Supplementary Materials

Supplementary Results

Optimization of propidium monoazide (PMA) concentration for O_pma

Two concentrations of PMA (10 μM and 50 μM) were tested on six bronchoalveolar lavage fluid (BALF) samples, and two concentrations of PMA (10 μM and 30 μM) were tested on six oropharyngeal (OP) samples. Higher concentrations of PMA were more effective at removing host DNA compared to 10 μM (Supplementary Fig. 1A, 1B). However, higher concentrations of PMA also resulted in a more severe loss of bacterial DNA, which was significant for 50 μM in BALF ($p < 0.05$, Wilcoxon matched pairs test, Supplementary Fig. 1C, 1D). Consequently, there was no difference in bacterial DNA percentage between the different PMA concentrations (Supplementary Fig. 1E, 1F). Therefore, 10 μM was selected as the optimal concentration of PMA.

Optimization of saponin concentration for S_ase

Saponin concentrations ranging from 0.025% to 2.5% have been reported in the literature (1-4). We tested three concentrations—0.025%, 0.1%, and 0.5%—using five BALF samples and 8 OP samples. The 0.025% saponin concentration resulted in the lowest human DNA content, the highest bacterial DNA content, and the highest bacterial DNA percentage after the host removal process, although not statistically significant (Supplementary Fig. 1G-L). Therefore, 0.025% was selected as the optimal saponin concentration.

Optimization of cryopreservation for BALF samples

Cryopreservation is often required before downstream processing of clinical samples. The freezing and thawing process can lead to cell damage of microbes, potentially affecting the host removal process. To address this

concern, we tested six-month cryopreservation with or without 25% glycerol using 11 BALF samples. After host removal with K_qia, bacterial DNA in samples cryopreserved with glycerol was remarkably higher than in those without glycerol (median 402.8 vs 6.4 pg/ml BALF, $p < 0.001$, Wilcoxon matched pairs test, Supplementary Fig. 1M). Although samples cryopreserved with glycerol also showed a trend of higher human DNA (Supplementary Fig. 1N), the bacterial DNA percentage was marginally higher in samples cryopreserved with glycerol (Supplementary Fig. 1O). Thus, cryopreservation with 25% glycerol was used for the storage of all samples in the study.

Comparison of the effectiveness among S_ase, MEM, and F_ase host DNA depletion methods

Eight BALF samples were used to assess host depletion efficiency using S_ase, F_ase, MEM, and R_ase methods. All methods demonstrated similar bacterial retention rates (Supplementary Fig. 9A). MEM and S_ase exhibited higher microbial proportion and species richness compared to F_ase and R_ase, all of which surpassed Raw (Supplementary Fig. 9B,C). F_ase demonstrated the highest accuracy, followed by MEM, whereas S_ase exhibited the highest contamination proportion and a significant distortion of the microbial community, notably with a substantial reduction in *Mycoplasma pneumoniae* (Supplementary Fig. 9D). MEM demonstrated the largest radar char area (Supplementary Fig. 9E).

Supplementary Table 1. Gender, age, and clinical diagnosis of enrolled patients.

Characteristic	N = 35 (%)
Gender, n (%)	
Male	22 (62.9)
Female	13 (27.1)
Age (years), n (%)	
17-30	10 (28.6)
31-50	7 (20)
51-70	12 (34.3)
71-82	6 (17.1)
Clinical diagnosis, n (%)	
Pneumonia	26 (74.3)
COPD	2 (5.7)
Bronchiectasis	1 (2.9)
Bronchitis	2 (5.7)
Lung cancer	1 (2.9)
Esophagus cancer	1 (2.9)
Rib fracture	1 (2.9)
Dysphagia	1 (2.9)

Supplementary Table 2. Cost and time of host DNA depletion methods.

	Price ($)			
	Total	Cell removal	DNA removal	DNA extraction
Raw	6.3	0.0	0.0	6.3
R_ase	10.4	0.0	4.1	6.3
O_pma	6.9	0.1	0.5	6.3
O_ase	10.5	0.1	4.1	6.3
S_ase	10.4	0.0	4.1	6.3
F_ase	15.3	4.8	4.1	6.3
K_qia	16.3	-	-	-
K_zym	13.6	-	-	-
	Time (min)			
Raw	38	0	0	38
R_ase	83-233	0	45-195*	38
O_pma	91	28	25	38
O_ase	111-261	28	45-195*	38
S_ase	95-245	12	45-195*	38
F_ase	84-234	1	45-195*	38
K_qia	179	42	60	77
K_zym	96	22	40	34

