Research Square

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

Pretreatment gut microbiome signatures in hospitalized patients with COPD: sequence-level characterization and cross-cohort validation

Yuejiao Sun
The Affiliated Hospital of Jiaxing University (The First Hospital of Jiaxing)
Jing Feng
Zhengzhou YIHE Hospital
Xixi Gao
The Affiliated Hospital of Jiaxing University (The First Hospital of Jiaxing)
Chaoping Zhu
The Affiliated Hospital of Jiaxing University (The First Hospital of Jiaxing)
Yan Liu
The First Affiliated Hospital of Zhejiang Chinese Medical University (Zhejiang Provincial Hospital of Chinese Medicine)
Yue Zhou
Zhejiang Gongshang University
Xiaolong Ma
maxiaolong068210@163.com
The Affiliated Hospital of Jiaxing University (The First Hospital of Jiaxing)

Research Article

Keywords: chronic obstructive pulmonary disease, gut microbiome, 16S rRNA gene sequencing, case-control discrimination, sequence-level bacterial features

Posted Date: September 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10738750/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Chronic obstructive pulmonary disease (COPD) has important extrapulmonary manifestations, but whether pretreatment intestinal bacterial profiles carry clinically relevant information in hospitalized patients remains unclear.

Methods

This cross-sectional study included 30 hospitalized patients with COPD and 30 patients hospitalized with other respiratory diseases. Fresh stool was collected within 24 h of admission and before antibiotics or systemic glucocorticoids were administered. Bacterial 16S rRNA gene V3–V4 sequencing was used to characterize community organization, identify exact-sequence associations and develop a classifier under repeated nested cross-validation. A complementary genome-level classifier was developed and evaluated in separately recruited cohorts from a public metagenomic study.

Results

Patients with COPD showed higher Simpson and inverse Simpson diversity (BH q = 0.0175 and 0.005, respectively) and altered Bray–Curtis community composition (PERMANOVA R^2 = 0.0550, p = 0.001). Among 56 directionally concordant exact-sequence candidates, 49 were supported by both technical-adjustment and centered-log-ratio analyses, and the 12 highest-ranked sequences retained their associations after non-chimeric reconstruction. The local exact-sequence classifier achieved a repeated nested cross-validated ROC-AUC of 0.8891 (95% CI, 0.8137–0.9510). In the public evaluation cohort, the complementary genome-level classifier achieved a ROC-AUC of 0.9205 (95% CI, 0.8182–0.9943), with 93.75% sensitivity and 72.73% specificity.

Conclusions

Pretreatment stool bacterial profiles captured a distinct COPD-associated intestinal phenotype and discriminated COPD from other respiratory diseases in hospitalized patients. The prioritized 12-sequence signature and complementary public metagenomic model provide convergent evidence for the clinical potential of stool microbial features as adjunctive markers of COPD.

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous respiratory disorder in which persistent airflow limitation coexists with systemic inflammation, metabolic disturbance, skeletal-muscle dysfunction and multimorbidity. These extrapulmonary manifestations contribute substantially to disease burden and show that COPD biology extends beyond the airways. The gut–lung axis provides a framework for this systemic dimension. Intestinal microorganisms and their metabolites influence mucosal and circulating immune responses. Conversely, smoking, diet, hypoxemia, inflammation and medication exposure can alter the intestinal ecosystem

[1,2]. The fecal microbiome may therefore capture clinically relevant features of both COPD biology and host state.

Human studies using fecal 16S rRNA gene sequencing, shotgun metagenomics and metabolomics have identified differences in diversity, bacterial taxa, microbial functions and metabolites in COPD [3–6]. Bowerman et al. described community, functional and metabolite changes in clinically stable COPD and observed related taxonomic signals in a second cohort [5]. Other studies have extended severity-stratified profiling and case-control classification [3,4,6]. These findings support a measurable intestinal microbial phenotype, although diversity patterns, implicated taxa and clinical associations vary across cohorts.

The clinical setting is likely to contribute to this heterogeneity. A microbiome measured during hospitalization can integrate chronic airflow limitation with acute symptom burden, infection-associated inflammation, multimorbidity and treatment exposure. Hospitalized patients with COPD are therefore clinically important but biologically distinct from untreated stable outpatients. Relatively few microbiome studies have linked sequence data from this setting to participant-level treatment, inflammatory and pulmonary records, although such linkage is essential for interpreting a case-control signal.

Analytical resolution creates a second challenge. Genus- or species-level summaries can conceal sequence heterogeneity, whereas assignments depend on the amplified region, reference database and reporting threshold. Sample processing, sequencing depth, primer choice and bioinformatic parameters also influence feature recovery [7–10]. In high-dimensional sparse data, preprocessing, technical adjustment, feature selection and model tuning must occur within resampling folds to limit optimistic performance estimates [11,12]. Host-variable imbalance and batch structure also require explicit consideration through fold-restricted preprocessing and complementary sensitivity analyses [13,14].

We therefore examined stool bacterial community organization and sequence-level features in hospitalized patients with COPD. The local 16S analysis integrated clinical characterization, community ecology, differential association, compositional and technical analyses, non-chimeric reconstruction and nested internal classification. We also assessed genome-level case-control classification using public metagenomic data.

Methods

Study design and analytical framework

The study comprised a local cross-sectional 16S rRNA gene analysis and a complementary public metagenomic classification analysis. The local analysis characterized stool bacterial community structure and exact-sequence associations in hospitalized patients with COPD, with repeated nested cross-validation used to evaluate internal case-control classification. Technical, compositional and non-chimeric analyses examined the stability of ecological associations, candidate features and classifier performance. Separately recruited public metagenomic cohorts were used to develop and evaluate genome-level case-control classification.

Local cohort, eligibility and clinical characterization

The local cohort comprised 30 patients hospitalized with COPD between October 2024 and April 2025 and 30 patients hospitalized with non-COPD respiratory diseases. COPD status was confirmed by a pulmonologist from

the longitudinal medical record, including previous or index-admission post-bronchodilator spirometry when available. For patients without contemporaneous spirometry, eligibility required an established specialist diagnosis supported by previous pulmonary-function records and the index discharge diagnosis. The index hospitalization was prompted by acute respiratory symptoms requiring inpatient evaluation or treatment. Spirometry was available for 19 patients with COPD and 17 controls and was summarized as recorded.

Inclusion criteria were: (1) age ≥ 18 years and hospitalization for COPD or another respiratory disease; (2) pulmonologist-confirmed COPD, or a documented non-COPD respiratory diagnosis with no history of COPD for controls; (3) collection of fresh stool within 24 h of admission and before the first inpatient dose of antibiotics or systemic glucocorticoids; and (4) complete paired-end 16S rRNA gene data with one-to-one linkage to the clinical record. Exclusion criteria were: (1) antibiotic or systemic glucocorticoid exposure before stool collection during the index admission; (2) inflammatory bowel disease, acute infectious gastroenteritis, colorectal malignancy, major intestinal surgery or another condition expected to substantially alter the intestinal microbiota; (3) uncertain case-control classification or coexisting COPD in a control participant; and (4) an inadequate stool specimen, incomplete paired-end sequence data or unresolved clinical-to-library linkage.

The documented control diagnoses comprised pneumonia (n = 12), pulmonary malignancy or a pulmonary mass (n = 9), pulmonary embolism (n = 5), bronchiectasis with infection (n = 2) and pleural effusion (n = 2).

Clinical characterization included demographic and smoking variables, pretreatment oxygenation, inflammatory and hematologic indices, pulmonary function, length of stay, subsequent inpatient treatment and CT findings. BMI was recalculated from source height and weight values when available. The available sample size is reported for each characteristic.

The study protocol was reviewed and approved by the Ethics Committee of Jiaxing First Hospital (approval no. 2024-LY-418) and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants or their legally authorized representatives before enrollment.

