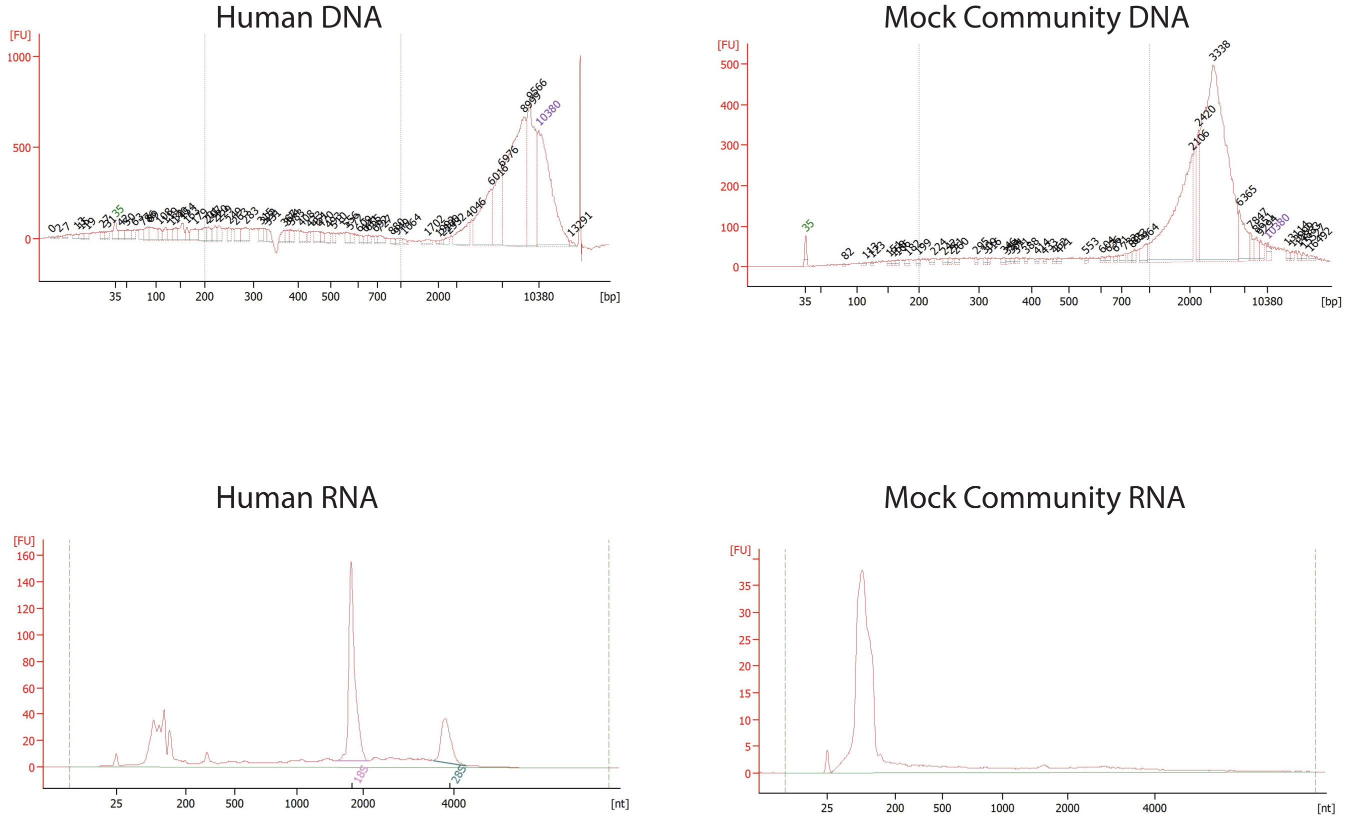
iScience, Volume *23*

Supplemental Information

sMETASeq: Combined Profiling of Microbiota and Host Small RNAs

Robin Mjelle, Kristin Roseth Aass, Wenche Sjursen, Eva Hofsli, and Pål Sætrom

Figure S1



Human DNA

Mock Community DNA

Human RNA

Mock Community RNA

Figure S2

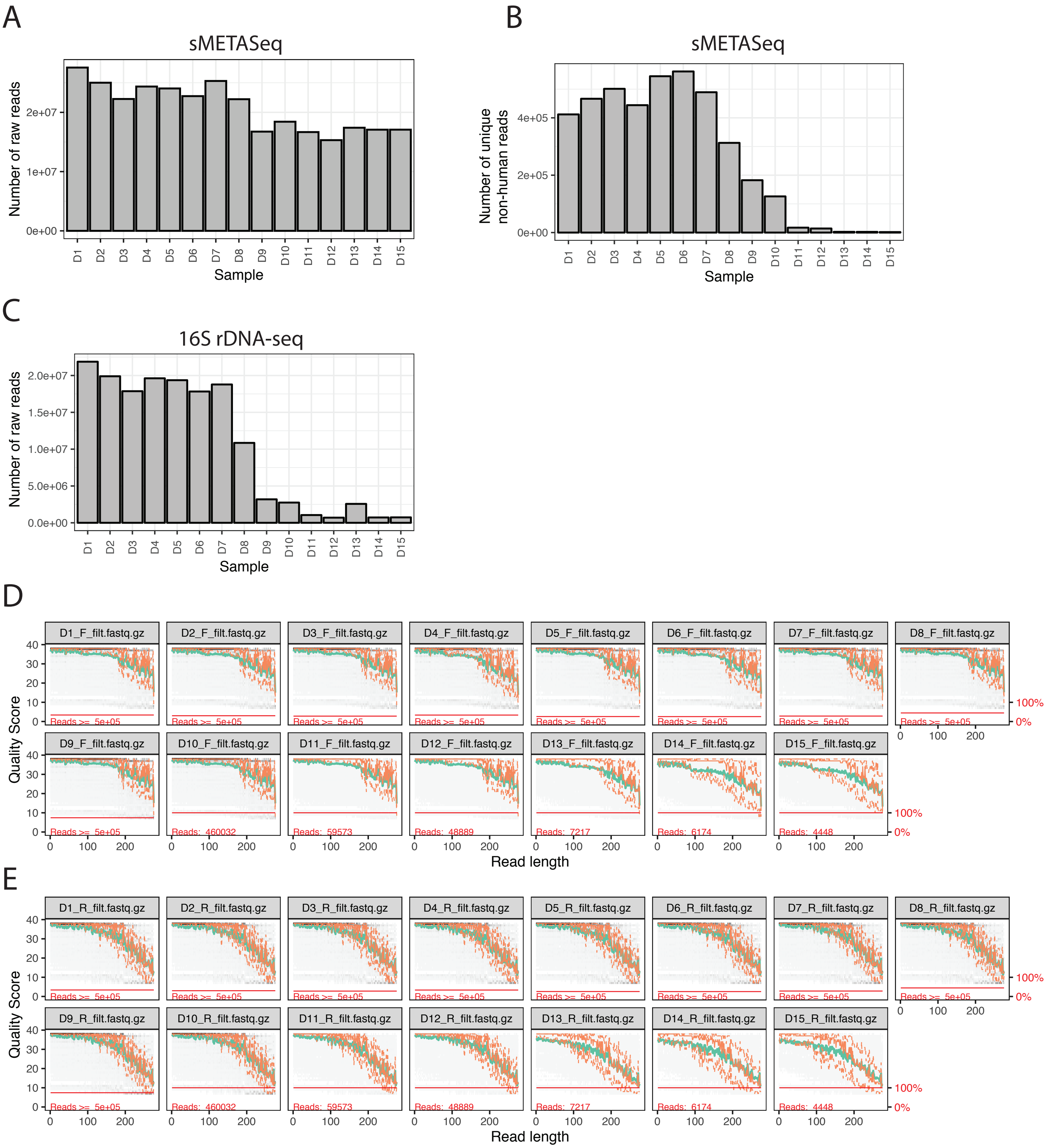

A
sMETASeq
Number of raw reads
0e+00
1e+07
2e+07
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
Sample
B
sMETASeq
Number of unique non-human reads
0e+00
2e+05
4e+05
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
Sample
C
16S rDNA-seq
Number of raw reads
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
D15
Sample
D
D1_F_filt.fastq.gz
D2_F_filt.fastq.gz
D3_F_filt.fastq.gz
D4_F_filt.fastq.gz
D5_F_filt.fastq.gz