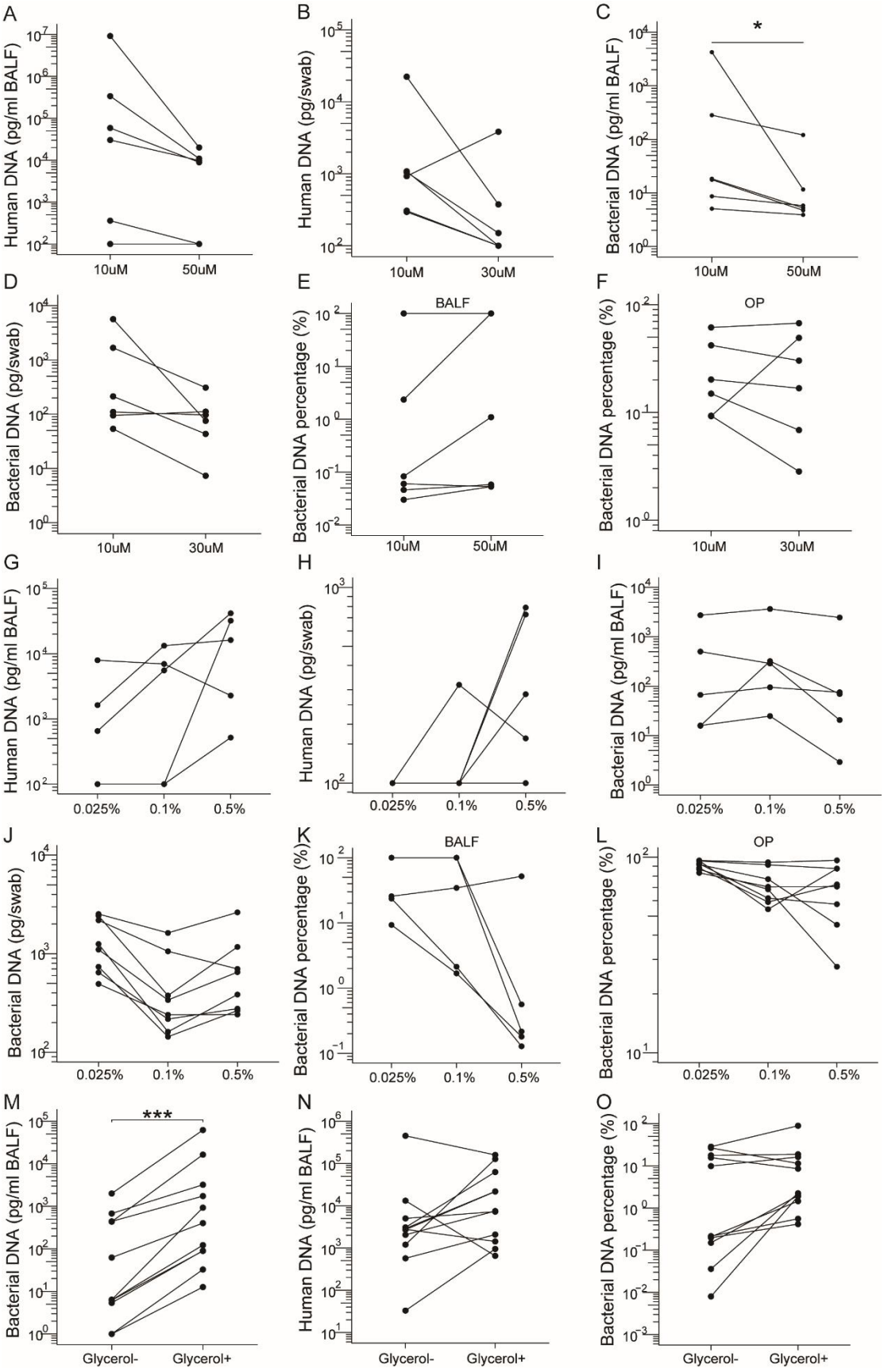
*The duration of endonuclease digestion can vary between 30 min and 180 min.

Supplementary Table 3. Composition of microbial mock community

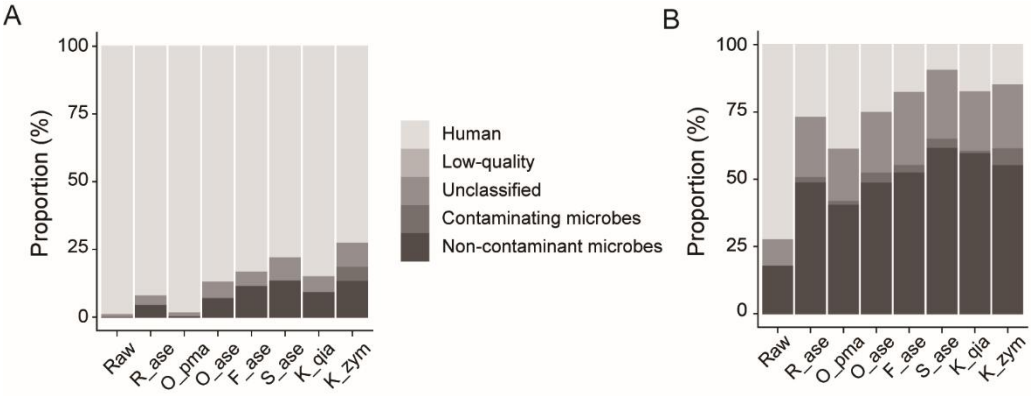
Taxon name	Proportion[a]
Streptococcus salivarius	17.8%
Staphylococcus aureus	14.0%
Achromobacter xylosoxidans	12.1%
Ralstonia pickettii	9.7%
Enterobacter cloacae	9.6%
Prevotella intermedia	9.5%
Pseudomonas putida	8.6%
Acinetobacter baumannii	5.6%
Pseudomonas aeruginosa	2.6%
Brucella anthropi	2.6%
Ligilactobacillus salivarius	2.4%
Aggregatibacter actinomycetemcomitans	2.2%
Neisseria subflava	0.8%
Moraxella catarrhalis	0.1%
Gemella haemolysans	0.1%

[a] The proportion was estimated from the number of sequencing reads.