Stool sampling, sequencing and bacterial feature construction

Fresh stool was collected from all participants within 24 h of admission and before the first inpatient administration of antibiotics or systemic glucocorticoids. Samples were collected in sterile, nuclease-free containers without contact with urine or toilet water, transported on ice, aliquoted within 2 h and stored at − 80°C until DNA extraction [7,15–17]. Genomic DNA was extracted from 200 mg of stool using the QIAamp DNA Stool Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The V3–V4 hypervariable region of the bacterial 16S rRNA gene was amplified with primers 341F (5'-CCTACGGGNGGCWGCAG-3') and 806R (5'-GGACTACHVGGGTATCTAAT-3'), yielding an expected amplicon of approximately 466 bp. Each 25-μl reaction contained 12.5 μl of 2× high-fidelity PCR master mix, 0.5 μM of each primer and approximately 10 ng of template DNA. Thermal cycling comprised 3 min at 95°C; 30 cycles of 30 s at 95°C, 30 s at 55°C and 45 s at 72°C; and a final extension for 10 min at 72°C. Case and control specimens were interleaved across extraction, PCR and indexing plates, and each batch included an extraction blank, a no-template PCR control and a commercial mock bacterial community. Amplicons were gel-purified, quantified with QuantiFluor, pooled in equimolar amounts, indexed and sequenced on an Illumina MiSeq instrument using a 2 × 250-bp paired-end protocol by Guangzhou Omicsmaster Biotechnology Co., Ltd.

Paired reads underwent integrity checking, quality-aware merging, primer removal and filtering by workflow-specified length and ambiguity criteria. Identical clean sequences were dereplicated within each sample. Features with at least 10 cohort-wide counts and detection in at least two samples were retained as observed exact-sequence features (ESVs). This term denotes the exact sequences observed after quality filtering. Chloroplast, mitochondrial, archaeal, eukaryotic and other non-bacterial assignments were removed before ecological, association and classification analyses.

As a sequence-processing sensitivity analysis, bacterial sequences underwent VSEARCH UNOISE abundance-skew denoising followed by UCHIME3 de novo chimera removal [18–20]. Ecological associations, candidate effects and internal classification were recomputed on the non-chimeric centroid matrix. Processing parameters and the bacterial 97% identity and genus-level sensitivity representations are provided in the Supplementary Methods.

Community ecology and sequence-level association analysis

Observed richness, Shannon diversity, Simpson diversity, inverse Simpson diversity and Pielou evenness were compared using two-sided label-permutation tests with 999 permutations. Benjamini–Hochberg correction was applied across the five alpha-diversity indices. Bray–Curtis dissimilarity was calculated from total-sum-scaled bacterial relative-abundance profiles and visualized by principal coordinates analysis. Group separation was assessed by PERMANOVA with 999 permutations, and within-group dispersion by PERMDISP with 9,999 permutations. Extended ordination, distance, rarefaction and rank-abundance analyses are described in the Supplementary Methods.

Bacterial counts were converted to within-sample relative abundances for feature-level analysis. Group differences were tested using two-sided Mann–Whitney U tests with Benjamini–Hochberg false-discovery-rate (FDR) correction. Log_2 fold change was calculated from group mean relative abundances after adding 1×10^{-8} to each mean, and Cliff's delta quantified effect size. A two-part analysis separately tested detection prevalence by Fisher's exact test and positive-sample abundance by Mann–Whitney U test.

Candidate prioritization and taxonomic characterization

Candidates first required cohort-wide FDR $q < 0.05$ and concordance between the differential-abundance direction and the mean coefficient direction from training-fold L1-regularized models. Eligible candidates were ranked by a composite score combining statistical strength, fold change, fold inclusion frequency and coefficient magnitude. The final panel comprised the 12 highest-ranked candidates that mapped to distinct non-chimeric centroids and met technical and centered-log-ratio (CLR) support at candidate-family $q < 0.10$.

Technical support required the original direction to persist after reduced-QC common-support restriction and candidate-level ridge residualization. CLR support required directional and statistical consistency across unadjusted, common-support and reduced-QC-residualized analyses. Candidate-family FDR $q < 0.10$ defined support in both frameworks. Exact sequences were globally aligned against Ribosomal Database Project (RDP) trainset 18 references using VSEARCH v2.31.0 [18]. Genus names were reported when identity and query coverage were both at least 95% and no near-best genus conflict was present; otherwise, a supported higher taxonomic level was used.

Clinical–microbiome relationships within the COPD cohort

Clinical variables were summarized as mean ± SD, median (IQR) or n (%), with the available number reported for each measure. Within the COPD cohort, ten clinical variables were evaluated against inverse Simpson diversity, observed feature richness, Bray–Curtis distance to the COPD centroid and a direction-aligned 12-candidate score. The score was the mean within-COPD z-score of $\log_{10}$ relative abundance across the 12 candidates after reversing the control-enriched candidate. Spearman rank correlations were calculated with binary variables coded 0/1, and Benjamini–Hochberg correction was applied across all 40 tests.

Nested internal case-control classification

The local classifier was a linear support vector machine evaluated by stratified five-fold cross-validation repeated ten times. Each repeat generated complete out-of-fold predictions for all 60 participants. ROC-AUC was the primary performance measure; precision–recall AUC and balanced accuracy were secondary measures. Sensitivity, specificity and balanced accuracy used the linear SVM decision threshold of zero. A 10,000-resample stratified participant bootstrap preserved repeat-level predictions within participants and provided the ROC-AUC confidence interval.

Within each outer training fold, ridge regression residualized bacterial features against six non-redundant sequencing-quality and depth variables. These were input read pairs, mean Read 1 quality, mean Read 2 quality, Q30 fraction, merge rate and clean-tag retention rate. Q30 was defined as the fraction of bases with a Phred score ≥ 30. Residual-variance feature selection, scaling, inner-loop tuning and model fitting then used training data only. Mean Read 2 quality alone and a broader 27-variable matrix were evaluated as adjustment sensitivities.

Sensitivity analyses used the non-chimeric UNOISE matrix, 1,000 label permutations and abundance and prevalence filters fitted within each outer training fold. We also repeated the analysis in the 51 participants within the reduced-QC common-support interval. Incremental models compared reduced-QC variables, residualized microbial features and their combination using the same linear SVM family and identical outer splits. Potential run-, lane- and tile-associated effects were examined using metadata from 1,000 headers sampled from each FASTQ file. Full model specifications and results are provided in the Supplementary Methods and Supplementary Tables 5–8.

Public metagenomic classification

Public genome-level abundance data from Bowerman et al. [5] comprised a development cohort of 28 patients with COPD and 29 controls and a separately recruited evaluation cohort of 16 patients with COPD and 22 controls. The cohorts shared 1,668 genome-level features. Genomes detected in at least 10% of development samples were retained before model fitting.

A pseudocount of 0.5 was added before sample-wise CLR transformation. Feature selection, standardization and class-balanced elastic-net logistic regression were tuned by repeated stratified five-fold cross-validation in the development cohort. The selected model was refitted to all development samples and applied once to the evaluation cohort. Evaluation ROC-AUC was the primary endpoint; its confidence interval was estimated using 2,000 stratified bootstrap resamples. Secondary endpoints were precision–recall AUC, balanced accuracy,

sensitivity, specificity, Brier score and descriptive calibration. Within-sample ranks, alternative pseudocounts, different feature-set sizes and leave-one-participant-out analyses assessed stability.

Statistical analysis and reporting

Baseline continuous variables are reported as mean ± SD or median (IQR) and were compared using Welch's t test or the Mann–Whitney U test, respectively; categorical variables were compared using Fisher's exact test. Baseline p values were descriptive and were not adjusted for multiple comparisons. All tests were two-sided with a significance threshold of 0.05. Benjamini–Hochberg correction was applied within each analysis family, and empirical permutation p values used the plus-one correction. Analyses used Python 3.9.6 with NumPy 2.0.2, pandas 2.3.3, SciPy 1.13.1 and scikit-learn 1.6.1. Reporting followed the STORMS guideline for human microbiome research [21]; the completed checklist is supplied separately.