D6_F_filt.fastq.gz
D7_F_filt.fastq.gz
D8_F_filt.fastq.gz
D9_F_filt.fastq.gz
D10_F_filt.fastq.gz
D11_F_filt.fastq.gz
D12_F_filt.fastq.gz
D13_F_filt.fastq.gz
D14_F_filt.fastq.gz
D15_F_filt.fastq.gz
Quality Score
0
10
20
30
40
Reads >= 5e+05
Reads: 460032
Reads: 59573
Reads: 48889
Reads: 7217
Reads: 6174
Reads: 4448
100%
0%
0
100
200
Read length
E
D1_R_filt.fastq.gz
D2_R_filt.fastq.gz
D3_R_filt.fastq.gz
D4_R_filt.fastq.gz
D5_R_filt.fastq.gz
D6_R_filt.fastq.gz
D7_R_filt.fastq.gz
D8_R_filt.fastq.gz
D9_R_filt.fastq.gz
D10_R_filt.fastq.gz
D11_R_filt.fastq.gz
D12_R_filt.fastq.gz
D13_R_filt.fastq.gz
D14_R_filt.fastq.gz
D15_R_filt.fastq.gz
Quality Score
0
10
20
30
40
Reads >= 5e+05
Reads: 460032
Reads: 59573
Reads: 48889
Reads: 7217
Reads: 6174
Reads: 4448
100%
0%
0
100
200
Read length