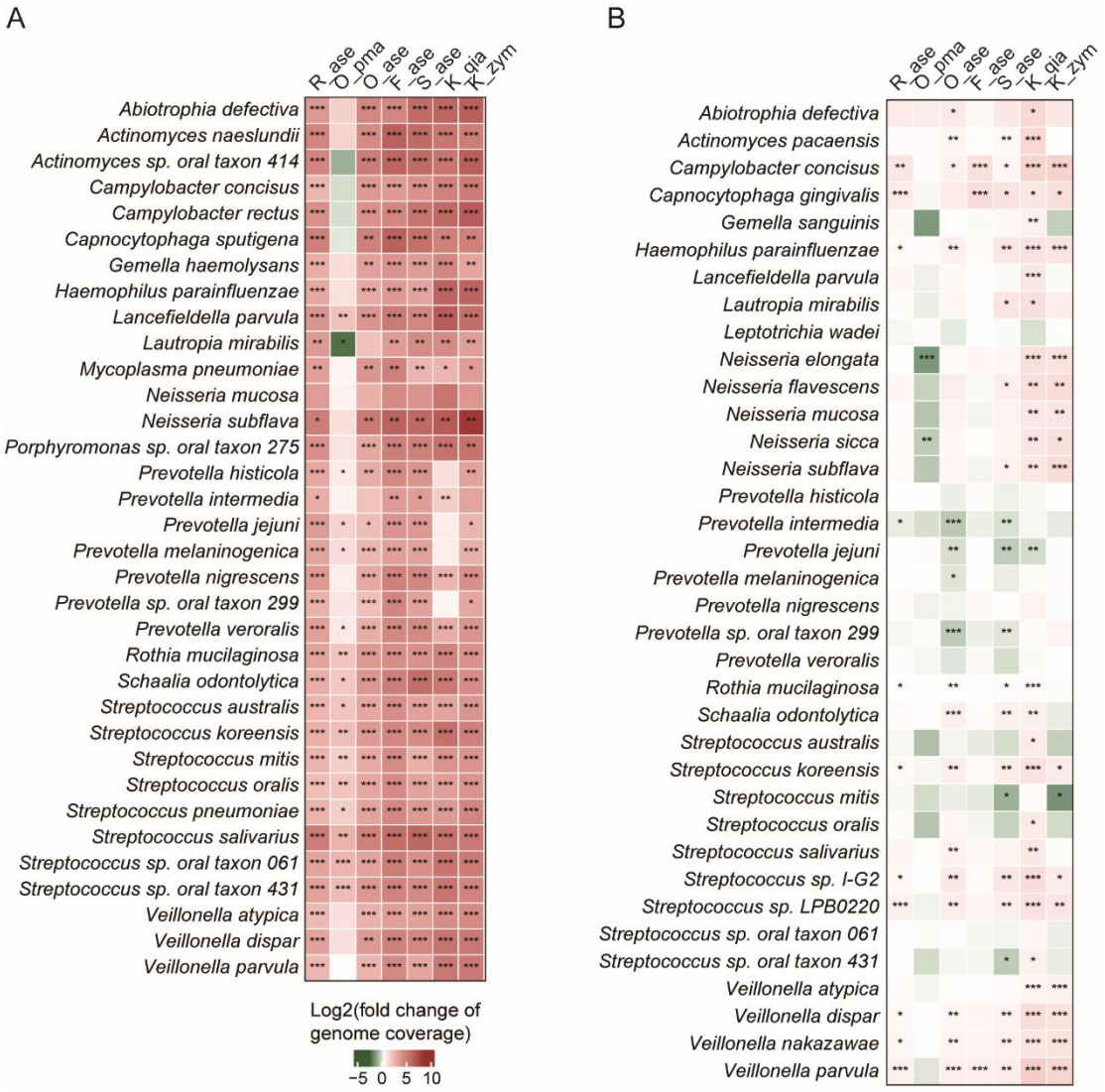
Supplementary Figure 1. Optimization of Experimental conditions. (A-F) Optimization of PMA concentration for O_pma. Human DNA load in BALF (A) and OP (B), bacterial DNA load in BALF (C) and OP (D), and bacterial DNA percentage in BALF (E) and OP (F) were evaluated for 10 μM, 30 μM (OP) and 50 μM (BALF) PMA. n = 6 for BALF and OP. (G-L) Optimization of saponin concentration for S_ase. Human DNA load in BALF (G) and OP (H), bacterial DNA load in BALF (I) and OP (J), and bacterial DNA percentage in BALF (K) and OP (L) were evaluated for 0.025%, 0.1%, and 0.5% saponin. n = 5 for BALF and n = 8 for OP. (M-O) Optimization of sample storage condition. Bacterial DNA load (M), human DNA load (N), and bacterial DNA percentage (O) were evaluated for BALF samples with or without 25% glycerol during storage. In M-O, K_qia was used to deplete host DNA and extract nucleic acids. n = 11. Each dot represents a sample. *$p < 0.05$, ***$p < 0.001$, Wilcoxon matched pairs tests.


A
Human DNA (pg/ml BALF)
10uM
50uM
B
Human DNA (pg/swab)
10uM
30uM
C
*
Bacterial DNA (pg/ml BALF)
10uM
50uM
D
Bacterial DNA (pg/swab)
10uM
30uM
E
BALF
Bacterial DNA percentage (%)
10uM
50uM
F
OP
Bacterial DNA percentage (%)
10uM
30uM
G
Human DNA (pg/ml BALF)
0.025%
0.1%
0.5%
H
Human DNA (pg/swab)
0.025%
0.1%
0.5%
I
Bacterial DNA (pg/ml BALF)
0.025%
0.1%
0.5%
J
Bacterial DNA (pg/swab)
0.025%
0.1%
0.5%
K
BALF
Bacterial DNA percentage (%)
0.025%
0.1%
0.5%
L
OP
Bacterial DNA percentage (%)
0.025%
0.1%
0.5%
M
***
Bacterial DNA (pg/ml BALF)
Glycerol-
Glycerol+
N
Human DNA (pg/ml BALF)
Glycerol-
Glycerol+
O
Bacterial DNA percentage (%)
Glycerol-
Glycerol+