Results

Cohort characteristics and sequencing yield

The analysis included 30 hospitalized patients with COPD and 30 non-COPD respiratory inpatients (Fig. 1a). Clinical records were linked to all 60 sequencing libraries. Mean age (74.5 ± 6.6 versus 71.5 ± 6.5 years; p = 0.076), BMI (24.1 ± 2.6 versus 23.4 ± 2.7 kg/m²; p = 0.339) and documented ever-smoking history (33.3% versus 26.7%; p = 0.779) were similar. Men accounted for 29 of 30 patients with COPD (96.7%) and 13 of 30 controls (43.3%; p < 0.001). Pretreatment oxygen saturation was higher in COPD (95.7 ± 2.3% versus 93.5 ± 3.6%; p = 0.006), while IL-6 was lower (median, 6.04 versus 17.41 pg/mL; p = 0.043). Among participants with available spirometry (19 with COPD and 17 controls), FEV_1 (1.12 ± 0.34 versus 1.85 ± 0.48 L) and FEV_1/FVC (48.6 ± 11.1% versus 80.5 ± 4.9%) were lower in COPD (both p < 0.001; Table 1).

Table 1
Baseline clinical characteristics of hospitalized patients with COPD and non-COPD respiratory inpatients.

Characteristic	COPD (n = 30)	Non-COPD respiratory inpatients (n = 30)	P value
Demographic and exposure characteristics			
Participants	30	30	—
Age, years	74.5 ± 6.6	71.5 ± 6.5	0.076
Male sex	29 (96.7%)	13 (43.3%)	< 0.001
BMI, kg/m²	24.1 ± 2.6 (n = 26)	23.4 ± 2.7 (n = 27)	0.339
Documented ever-smoking history	10 (33.3%)	8 (26.7%)	0.779
Pack-years among ever-smokers	52.5 (50.0–95.0) (n = 10)	37.5 (27.5–52.5) (n = 8)	0.067
Pretreatment clinical and inflammatory characteristics			
Peripheral oxygen saturation, %	95.7 ± 2.3	93.5 ± 3.6	0.006
C-reactive protein, mg/L	12.7 (3.8–56.2)	24.7 (14.2–56.6)	0.178
White blood cell count, ×10⁹/L	7.77 (5.61–9.61)	7.94 (6.50–8.95)	0.912
Neutrophil count, ×10⁹/L	5.65 (3.73–7.10)	4.75 (3.90–5.70)	0.137
Lymphocyte count, ×10⁹/L	1.20 (0.72–1.50)	1.33 (0.92–1.66)	0.246
Interleukin-6, pg/mL	6.04 (1.50–55.24) (n = 29)	17.41 (10.87–46.94) (n = 27)	0.043
Pulmonary physiology			
FVC, L	2.35 ± 0.55 (n = 19)	2.30 ± 0.55 (n = 17)	0.761
FEV_1, L	1.12 ± 0.34 (n = 19)	1.85 ± 0.48 (n = 17)	< 0.001
FEV_1/FVC, %	48.6 ± 11.1 (n = 19)	80.5 ± 4.9 (n = 17)	< 0.001
Hospitalization course			
Length of hospital stay, days	7.0 (5.0–8.0)	7.5 (6.0–9.8)	0.088

Values are mean ± SD, median (IQR) or n (%); available sample sizes are shown when fewer than 30. P values were descriptive, were not adjusted for multiple comparisons, and were calculated using Welch's t test, the Mann–Whitney U test or Fisher's exact test, as appropriate. Stool was collected within 24 h of admission and before the first administration of antibiotics or systemic glucocorticoids.

The 60 samples generated 7,430,449 input read pairs, of which 98.62% were merged. Primer removal and quality filtering retained 6,899,575 clean sequences, representing 92.86% of input read pairs. Abundance and prevalence filtering retained 2,361,872 counts (34.2% of clean sequences) across 25,980 features. Taxonomic filtering removed 433 non-bacterial features. The final bacterial matrix contained 25,547 observed exact-sequence features and 2,328,541 counts, equivalent to 33.7% of clean sequences (Fig. 1b,c,e,f and Supplementary Table 1).

Within the COPD cohort, C-reactive protein showed the strongest exploratory association with Bray–Curtis distance to the COPD centroid (Spearman ρ = 0.403, p = 0.027). The broader clinical–microbiome correlation pattern was weak after multiplicity correction (minimum q = 0.607; Supplementary Fig. 7f).

Stool bacterial communities were reorganized in hospitalized patients with COPD

Simpson and inverse Simpson diversity were higher in patients with COPD (BH q = 0.0175 and 0.005, respectively), whereas Shannon diversity, observed richness and Pielou evenness were not significant after correction (all BH q = 0.215; Fig. 2a–c and Supplementary Table 2). Median inverse Simpson diversity was 42.83 in COPD and 26.26 in controls. Bray–Curtis dissimilarity calculated from total-sum-scaled relative-abundance profiles showed a modest group-associated difference. Case–control grouping explained 5.50% of compositional variation (PERMANOVA R^2 = 0.0550, p = 0.001), while dispersion was similar (PERMDISP F = 0.0254, p = 0.8911; Fig. 2d,g).

The ecological findings persisted in the reduced-QC common-support subset of 21 patients and 30 controls. Inverse Simpson diversity remained higher in COPD (median, 41.73 versus 26.26; BH q = 0.010). Bray–Curtis separation was similar (PERMANOVA R^2 = 0.0586, p = 0.001), and dispersion remained comparable (PERMDISP p = 0.8686; Supplementary Table 2).

Non-chimeric UNOISE reconstruction retained the inverse Simpson association and produced a similar group effect on total-sum-scaled Bray–Curtis composition (R^2 = 0.0621, p = 0.001; PERMDISP p = 0.8519). Together, these results localized the group difference to community dominance and configuration (Supplementary Fig. 2 and Supplementary Table 2).

A prioritized sequence-level candidate panel characterized hospitalized patients with COPD

Differential analysis identified 171 FDR-significant bacterial features, including 137 enriched in COPD and 34 enriched in controls (Fig. 3a). Fifty-six candidates also showed concordant differential-abundance and L1-model directions. Technical adjustment supported 50 candidates, CLR analysis supported 52 and both analyses supported 49 at candidate-family q < 0.10 (Fig. 3b; Supplementary Tables 3 and 4). Composite ranking identified the 12 highest-ranked candidates that satisfied the technical, CLR and distinct non-chimeric-centroid criteria (Fig. 3c,d and Table 2).

Table 2
Prioritized bacterial exact-sequence candidate panel associated with hospitalized patients with COPD.

Provisional annotation	Feature ID	Enriched group	COPD prevalence	Control prevalence	log_2 fold change	BH q value	Cliff's delta
Saccharibacteria genera incertae sedis	ESV000599	COPD	80.0%	16.7%	5.49	0.000393	0.73
Fusobacterium	ESV000208	COPD	86.7%	16.7%	5.62	0.000444	0.73
Fusobacterium	ESV000269	COPD	90.0%	33.3%	4.92	0.000592	0.74
Unclassified Bacteroidetes-lineage sequence	ESV000624	COPD	90.0%	26.7%	4.50	0.000462	0.74
Catonella	ESV001314	COPD	86.7%	20.0%	3.50	0.000296	0.76
Unclassified Firmicutes-lineage sequence	ESV000521	COPD	93.3%	43.3%	3.94	0.000444	0.77
Prevotella	ESV000072	COPD	93.3%	46.7%	3.47	0.000651	0.74
Filifactor	ESV000089	COPD	100.0%	76.7%	3.30	0.000296	0.80
Oribacterium	ESV000133	COPD	100.0%	76.7%	3.34	0.000296	0.80
Prevotella	ESV000201	COPD	86.7%	30.0%	4.14	0.00235	0.68
Neisseria	ESV000187	Control	10.0%	66.7%	−5.36	0.0103	−0.55
Treponema	ESV001689	COPD	63.3%	13.3%	5.05	0.00793	0.57

Candidates first required cohort-wide FDR q < 0.05 and concordant L1 direction. The panel comprised the 12 highest-ranked candidates that mapped to distinct non-chimeric centroids and met technical and CLR support at candidate-family q < 0.10. Eleven candidates also met q < 0.05 in both support frameworks, and all 12 retained their association direction with q < 0.05 after denoised reanalysis. Detailed abundance, taxonomic alignment and sensitivity statistics are provided in Supplementary Tables 11 and 12.