Figure S3

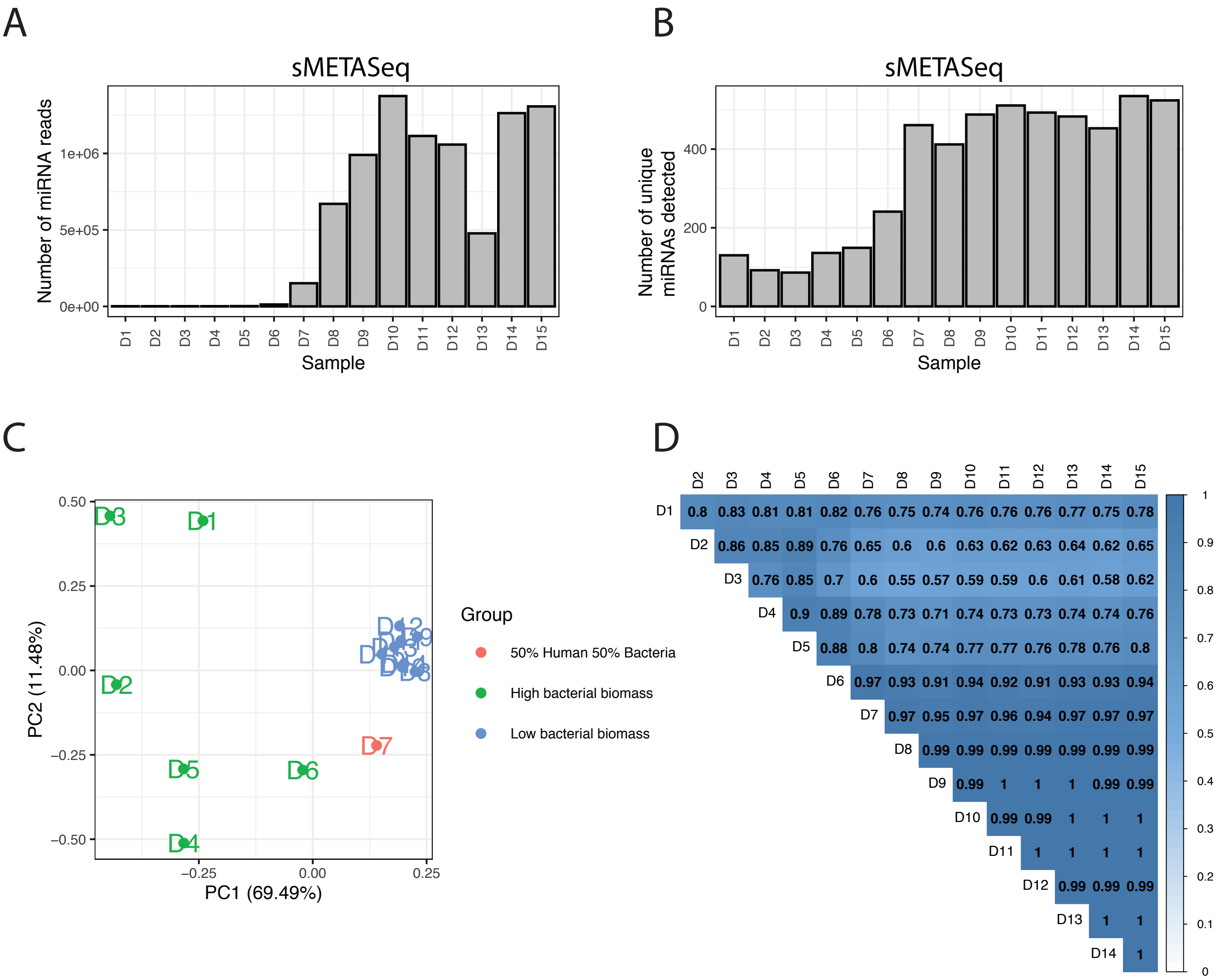

A
sMETASeq
Number of miRNA reads
0e+00
5e+05
1e+06
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
B
sMETASeq
Number of unique miRNAs detected
0
200
400
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
C
PC2 (11.48%)
PC1 (69.49%)
0.50
0.25
0.00
−0.25
−0.50
−0.25
0.00
0.25
D3
D1
D2
D5
D6
D7
D4
D12
D9
D8
Group
50% Human 50% Bacteria
High bacterial biomass
Low bacterial biomass
D
D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
D1 0.8 0.83 0.81 0.81 0.82 0.76 0.75 0.74 0.76 0.76 0.76 0.77 0.75 0.78
D2 0.86 0.85 0.89 0.76 0.65 0.6 0.6 0.63 0.62 0.63 0.64 0.62 0.65
D3 0.76 0.85 0.7 0.6 0.55 0.57 0.59 0.59 0.6 0.61 0.58 0.62
D4 0.9 0.89 0.78 0.73 0.71 0.74 0.73 0.73 0.74 0.74 0.76
D5 0.88 0.8 0.74 0.74 0.77 0.77 0.76 0.78 0.76 0.8
D6 0.97 0.93 0.91 0.94 0.92 0.91 0.93 0.93 0.94
D7 0.97 0.95 0.97 0.96 0.94 0.97 0.97 0.97
D8 0.99 0.99 0.99 0.99 0.99 0.99 0.99
D9 0.99 1 1 1 0.99 0.99
D10 0.99 0.99 1 1 1
D11 1 1 1 1
D12 0.99 0.99 0.99
D13 1 1
D14 1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure S4