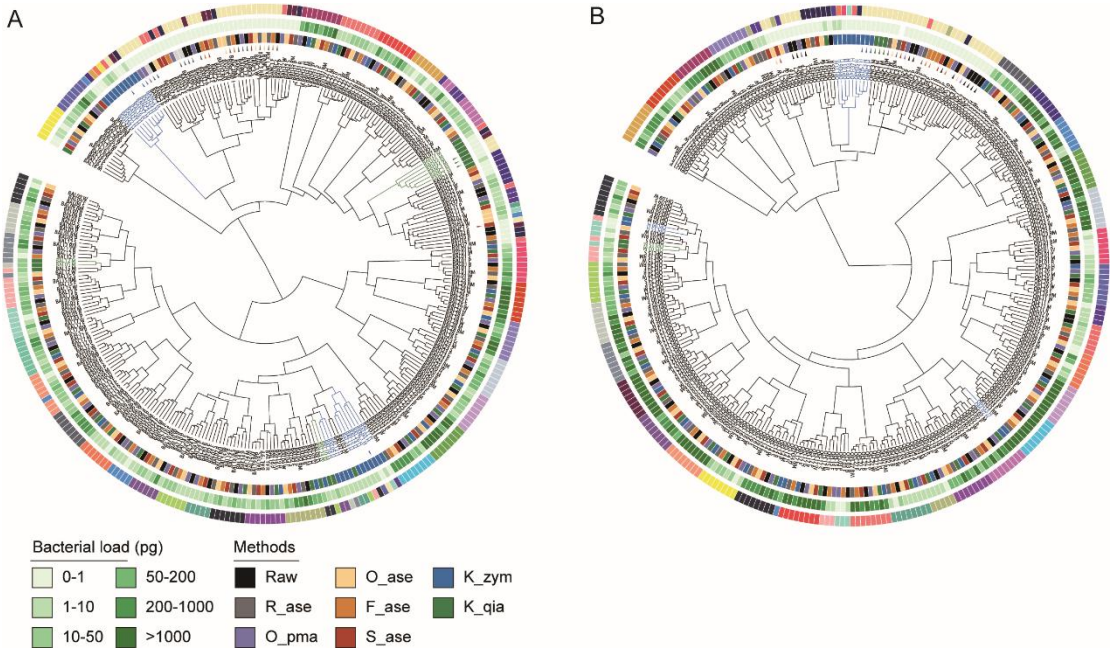
Supplementary Figure 2. Composition of sequencing reads in raw sequencing data. (A) Composition for BALF samples; (B) Composition for OP samples. Microbes refer to species assigned to the taxa including archaea, bacteria, fungi, and viruses.



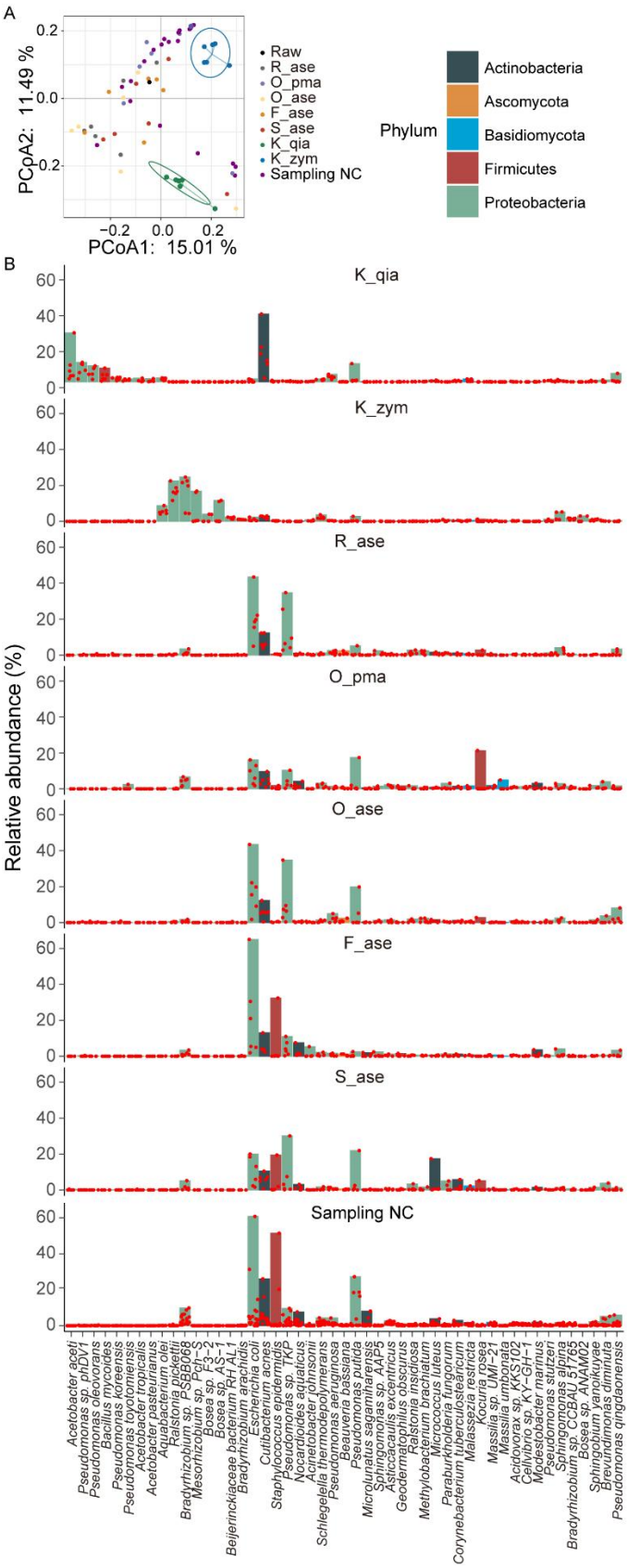
Supplementary Figure 3. Impact of host DNA depletion on microbial genome coverage. (A-B) Fold changes of microbial genome coverage between host DNA depletion methods and Raw samples for BALF (A) and OP (B). Median values are shown. Data from eight samples from the same patient was rarefied to the lowest sequencing depth among them. **p*.adj < 0.05, ***p*.adj < 0.01, ****p*.adj < 0.001, Wilcoxon matched pairs tests. n = 35 for BALF and n = 34 for OP.