The COPD-enriched panel included *Fusobacterium*-, *Prevotella*-, *Catonella*-, *Filifactor*-, *Oribacterium*- and *Treponema*-associated sequences; a *Neisseria*-associated sequence was enriched in controls. Most candidates differed in detection prevalence, and some also differed in positive-sample abundance (Fig. 3e,f). Ten candidates met the genus-level reporting rule, whereas two retained higher-level labels (Fig. 3g).

UNOISE/UCHIME processing yielded 16,656 non-chimeric centroids. Of the 56 candidates, 52 mapped to non-chimeric centroids and 48 retained their original direction with FDR significance in the denoised reanalysis. All 12 prioritized candidates mapped to distinct non-chimeric centroids and retained their original direction with q < 0.05 in the denoised analysis. Each met the technical and CLR support threshold of candidate-family q < 0.10; 11 also met q < 0.05 in both frameworks (Supplementary Tables 11 and 12 and Supplementary Figs. 3 and 4).

Local and public classifiers distinguished COPD cases from controls

The primary reduced-QC-adjusted exact-sequence classifier achieved a mean out-of-fold ROC-AUC of 0.8891 (95% CI, 0.8137–0.9510; Table 3 and Fig. 4a–d). Precision–recall AUC was 0.8974, balanced accuracy was 0.8100, and mean sensitivity and specificity at the SVM decision threshold of zero were 0.7867 and 0.8333, respectively. The non-chimeric UNOISE classifier achieved a ROC-AUC of 0.8770 (95% CI, 0.7943–0.9452). The common-support analysis included 51 participants and achieved a ROC-AUC of 0.8867 (95% CI, 0.7997–0.9556; Fig. 4g). In 1,000 label permutations that repeated all model-development steps, no permuted estimate reached either observed full-cohort ROC-AUC (both empirical p = 0.0010; Fig. 4f).

Table 3
Core classification performance across the local 16S and public metagenomic analyses.

Analysis	Cohort and performance context	Model	ROC-AUC (95% CI)	PR-AUC	Balanced accuracy	Sensitivity	Specificity
Local 16S primary analysis	30 hospitalized patients with COPD / 30 controls; repeated nested five-fold internal cross-validation	Reduced-QC-residualized exact-sequence linear SVM	0.8891 (0.8137–0.9510)	0.8974	0.8100	0.7867	0.8333
Local 16S denoising sensitivity	30 hospitalized patients with COPD / 30 controls; repeated nested five-fold internal cross-validation	Reduced-QC-residualized non-chimeric UNOISE linear SVM	0.8770 (0.7943–0.9452)	0.8872	0.8100	0.7967	0.8233
Local 16S common-support sensitivity	21 hospitalized patients with COPD / 30 controls within the reduced-QC overlap; repeated nested five-fold internal cross-validation	Reduced-QC-residualized exact-sequence linear SVM	0.8867 (0.7997–0.9556)	0.8204	0.8119	0.8238	0.8000
Public metagenomic evaluation	Development: 28 COPD / 29 controls; evaluation: 16 COPD / 22 controls	Locked CLR elastic-net logistic regression	0.9205 (0.8182–0.9943)	0.8639	0.8324	0.9375	0.7273

Local estimates are means from ten repeats with 10,000-resample participant-bootstrap confidence intervals. Local sensitivity, specificity and balanced accuracy used the linear SVM decision threshold of zero. The common-support analysis included 21 patients with COPD and 30 controls. The public model was fitted in the complete development cohort and applied once to the evaluation cohort; its confidence interval used 2,000 stratified bootstrap resamples. Public sensitivity and specificity used the fixed 0.5 threshold. Complete comparator and sensitivity results are provided in Supplementary Tables 5, 9 and 13.

Sequencing-quality variables alone yielded a ROC-AUC of 0.8186. On identical outer splits, residualized microbial features achieved a ROC-AUC of 0.8891, while the combined microbial and quality model reached 0.9610, an increase of 0.1424 over the quality-only model (95% CI, 0.0310–0.2701; Fig. 4e). Microbial classification remained similar with outer-training-fold filtering (ROC-AUC, 0.8850), non-chimeric reconstruction, common-support

restriction and male-only retraining (ROC-AUC, 0.8737; Fig. 4c,g; Supplementary Tables 5–8 and Supplementary Fig. 5).

In the public evaluation cohort, the CLR elastic-net classifier achieved a ROC-AUC of 0.9205 (95% CI, 0.8182–0.9943; Table 3 and Fig. 5a–d). Precision–recall AUC was 0.8639 and balanced accuracy was 0.8324. At the fixed 0.5 threshold, 15 of 16 patients with COPD and 16 of 22 controls were correctly classified, corresponding to 93.75% sensitivity and 72.73% specificity. The within-sample-rank analysis achieved a ROC-AUC of 0.8864. Estimates across pseudocount, feature-set-size and leave-one-participant-out analyses ranged from 0.8580 to 0.9583 (Fig. 5e–g; Supplementary Tables 9 and 10 and Supplementary Fig. 6).

Discussion

This study identified a pretreatment stool bacterial phenotype in hospitalized patients with COPD that combined altered community organization, a prioritized 12-sequence signature and internally validated case–control discrimination. A complementary genome-level classifier also achieved high discrimination in a separately recruited public evaluation cohort. Together, these findings connect an ecologically defined intestinal phenotype with sequence-resolved markers and participant-level classification.

Most previous COPD microbiome studies have compared stable outpatients with healthy participants and have reported differences in taxonomic composition, microbial function or circulating and fecal metabolites [3–6]. By contrast, the present comparator group comprised patients hospitalized with pneumonia, pulmonary malignancy or mass, pulmonary embolism, bronchiectasis with infection or pleural effusion. This clinically active comparison extends earlier observations to an inpatient differential-diagnosis setting in which respiratory inflammation, multimorbidity and acute symptoms are shared features rather than attributes unique to COPD. Sampling before inpatient antibiotics or systemic glucocorticoids further linked the observed profiles to the pretreatment state.

Higher Simpson and inverse Simpson diversity without corresponding differences in Shannon diversity, richness or Pielou evenness indicates a change in dominance-sensitive community organization rather than a uniform gain in taxa. Case–control status explained 5.50% of Bray–Curtis variation, with comparable dispersion between groups. Previous COPD cohorts have reported variable alpha-diversity patterns [5,22], whereas the present combination of dominance-sensitive diversity, multivariate composition and non-chimeric reconstruction provides a more specific description of community reorganization in hospitalized patients.

The prioritized panel included *Fusobacterium*-, *Prevotella*-, *Catonella*-, *Filifactor*-, *Oribacterium*- and *Treponema*-associated sequences enriched in COPD and a *Neisseria*-associated sequence enriched in controls. The *Fusobacterium* signal is consistent with earlier COPD observations [4]. Several represented genera are common members of the oral microbiota, placing oral health, swallowing and oral–gut microbial exchange within a plausible biological framework for the gut–lung axis. Retaining the markers as exact sequences, rather than relying only on genus-level summaries, provides stable molecular targets for replication and targeted assay development.

