, representing a performance decrease comparable to the all-taxa-to-genus drop (~0.04-0.05 AUC points). The remaining single-taxonomy models exhibited more substantial performance degradation, establishing a clear hierarchy: all-taxa > genus > species > family > order > class ≈ phylum, with class and phylum models showing equivalent poor performance (AUC ~0.64).

Stage 2: model performance from LONO and 10-fold cross-validation on the training datasets

Using optimal lambda parameters identified in Stage 1, we evaluated all-taxa and genus-level models through both LONO and 10-fold cross-validation (10 repeats) across five prevalence thresholds (**Table S12** and **Figure 6A**). LONO validation revealed consistent performance degradation with increasing prevalence thresholds: all-taxa all-conditions models achieved mean AUC of 0.877, 0.874, and 0.857 at 0%, 1%, and 5% thresholds respectively, with marked declines at 10% (0.826) and 20% (0.816). 10-fold cross-validation yielded systematically higher performance across all configurations, with corresponding AUC values of 0.907, 0.902, 0.889, 0.855, and 0.845, representing a consistent ~0.03 improvement over LONO estimates. Viral infection-specific models marginally outperformed all-conditions models by 0.008-0.016 AUC points across both validation strategies. Genus-level models maintained competitive performance, achieving mean AUC values 0.03-0.04 lower than all-taxa models, with 0.814 (LONO) and 0.847 (10-fold) at the 5% threshold for all-conditions classification.

Balanced accuracy metrics paralleled AUC trends across both validation strategies. LONO cross-validation for all-taxa all-conditions models yielded balanced accuracies of 0.792, 0.791, and 0.773 at 0%, 1%, and 5% prevalence thresholds, with notable decreases to 0.743 and 0.738 at 10% and 20% thresholds respectively. Viral infection-specific models demonstrated consistent but modest improvements, with balanced accuracies 0.012-0.018 higher than corresponding all-conditions models across all thresholds (**Table S12**).

10-fold cross-validation produced systematically elevated balanced accuracies, achieving 0.825, 0.821, and 0.804 for all-taxa all-conditions models at 0%, 1%, and 5% thresholds, representing improvements of around 0.03 over LONO estimates. The convergent performance patterns across both validation approaches support selection of the 5% prevalence threshold, which optimally balances discriminatory power with model generalizability by capturing community-level microbial signatures while filtering rare taxa present in <5% of samples.

Stage 3: final model evaluation using the full training data

Having established the 5% prevalence threshold, we trained four final models (all-taxa and genus-level for both all-conditions and viral infections) and evaluated NMHI distributions across disease categories. Our dataset encompassed diverse pathogen types: viral infections (e.g., SARS-CoV-2, rhinovirus, influenza), bacterial infections (e.g., meningococcal disease, tuberculosis, pneumococcal disease), and polymicrobial conditions of mixed etiology (e.g., otitis media and rhinosinusitis). Initial pairwise comparisons between healthy controls and individual disease states yielded significant differences for all conditions (Wilcoxon rank-sum test, FDR-adjusted $p < 0.001$). When aggregated by pathogen category, NMHI demonstrated strong discriminatory power with large effect sizes: viral infections (Cohen's $d = 1.88$), bacterial infections ($d = 1.58$), and mixed infections ($d = 1.50$), with all category-level comparisons remaining highly significant (FDR-adjusted $p < 0.0001$). These substantial effect sizes across diverse infectious etiologies validate NMHI's capacity to capture general health index signatures independent of specific pathogen type.