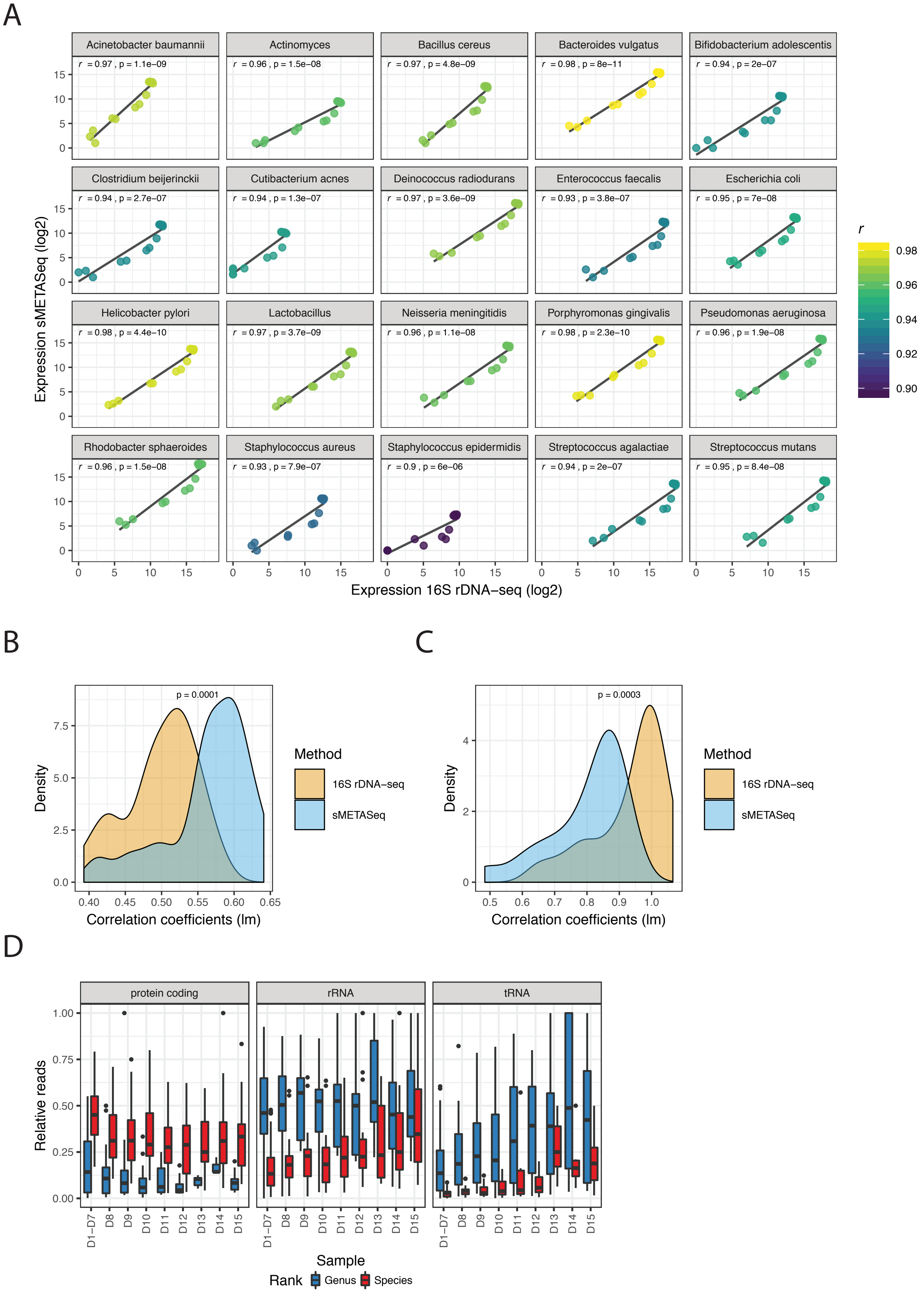

A
Acinetobacter baumannii
r = 0.97 , p = 1.1e−09
Actinomyces
r = 0.96 , p = 1.5e−08
Bacillus cereus
r = 0.97 , p = 4.8e−09
Bacteroides vulgatus
r = 0.98 , p = 8e−11
Bifidobacterium adolescentis
r = 0.94 , p = 2e−07
Clostridium beijerinckii
r = 0.94 , p = 2.7e−07
Cutibacterium acnes
r = 0.94 , p = 1.3e−07
Deinococcus radiodurans
r = 0.97 , p = 3.6e−09
Enterococcus faecalis
r = 0.93 , p = 3.8e−07
Escherichia coli
r = 0.95 , p = 7e−08
Helicobacter pylori
r = 0.98 , p = 4.4e−10
Lactobacillus
r = 0.97 , p = 3.7e−09
Neisseria meningitidis
r = 0.96 , p = 1.1e−08
Porphyromonas gingivalis
r = 0.98 , p = 2.3e−10
Pseudomonas aeruginosa
r = 0.96 , p = 1.9e−08
Rhodobacter sphaeroides
r = 0.96 , p = 1.5e−08
Staphylococcus aureus
r = 0.93 , p = 7.9e−07
Staphylococcus epidermidis
r = 0.9 , p = 6e−06
Streptococcus agalactiae
r = 0.94 , p = 2e−07
Streptococcus mutans
r = 0.95 , p = 8.4e−08
Expression sMETASeq (log2)
0
5
10
15
Expression 16S rDNA−seq (log2)
r
0.98
0.96
0.94
0.92
0.90
B
p = 0.0001
Density
0.0
2.5
5.0
7.5
0.40
0.45
0.50
0.55
0.60
0.65
Correlation coefficients (lm)
Method
16S rDNA−seq
sMETASeq
C
p = 0.0003
Density
0
2
4
0.5
0.6
0.7
0.8
0.9
1.0
Correlation coefficients (lm)
Method
16S rDNA−seq
sMETASeq
D
protein coding
rRNA
tRNA
Relative reads
0.00
0.25
0.50
0.75
1.00
D1−D7
D8
D9
D10
D11
D12
D13
D14
D15
Sample
Rank
Genus
Species