The primary exact-sequence classifier distinguished hospitalized cases from respiratory controls under repeated nested validation, with similar estimates after non-chimeric reconstruction, common-support restriction and male-only retraining. This performance links the classification result to the sequence-level community signal and shows that a compact, traceable feature space can retain clinically relevant information. The public genome-level classifier achieved a ROC-AUC of 0.9205 in a separately recruited evaluation cohort, complementing the local 16S

analysis with evidence at a different molecular resolution. The convergence of ecological change, exact-sequence prioritization and genome-level discrimination is the principal translational value of the study.

The study was limited by its single-centre, cross-sectional design, modest sample size, male predominance in the COPD group, heterogeneous respiratory control diagnoses and incomplete spirometry. Amplicon sequencing also limits functional inference and cannot determine the anatomical origin of stool-associated sequences. Future prospective studies should use sex-balanced recruitment, standardized post-bronchodilator phenotyping, harmonized laboratory processing, absolute bacterial quantification, paired oral and fecal sampling and longitudinal assessment across hospitalization and stable follow-up [9,10,13,14,23].

Conclusion

Pretreatment stool profiles from hospitalized patients with COPD showed altered dominance-sensitive diversity, distinct Bray–Curtis composition and a prioritized 12-sequence bacterial signature relative to other respiratory inpatients. The local exact-sequence and public genome-level classifiers achieved ROC-AUCs of 0.889 and 0.921, respectively. These findings define a sequence-resolved intestinal bacterial phenotype of hospitalized COPD and support the potential of stool microbial features as adjunctive markers for differentiating COPD from other respiratory diseases.

Declarations

Data availability

De-identified data and source data supporting the findings of this study—including processed figure and table data, local clinical metadata, case-to-sample linkage metadata, clinical–microbiome association results, participant-level model estimates and FASTA files for the 12 prioritized, 49 dual-supported and 56 directionally concordant bacterial candidates—have been assigned the Zenodo DOI https://doi.org/10.5281/zenodo.21897586 and will be made publicly available upon publication. Direct identifiers are not included. Public genome-level abundance data reused for the complementary classifier were obtained from Bowerman et al. [5] and the source-data file associated with that article.

Code availability

Code used for sequence processing, statistical analysis, machine learning and figure generation will be archived with the study data in Zenodo (https://doi.org/10.5281/zenodo.21897586) and will be made publicly available upon publication with environment and execution documentation.

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Ethics Committee of Jiaxing First Hospital (approval no. 2024-LY-418) and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants or their legally authorized representatives before enrollment.

Data-sharing authorization

Only de-identified study data and analysis outputs are included in the public release. The shared files exclude direct identifiers and are released in accordance with participant consent, ethics approval and institutional data-

governance requirements.

Competing interests

None.

Author contributions

Yuejiao Sun and Jing Feng contributed equally to this work. Yuejiao Sun and Jing Feng conducted the study, collected and analyzed the data, and prepared the original draft of the manuscript. Xixi Gao and Chaoping Zhu assisted with data collection. Yan Liu supervised the study and contributed to the review and editing of the manuscript. Yue Zhou contributed to the investigation, data curation, formal analysis, and visualization. Xiaolong Ma conceptualized and supervised the study and critically reviewed and revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Clinical trial number

Not applicable.

Funding

This work was funded by National Oncology Clinical Key Speciality 2023-GJZK-001),Jiaxing Key Laboratory of Precision Treatment for Lung Cancer and Key Construction Disciplines of Provincial and Municipal Co construction of Zhejiang(NO.2023-SSGJ-002)Science and Technology Project of Jiaxing (NO.2024AD10018) .

Acknowledgements

We thank the clinical and laboratory staff of Jiaxing First Hospital for assistance with participant recruitment, specimen collection and clinical-record review. We thank Guangzhou Omicsmaster Biotechnology Co., Ltd. for amplicon sequencing and initial bioinformatic support. We are grateful to all participants and their families.

References

1. Ozcam M, Lynch SV. The gut-airway microbiome axis in health and respiratory diseases. Nat Rev Microbiol. 2024;22:492–506. https://doi.org/10.1038/s41579-024-01048-8.
2. Dang AT, Marsland BJ. Microbes, metabolites, and the gut-lung axis. Mucosal Immunol. 2019;12:843–50. https://doi.org/10.1038/s41385-019-0160-6.
3. Li N, et al. The gut microbiome as a potential source of non-invasive biomarkers of chronic obstructive pulmonary disease. Front Microbiol. 2023;14:1173614. https://doi.org/10.3389/fmicb.2023.1173614.
4. Chiu Y-C, et al. Comprehensive profiling of the gut microbiota in patients with chronic obstructive pulmonary disease of varying severity. PLoS ONE. 2021;16:e0249944. https://doi.org/10.1371/journal.pone.0249944.
5. Bowerman KL, et al. Disease-associated gut microbiome and metabolome changes in patients with chronic obstructive pulmonary disease. Nat Commun. 2020;11:5886. https://doi.org/10.1038/s41467-020-19701-0.
6. Li N, et al. Gut microbiota dysbiosis contributes to the development of chronic obstructive pulmonary disease. Respir Res. 2021;22:274. https://doi.org/10.1186/s12931-021-01872-z.

7. Costea PI, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35:1069–76. https://doi.org/10.1038/nbt.3960.
8. Sinha R, et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control project consortium. Nat Biotechnol. 2017;35:1077–86. https://doi.org/10.1038/nbt.3981.
9. Knight R, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16:410–22. https://doi.org/10.1038/s41579-018-0029-9.
10. Abellan-Schneyder I et al. Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. mSphere 6, e01202-20 (2021). https://doi.org/10.1128/mSphere.01202-20
11. Asnicar F, Thomas AM, Passerini A, Waldron L, Segata N. Machine learning for microbiologists. Nat Rev Microbiol. 2024;22:191–205. https://doi.org/10.1038/s41579-023-00984-1.
12. Papoutsoglou G, et al. Machine learning approaches in microbiome research: challenges and best practices. Front Microbiol. 2023;14:1261889. https://doi.org/10.3389/fmicb.2023.1261889.
13. Vujkovic-Cvijin I, et al. Host variables confound gut microbiota studies of human disease. Nature. 2020;587:448–54. https://doi.org/10.1038/s41586-020-2881-9.
14. Gibbons SM, Duvallet C, Alm EJ. Correcting for batch effects in case-control microbiome studies. PLoS Comput Biol. 2018;14:e1006102. https://doi.org/10.1371/journal.pcbi.1006102.
15. Vandeputte D, et al. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut. 2016;65:57–62. https://doi.org/10.1136/gutjnl-2015-309618.
16. Choo JM, Leong LEX, Rogers GB. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep. 2015;5:16350. https://doi.org/10.1038/srep16350.
17. Song SJ et al. Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. mSystems 1, e00021-16 (2016). https://doi.org/10.1128/mSystems.00021-16
18. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4, e2584 (2016). https://doi.org/10.7717/peerj.2584
19. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv. 2016;081257. https://doi.org/10.1101/081257.
20. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics. 2011;27:2194–200. https://doi.org/10.1093/bioinformatics/btr381.
21. Mirzayi C, et al. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med. 2021;27:1885–92. https://doi.org/10.1038/s41591-021-01552-x.
22. Rotevatn AØ, et al. Gut microbiota in chronic obstructive pulmonary disease varies by CT-verified emphysema status. Eur Clin Respir J. 2025;12:2470499. https://doi.org/10.1080/20018525.2025.2470499.
23. Vandeputte D, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017;551:507–11. https://doi.org/10.1038/nature24460.