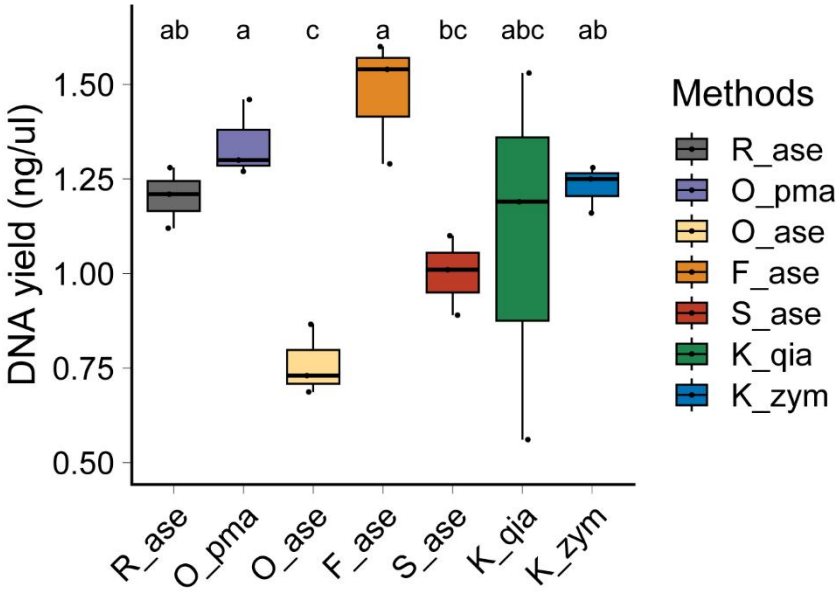
Supplementary Figure 4. Influence of host DNA depletion on the fidelity of microbiome profiling. (A-B) Hierarchical clustering of BALF (A) and OP (B) samples treated with different host DNA depletion methods. The layers from inside to outside indicate host DNA depletion methods, bacterial load, and patient identify, respectively. Triangles indicate negative control samples. The dendrograms were visualized using iTOLs (5). Jensen-Shannon distance (JSD) was calculated using all microbes, including suspected contaminants. n = 35 for BALF and n = 34 for OP.



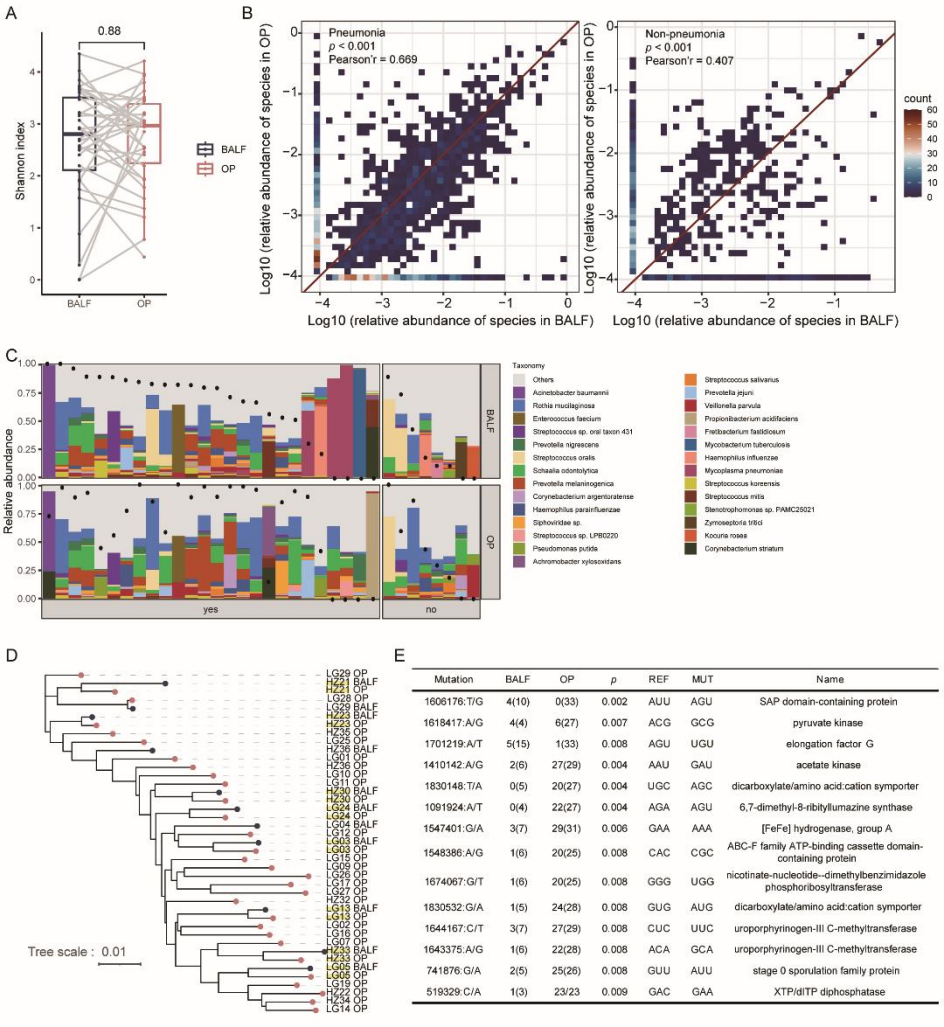
Supplementary figure 5. Contaminants introduced by host DNA depletion methods. (A) PCoA plot based on JSD of the microbial composition of negative control samples. (B) Abundant microbial species in negative control samples. The top 15 most abundant species in negative control samples for each method are shown, resulting in a total of 49 species. The highest abundances are indicated by the bars, and the abundances of individual samples are indicated by the dots. The color of the bars reflects the phylum to which each species belongs. Sampling NC refers to the negative controls for the BALF and OP samples processed using the Raw method, while other samples were deionized water processed by different host DNA depletion methods (labeled on top of each panel). n = 6 replicates.