Figure S5

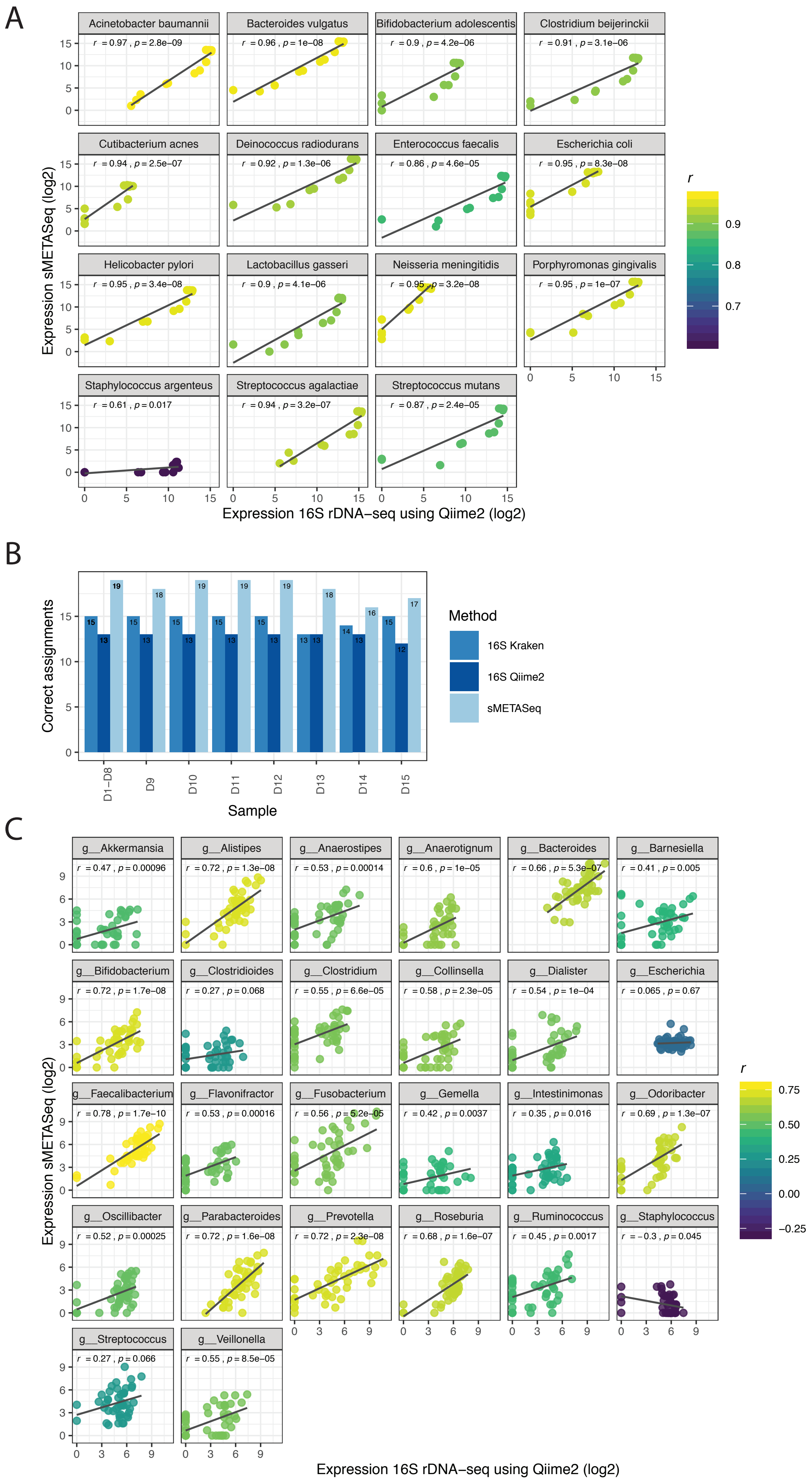

A
Acinetobacter baumannii
r = 0.97 , p = 2.8e−09
Bacteroides vulgatus
r = 0.96 , p = 1e−08
Bifidobacterium adolescentis
r = 0.9 , p = 4.2e−06
Clostridium beijerinckii
r = 0.91 , p = 3.1e−06
Cutibacterium acnes
r = 0.94 , p = 2.5e−07
Deinococcus radiodurans
r = 0.92 , p = 1.3e−06
Enterococcus faecalis
r = 0.86 , p = 4.6e−05
Escherichia coli
r = 0.95 , p = 8.3e−08
Helicobacter pylori
r = 0.95 , p = 3.4e−08
Lactobacillus gasseri
r = 0.9 , p = 4.1e−06
Neisseria meningitidis
r = 0.95 , p = 3.2e−08
Porphyromonas gingivalis
r = 0.95 , p = 1e−07
Staphylococcus argenteus
r = 0.61 , p = 0.017
Streptococcus agalactiae
r = 0.94 , p = 3.2e−07
Streptococcus mutans
r = 0.87 , p = 2.4e−05
Expression sMETASeq (log2)
Expression 16S rDNA−seq using Qiime2 (log2)
0
5
10
15
r
0.9
0.8
0.7
B
Correct assignments
0
5