Figures

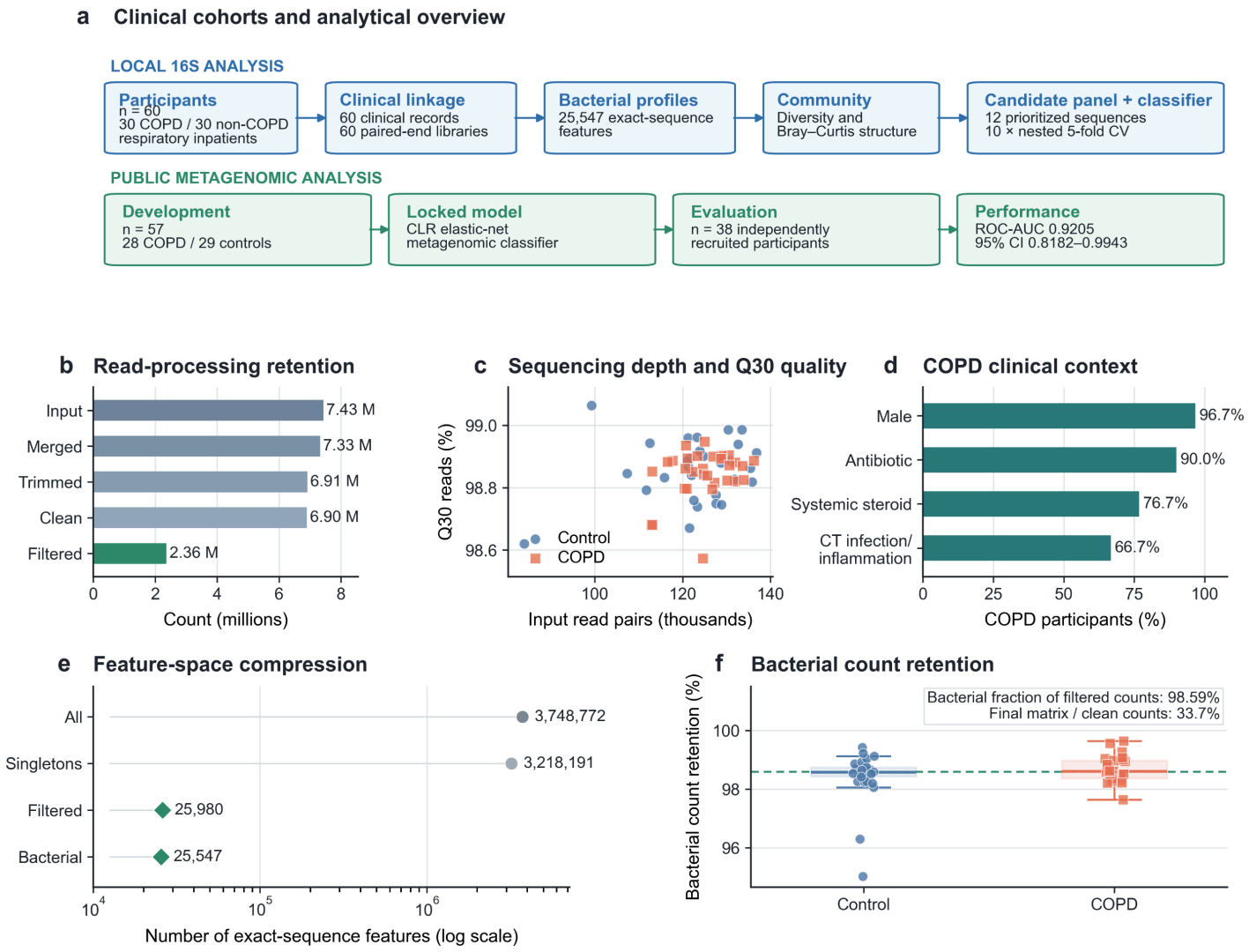


Figure 1

Study design, clinical context and sequencing characteristics. a, Study populations and analytical overview. The local 16S analysis comprised 30 hospitalized patients with COPD and 30 hospitalized non-COPD respiratory controls; the public analysis comprised 57 development and 38 evaluation participants. **b,** Read-processing retention from input pairs to the filtered bacterial matrix. **c,**Participant-level input depth and Q30 fraction. **d,** Selected COPD clinical characteristics; treatment variables describe medications recorded after pretreatment sampling during hospitalization. **e,** Compression from all observed exact sequences to 25,547 bacterial features. **f,** Participant-level bacterial-count retention after abundance/prevalence filtering; 98.59% of filtered counts remained bacterial, corresponding to 33.7% of clean sequences in the final matrix.

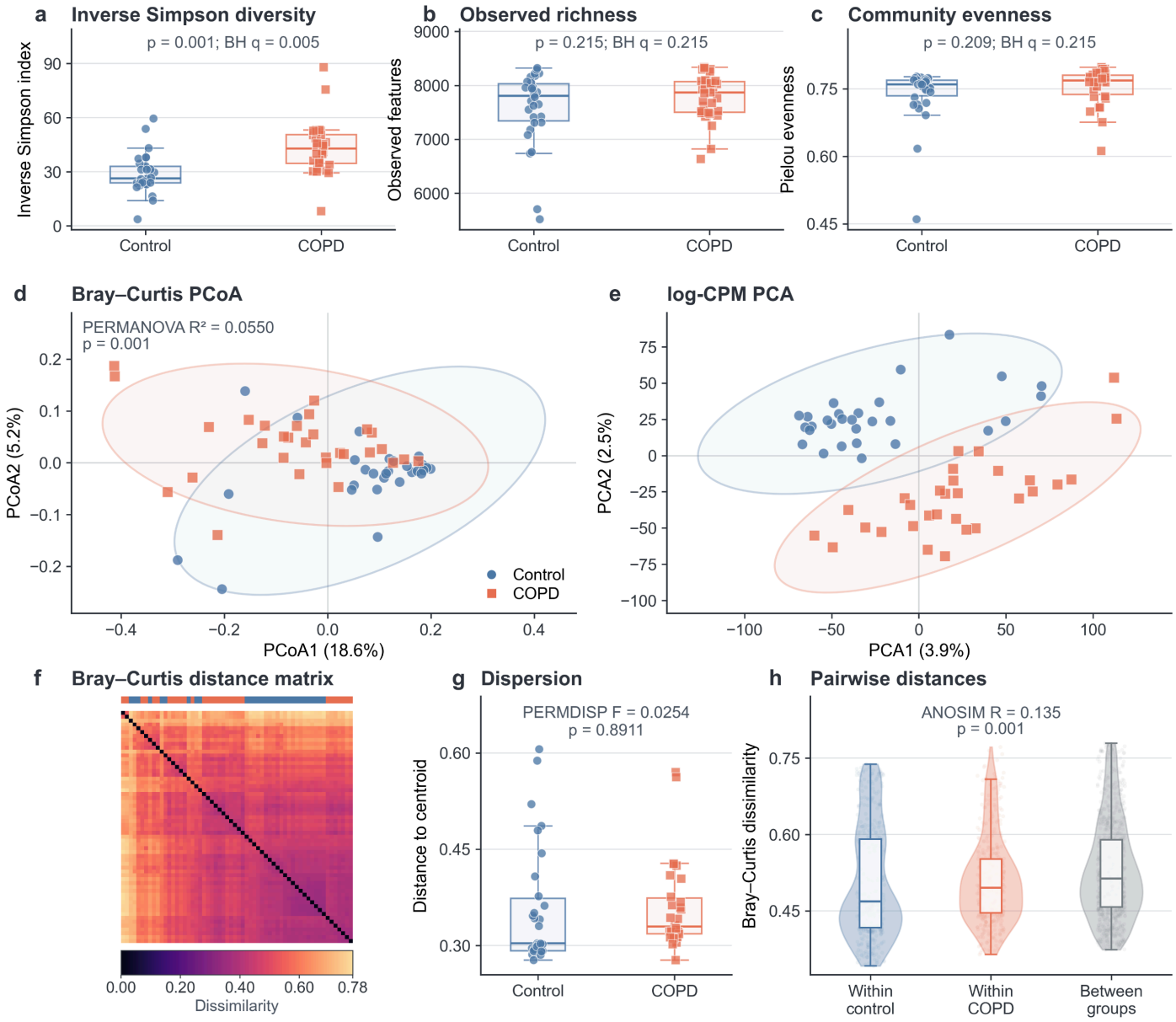


Figure 2

Bacterial community organization in hospitalized patients with COPD. a,Inverse Simpson diversity. **b,**Observed exact-sequence richness. **c,**Pielou evenness. **d,** Principal coordinates analysis of Bray–Curtis dissimilarity calculated from total-sum-scaled relative-abundance profiles, with PERMANOVA effect size. **e,** Principal-component analysis of log-CPM profiles. **f,**Participant-level Bray–Curtis distance matrix ordered by unsupervised average-linkage clustering; group is shown in the top annotation strip. **g,** Distance to the group centroid and PERMDISP result. **h,** Within-control, within-COPD and between-group Bray–Curtis distances with ANOSIM. The PCoA and PCA overlays are 95% covariance ellipses. Simpson and inverse Simpson BH q = 0.0175 and 0.005; Shannon diversity, richness and Pielou evenness all BH q = 0.215. PERMANOVA R^2 = 0.0550, p = 0.001; PERMDISP F = 0.0254, p = 0.8911; ANOSIM R = 0.135, p = 0.001.