A
PCoA2: 11.49 %
PCoA1: 15.01 %
Raw
R_ase
O_pma
O_ase
F_ase
S_ase
K_qia
K_zym
Sampling NC
Phylum
Actinobacteria
Ascomycota
Basidiomycota
Firmicutes
Proteobacteria
B
Relative abundance (%)
K_qia
K_zym
R_ase
O_pma
O_ase
F_ase
S_ase
Sampling NC
Acetobacter aceti
Pseudomonas sp. phDV1
Pseudomonas oleovorans
Bacillus mycoides
Pseudomonas koreensis
Pseudomonas lopnurensis
Acetobacter tropicalis
Acetobacter pasteurianus
Aquabacterium olei
Ralstonia pickettii
Bradyrhizobium sp. PSBB068
Mesorhizobium sp. Pch−3
Bosea sp. FS−2
Bosea sp. AS−1
Beijerinckiaceae bacterium RH AL1
Bradyrhizobium arachidis
Escherichia coli
Cutibacterium acnes
Staphylococcus epidermidis
Pseudomonas sp. TKP
Nocardioides aquaticus
Actinotalea ferrariae
Schlegelella thermodepolymerans
Pseudomonas aeruginosa
Beauveria bassiana
Pseudomonas putida
Microlunatus sagamiharensis
Sphingomonas sp. AAP5
Asticcacaulis excentricus
Geodermatophilus obscurus
Ralstonia insidiosa
Methylobacterium brachiatum
Micrococcus luteus
Paraburkholderia fungorum
Corynebacterium tuberculostearicum
Malassezia restricta
Kocuria rosea
Massilia sp. UMI−21
Massilia timonae
Acidovorax sp. KKS102
Cellvibrio sp. KY−GH−1
Modestobacter marinus
Pseudomonas stutzeri
Sphingomonas alpina
Bradyrhizobium sp. CCBAU 51765
Bosea sp. AVAM02
Sphingobium yanoikuyae
Brevundimonas diminuta
Pseudomonas qingdaonensis

Supplementary figure 6. DNA yield of the mock microbial community after being treated with different host DNA depletion methods. The concentration of extracted DNA was measured using Qubit Fluorometer. The center line of the boxplot represents the median, box limits represent upper and lower quartiles, and whiskers represent the 1.5x interquartile range. Different letters on the top indicate statistically significant differences among methods (Student's t-tests followed by FDR adjustment). n = 3 replicates.