With the final models established, we evaluated non-zero coefficients across all four model configurations to identify key taxa driving NMHI predictions (**Figure S22-23**). Coefficients were classified as health-promoting (>0) or disease-associated (<0), revealing 99 health-promoting and 98 disease-associated taxa in the All-Taxa All-conditions model, 103/96 in the All-Taxa Viral Infection model, 37/31 in the Genus-Only All-conditions model, and 38/36 in the Genus-Only Viral Infection model. To focus on taxa with substantial predictive influence, we analyzed features with any taxa that contain absolute coefficients ≥0.5, identifying 34 health-promoting and 24 disease-associated taxa that demonstrate varying prevalence patterns across nasopharyngeal community structures (**Figure 6D**). Comparative abundance analysis of these key markers across control, all-disease, and viral infection groups revealed that few taxa exhibit complete absence or presence patterns between healthy and diseased states (**Figures S20-S21**). Even statistically significant markers such as g_*Cutibacterium* (health-promoting), g_*Moraxella* (health-promoting) and g_*Haemophilus influenzae* (disease-associated) showed overlapping abundance

distributions across groups, with both taxa present in substantial proportions of control and disease samples. These findings validate our composite index approach rather than reliance on individual microbial markers, as no single taxon provides definitive discrimination between healthy and diseased nasopharyngeal microbiomes.

Figure S22. Disease-associated taxa abundance across infection categories. Boxplots show the relative abundance distributions of disease-associated microbial taxa across Control, All Disease (combined bacterial and viral infections), and Viral Infection groups, with sample sizes (n) displayed below each category. The total number of samples for control, all disease and viral infections are 3,344, 2,091, and 1,725, respectively. Statistical

significance bars indicate Wilcoxon rank-sum test results comparing Control vs All Disease and Control vs Viral Infection (FDR-corrected; *p<0.05, **p<0.01, ***p<0.001), with colors representing Control (green), All Disease (blue), and Viral Infection (yellow).

Figure S23. Health-promoting taxa abundance across infection categories. Boxplots show the relative abundance distributions of health-promoting microbial taxa across Control, All Disease (combined bacterial and viral infections), and Viral Infection groups, with sample sizes (n) displayed below each category. The total number of samples for control, all disease and viral infections are 3,344, 2,091, and 1,725, respectively. Statistical significance bars indicate Wilcoxon rank-sum test results comparing Control vs All Disease

and Control vs Viral Infection (FDR-corrected; *p<0.05, **p<0.01, ***p<0.001), with colors representing Control (green), All Disease (blue), and Viral Infection (yellow).

Stage 4: NMHI threshold evaluation

Once NMHI scores were generated across all samples, we proceeded with threshold optimization to maximize classification accuracy. While zero represents the theoretical neutral point, it does not necessarily provide optimal discriminatory performance between healthy and disease states. Therefore, we determined optimal thresholds for the complete training dataset by maximizing balanced accuracy across threshold values ranging from -5 to 5 (**Figure 6E**). Given the compositional distinctiveness of NPCSTs, we additionally evaluated NPCST-specific thresholds to account for community-structure variations in optimal classification boundaries. The analysis revealed strong clinical differentiation between healthy controls and disease groups, with Cohen's d values ranging from 1.661 to 2.012 across individual NPCSTs and 1.887 for the combined dataset (**Figure 6E**). These large effect sizes demonstrate robust separation between healthy and diseased populations, with all Wilcoxon rank-sum comparisons achieving statistical significance (FDR-adjusted $p < 0.001$).

Stage 5: external validation

With final model coefficients and optimal thresholds established, we evaluated NMHI performance on independent validation datasets excluded from all prior training stages (**Figure 6F**). The external validation cohort was dominated by SARS-CoV-2 cases, reflecting continued research focus following the 2019 pandemic, with disease severity ranging from standard qPCR-confirmed SARS-CoV-2 to critical cases requiring ICU admission and mechanical ventilation. Additional validation samples included limited numbers of lower respiratory tract infections (LTRI, n=5), critically ill SARS-CoV-2-negative patients (n=31), and symptomatic individuals with suspected but confirmed-negative SARS-CoV-2 (n=15). External validation maintained strong discriminatory performance, with Cohen's d values ranging from 1.598-2.368 across all diagnostic categories when compared to external healthy controls, indicating large effect sizes and robust separation between healthy and diseased populations. Notably, mechanically ventilated SARS-CoV-2 patients showed the lowest effect size, potentially due to procedural artifacts affecting the nasopharyngeal microbiome during or after intubation. Nevertheless, all disease categories demonstrated statistically significant distributional differences from healthy populations, confirming NMHI's broad applicability across diverse pathological conditions. AUC analysis further validated model performance, achieving 0.922 for the combined dataset and NPCST-specific values ranging from 0.848-0.953, with NPCST classifications determined using the random forest model developed in earlier stages to ensure reproducible workflow

implementation (**Figure 6G**). NPCST V was excluded from analysis as all samples belonged to the disease group, precluding meaningful AUC calculation.