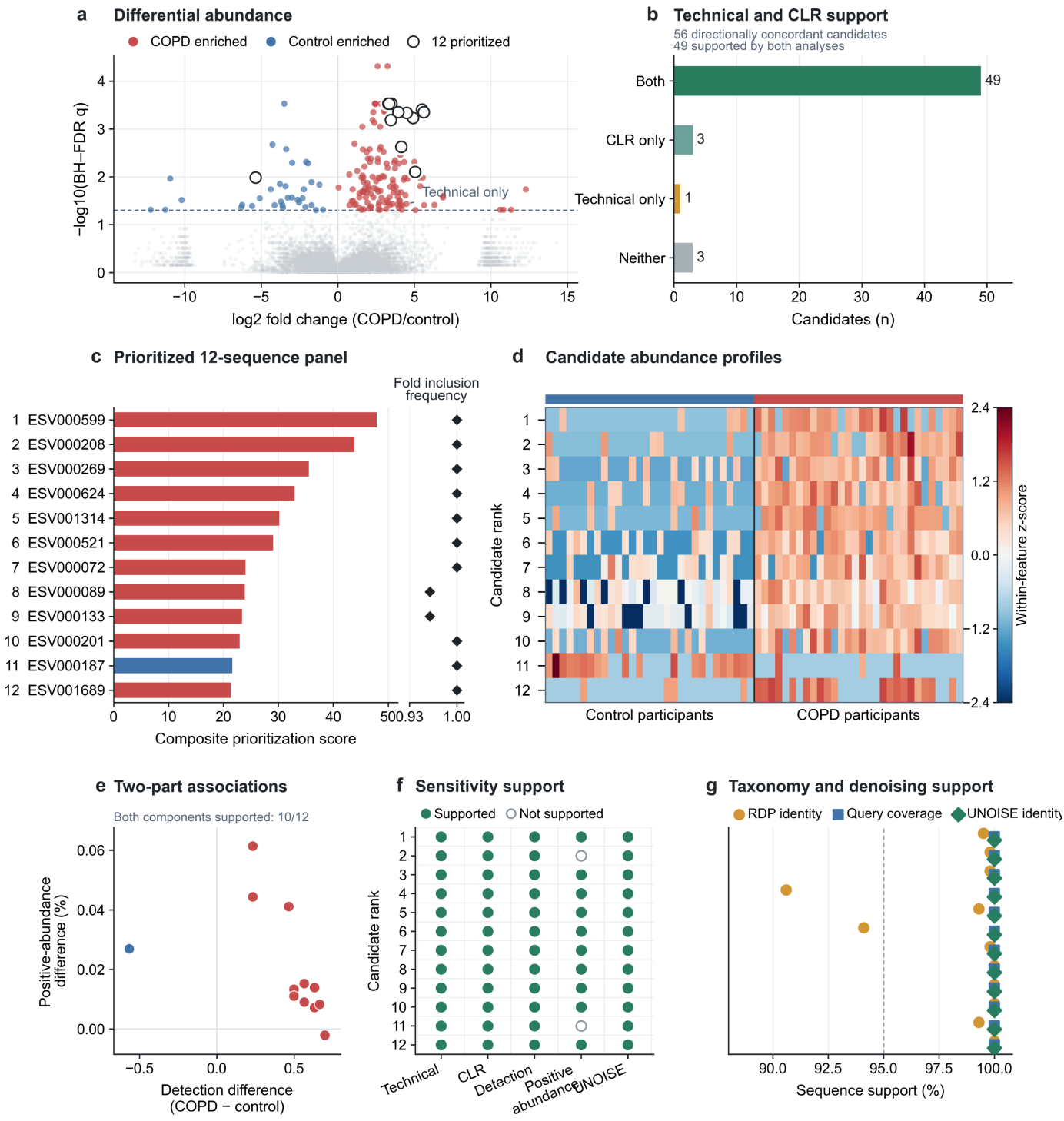


Figure 3

Candidate prioritization across complementary within-cohort analyses. a,Differential effects among 25,547 bacterial observed exact-sequence features; the dashed line denotes BH-FDR q = 0.05 and open circles identify the 12 prioritized candidates. **b,**Intersection of technical-adjustment and centered-log-ratio (CLR) support among 56 directionally concordant candidates. Both analyses supported 49 candidates, CLR alone supported three, technical adjustment alone supported one and three had neither support. **c,** Composite scores for the 12 highest-ranked candidates that met the technical, CLR and distinct non-chimeric-centroid criteria; diamonds show fold

inclusion frequency. **d,** Feature-wise standardized participant abundance profiles, ordered by group; rows follow the ranking in panel c. **e,**Detection-prevalence differences versus positive-sample abundance differences. **f,** Candidate-level technical, CLR, detection, positive-abundance and UNOISE support. **g,** RDP reference identity, query coverage and sequence-to-UNOISE-centroid identity.