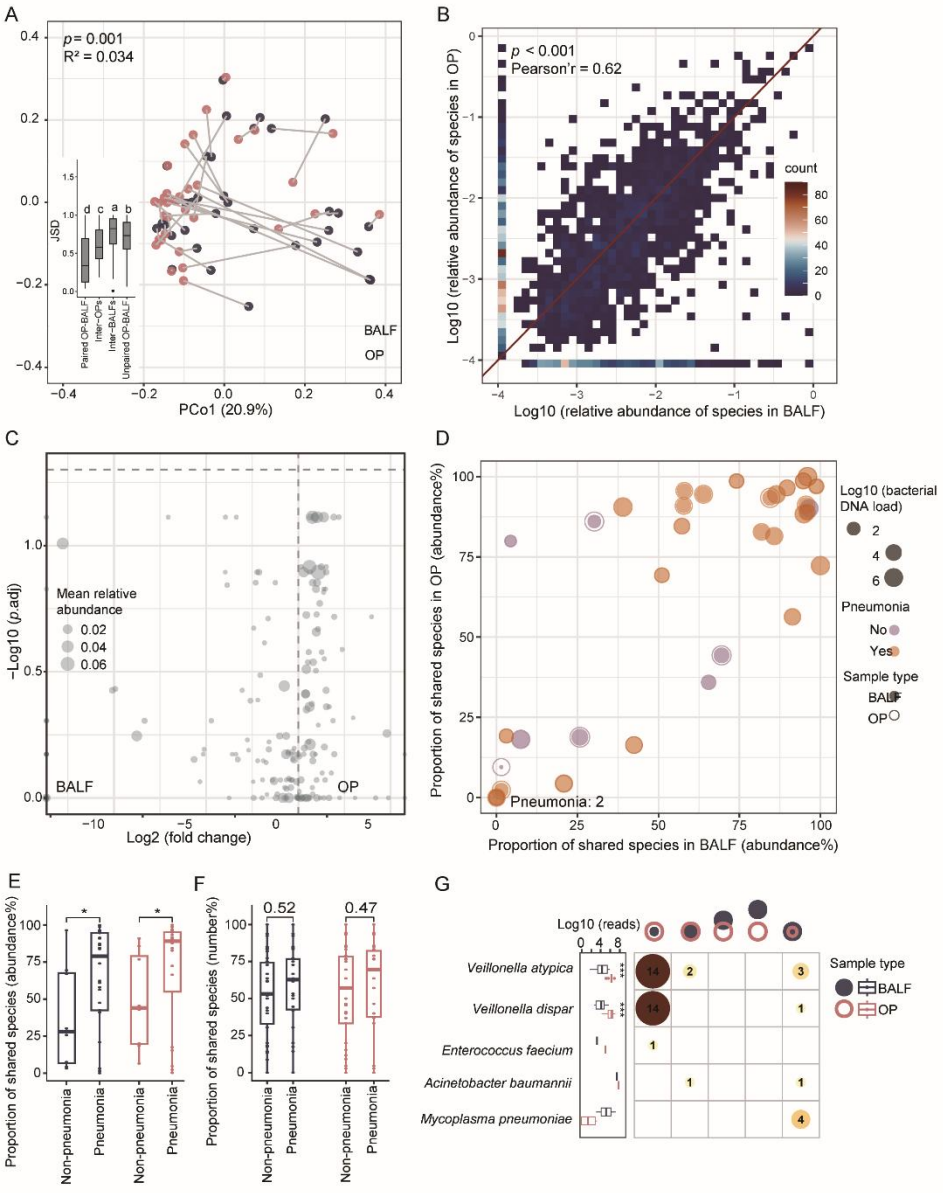


Supplementary figure 7. Comparison between the upper and lower respiratory tract microbiome. (A) Shannon index at the species level. Each dot represents a sample, and samples from the same individuals were connected with lines. n = 34. The center line of the boxplot represents the median, box limits represent upper and lower quartiles, and whiskers represent the 1.5x interquartile range. *p* values from Wilcoxon matched pairs tests were shown on the top. (B) 2D Density plot illustrating the correlation of species abundances between BALF and OP samples in pneumonia (left) and non-pneumonia cases (right). Pearson coefficient and *p* value are shown. (C) Composition of species in BALF and OP samples. The most abundant species from each sample was chosen and displayed in all samples. The abundance of BALF-OP-shared species relative to the total abundance of abundant species is indicated by black dots. In B-C, n = 26 for pneumonia and n = 8 for non-pneumonia diseases. (D) Phylogenetic tree of *V. parvula* genome consensus sequences in individual samples. Red dots represent sequences recovered from OP samples, while blue dots represent sequences recovered from BALF samples. Samples from the same individuals were highlighted on the right. n = 11 for BALF and n = 30 for OP. (E) Nonsynonymous mutations on the *V. parvula* genome showing differential prevalence in BALF and OP samples. Mutations with p values < 0.01 are shown (Fisher's exact tests).