These comprehensive external validation results demonstrate NMHI's robust generalizability across diverse clinical populations and nasopharyngeal community structures, confirming its utility as a reliable monitoring tool for tracking disease progression and healthy nasopharyngeal microbial composition. The consistent discriminatory performance across independent datasets, varying disease severities, and distinct community types validates NMHI's potential for clinical implementation as a standardized biomarker for respiratory health assessment.

PRISMA checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Introduction
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Introduction
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods “Study screening and metadata evaluation“ & Figure 1B
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods “Study screening and metadata evaluation“
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Methods “Study screening and metadata evaluation“
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods “Study screening and metadata evaluation“
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods “16S data processing”
Data items	10a	List and define all outcomes for which data were sought. Specify	Methods

Section and Topic	Item #	Checklist item	Location where item is reported
		whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	& Table 1
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Table 1
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Methods & Discussion (Limitations)
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Methods "Statistical Methods"
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods "16S data processing"
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods "16S data processing" & "Nasopharyngeal background decontamination protocol"
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Methods (multiple sections)
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods (multiple sections)
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Methods (multiple sections) & Supplementary File
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Methods (multiple sections) & Supplementary File
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Methods
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	N/A
RESULTS			
Study	16a	Describe the results of the search and selection process, from the	Methods "Study

Section and Topic	Item #	Checklist item	Location where item is reported
selection		number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	screening and metadata evaluation"
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	N/A
Study characteristics	17	Cite each included study and present its characteristics.	Table 1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	N/A
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	N/A
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results (multiple sections) & Supplementary File
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Results (multiple sections) & Supplementary File
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results (multiple sections) & Supplementary File
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Results (multiple sections) & Supplementary File
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	N/A
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Results (multiple sections on validations) & Supplementary File on validations
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion
	23b	Discuss any limitations of the evidence included in the review.	Discussion
	23c	Discuss any limitations of the review processes used.	Discussion
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion
OTHER INFORMATION			
Registration	24a	Provide registration information for the review, including register name	N/A

Section and Topic	Item #	Checklist item	Location where item is reported
and protocol		and registration number, or state that the review was not registered.	
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	N/A
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	N/A
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding statement
Competing interests	26	Declare any competing interests of review authors.	Conflict of Interest
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data avalibility

Reference

1. Anthropic. Claude. (2022).
2. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *Bmc Biol* **12**, 87 (2014).
3. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol* **27**, 105–117 (2019).
4. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2021).
5. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013).
6. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
7. Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).

8. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581–583 (2016).
9. Pascoal, F., Branco, P., Torgo, L., Costa, R. & Magalhães, C. Definition of the microbial rare biosphere through unsupervised machine learning. *Commun Biol* **8**, 544 (2025).
10. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
11. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, (2010).
12. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. (2024).
13. Kuhn & Max. Building Predictive Models in R Using the caret Package. *J Stat Softw* **28**, 1–26 (2008).
14. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets,. *Ieee T Vis Comput Gr* **20**, 1983–1992 (2014).
15. Csárdi, G. *et al. Igraph: Network Analysis and Visualization in R*. (2025). doi:10.5281/zenodo.7682609.
16. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *Bmc Bioinformatics* **12**, 77 (2011).
17. Chang, D. *et al.* Gut Microbiome Wellness Index 2 enhances health status prediction from gut microbiome taxonomic profiles. *Nat Commun* **15**, 7447 (2024).