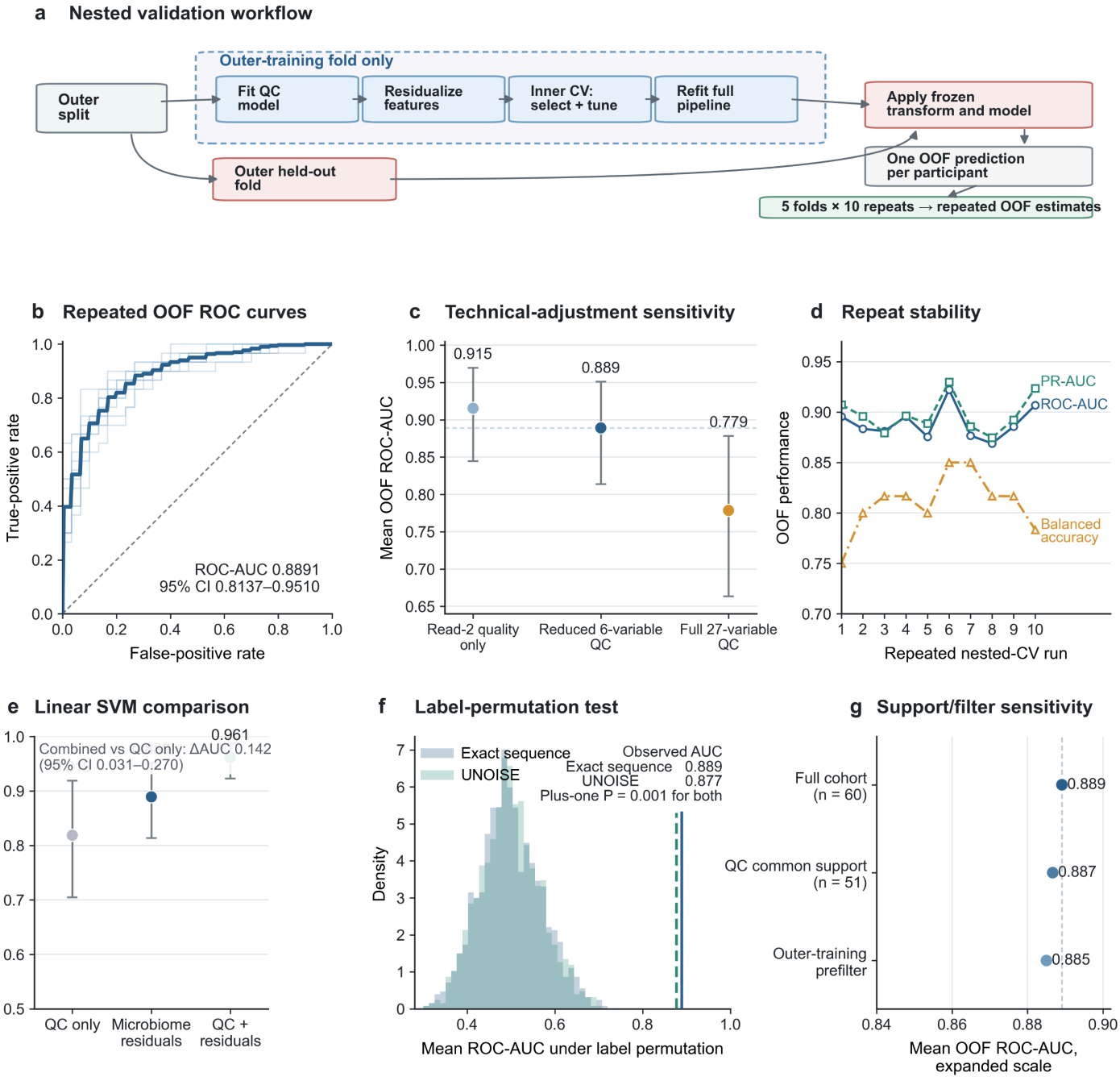


Figure 4

Nested internal classification and technical sensitivity analyses. a,Nested-validation design with outer-fold residualization, inner-loop feature selection and tuning, and held-out prediction. **b,** Complete out-of-fold ROC curves for ten repeated nested five-fold evaluations of the primary reduced-QC residual-variance SVM. **c,** Mean ROC-AUC and stratified participant-bootstrap 95% confidence intervals for Read 2-quality-only, reduced six-

variable QC and full 27-variable QC residualization sensitivities. **d,** ROC-AUC, precision–recall AUC and balanced accuracy across ten repeats of the primary model. **e,** Reduced-QC-only, residualized-microbiome-only and combined linear SVM performance on the same outer splits; error bars are stratified participant-bootstrap 95% confidence intervals. **f,** Label-permutation distributions obtained after repeating all model-development steps for the exact-sequence and UNOISE analyses; vertical lines denote observed estimates and both plus-one empirical p values were 0.001. **g,** Full-cohort, reduced-QC common-support and outer-training-fold filtering estimates. The primary mean ROC-AUC was 0.8891 (95% CI, 0.8137–0.9510), and the common-support estimate was 0.8867 (95% CI, 0.7997–0.9556).

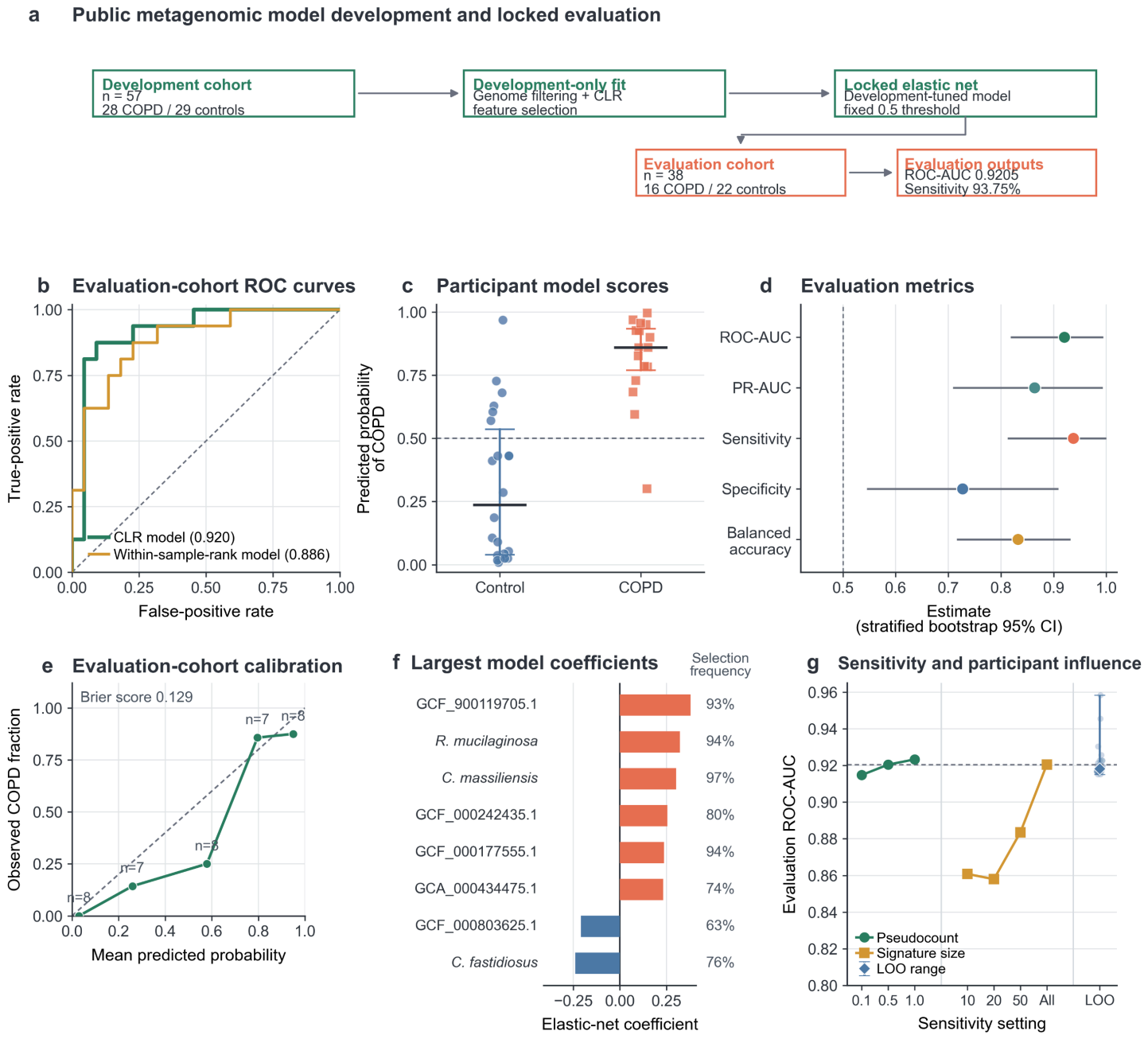


Figure 5

Public metagenomic classifier development and evaluation. a, Development-cohort processing, locked model fitting and evaluation-cohort assessment. **b,** Receiver operating characteristic curves for the primary CLR model and within-sample-rank sensitivity model. **c,** Participant-level CLR model scores; horizontal marks show medians and interquartile ranges and the dashed line denotes the fixed 0.5 threshold. **d,**Evaluation metrics with stratified bootstrap 95% confidence intervals. Sensitivity, specificity and balanced accuracy used the fixed threshold. **e,** Descriptive calibration in five quantile-based groups in the evaluation cohort; labels denote participants per group and the diagonal indicates perfect calibration. **f,** Eight largest absolute non-zero CLR coefficients; percentages denote selection frequency across 1,000 development-cohort bootstraps. **g,** Pseudocount, feature-set-size and leave-one-participant-out sensitivity analyses. Development n = 57; evaluation n = 38.