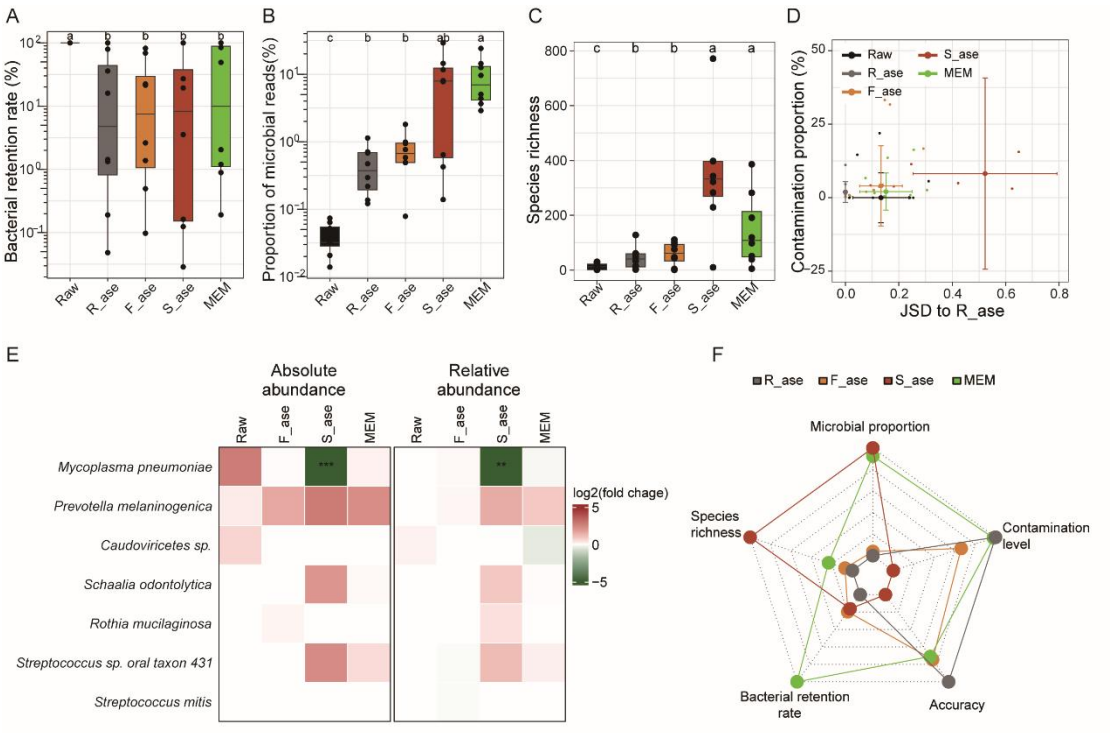


Mutation	BALF	OP	p	REF	MUT	Name
1606176:T/G	4(10)	0(33)	0.002	AUU	AGU	SAP domain-containing protein
1618417:A/G	4(4)	6(27)	0.007	ACG	GCG	pyruvate kinase
1701219:A/T	5(15)	1(33)	0.008	AGU	UGU	elongation factor G
1410142:A/G	2(6)	27(29)	0.004	AAU	GAU	acetate kinase
1830148:T/A	0(5)	20(27)	0.004	UGC	AGC	dicarboxylate/amino acid:cation symporter
1091924:A/T	0(4)	22(27)	0.004	AGA	AGU	6,7-dimethyl-8-ribityllumazine synthase
1547401:G/A	3(7)	29(31)	0.006	GAA	AAA	[FeFe] hydrogenase, group A
1548386:A/G	1(6)	20(25)	0.008	CAC	CGC	ABC-F family ATP-binding cassette domain-containing protein
1674067:G/T	1(6)	20(25)	0.008	GGG	UGG	nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase
1830532:G/A	1(5)	24(28)	0.008	GUG	AUG	dicarboxylate/amino acid:cation symporter
1644167:C/T	3(7)	27(29)	0.008	CUC	UUC	uroporphyrinogen-III C-methyltransferase
1643375:A/G	1(6)	22(28)	0.008	ACA	GCA	uroporphyrinogen-III C-methyltransferase
741876:G/A	2(5)	25(26)	0.008	GUU	AUU	stage 0 sporulation family protein
519329:C/A	1(3)	23/23	0.009	GAC	GAA	XTP/dITP diphosphatase

Supplementary figure 8. Comparison between the upper and lower respiratory tract microbiome using data from untreated samples. This figure corresponds to Figure 6, where Figure 6 displayed results from F_ase treated samples, while this figure displayed results from untreated samples. (A) Principal Co-ordinates Analysis (PCoA) plot based on JSD of microbial compositions. R2 and p values from PERMANOVA are shown. BALF and OP samples from the same individuals were connected with lines. Inserted boxplots show the JSD between different sample types, and different letters on the top indicate statistically significant differences between groups. (B) 2D density plot showing the correlation of species abundances between BALF and OP. Pearson correlation coefficient and p value are shown. (C) Volcano plot showing species with differential relative abundances between BALF and OP samples. The horizontal dash line indicates the adjusted p value of 0.05. In A-B, Wilcoxon matched pairs test followed by FDR adjustment was used. (D) The abundance of OP-BALF-shared species relative to the total species abundance. (E-F) Comparison of the proportion of OP-BALF-shared species between pneumonia and non-pneumonia cases, as per abundance (E) and number (F) of species. In D-F, only abundant species with relative abundances higher than 0.01 were included. (G) The relationship between strains detected in the upper and lower respiratory tracts. The number as well as the color and size of dots indicate the number of strains. Symbols on the top indicate the following relationships: strains in BALF were a subset of those in OP, BALF and OP had the same strains, BALF and OP had partially overlapping strains, BALF and OP had completely different strains, strains in OP were a subset of those in BALF. The boxplots on the left indicate the read number of species. In E-G, asterisks indicate significant differences, *p.adj < 0.05, **p.adj < 0.01, ***p.adj < 0.001, Wilcoxon matched pairs tests. In A and E-G, the center line of the boxplot represents the median, box limits represent upper and lower quartiles, and whiskers represent the 1.5x interquartile range. F_ase treated samples were used. n = 34.


A
p = 0.001
R² = 0.034
JSD
d c a b
Paired OP-BALF
Inter−OPs
Inter−BALFs
Unpaired OP-BALF
BALF
OP
PCo1 (20.9%)
B
p < 0.001
Pearson'r = 0.62
count
Log10 (relative abundance of species in OP)
Log10 (relative abundance of species in BALF)
C
Mean relative abundance
0.02
0.04
0.06
−Log10 (p.adj)
BALF
OP
Log2 (fold change)
D
Log10 (bacterial DNA load)
Pneumonia
No
Yes
Sample type
BALF
OP
Pneumonia: 2
Proportion of shared species in OP (abundance%)
Proportion of shared species in BALF (abundance%)
E
Proportion of shared species (abundance%)
Non-pneumonia
Pneumonia
F
0.52
0.47
Proportion of shared species (number%)
G
Log10 (reads)
Veillonella atypica
Veillonella dispar
Enterococcus faecium
Acinetobacter baumannii
Mycoplasma pneumoniae
Sample type
BALF
OP

Supplementary figure 9. Comparison of the effectiveness of host DNA depletion methods on BALF samples. (A) Bacterial DNA retention rate. (B) Proportion of non-contaminant microbial reads in raw sequencing data. (C) The number of observed species. In A-C, the center line of the boxplot represents the median, box limits represent upper and lower quartiles, and whiskers represent the 1.5x interquartile range. Different letters on the top indicate statistically significant differences among methods (Student's t-tests followed by FDR adjustment). n = 8. (D) Similarity in microbial composition between R_ase and other methods, and proportion of contaminating microbial reads in total microbial reads. Dots indicate median values of different methods, and error bars indicate standard deviations. (E) Absolute and relative abundance change of common species. Fold changes in abundances were calculated between different methods and R_ase. Asterisks indicate significant differences in abundances between different methods and R_ase, **p.adj < 0.01, ***p.adj < 0.001, Wilcoxon matched pairs tests. n = 8. (F) Radar charts showing the performance of host DNA depletion methods on five evaluation metrics.



Reference

1. Hasan MR, Rawat A, Tang P, Jithesh PV, Thomas E, Tan R, et al. Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing. J Clin Microbiol. 2016;54(4):919-27.
2. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat Biotechnol. 2019;37(7):783-92.
3. Fong W, Rockett R, Timms V, Sintchenko V. Optimization of sample preparation for culture-independent sequencing of Bordetella pertussis. Microb Genom. 2020;6(3).
4. Kok NA, Peker N, Schuele L, de Beer JL, Rossen JWA, Sinha B, et al. Host DNA depletion can increase the sensitivity of Mycobacterium spp. detection through shotgun metagenomics in sputum. Front Microbiol. 2022;13:949328.
5. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research. 2016;44(W1):W242-5.