10
15
Sample
D1−D8
D9
D10
D11
D12
D13
D14
D15
15
13
19
15
13
18
15
13
19
15
13
19
15
13
19
13
13
18
14
13
16
15
12
17
Method
16S Kraken
16S Qiime2
sMETASeq
C
g__Akkermansia
r = 0.47 , p = 0.00096
g__Alistipes
r = 0.72 , p = 1.3e−08
g__Anaerostipes
r = 0.53 , p = 0.00014
g__Anaerotignum
r = 0.6 , p = 1e−05
g__Bacteroides
r = 0.66 , p = 5.3e−07
g__Barnesiella
r = 0.41 , p = 0.005
g__Bifidobacterium
r = 0.72 , p = 1.7e−08
g__Clostridioides
r = 0.27 , p = 0.068
g__Clostridium
r = 0.55 , p = 6.6e−05
g__Collinsella
r = 0.58 , p = 2.3e−05
g__Dialister
r = 0.54 , p = 1e−04
g__Escherichia
r = 0.065 , p = 0.67
g__Faecalibacterium
r = 0.78 , p = 1.7e−10
g__Flavonifractor
r = 0.53 , p = 0.00016
g__Fusobacterium
r = 0.56 , p = 5.2e−05
g__Gemella
r = 0.42 , p = 0.0037
g__Intestinimonas
r = 0.35 , p = 0.016
g__Odoribacter
r = 0.69 , p = 1.3e−07
g__Oscillibacter
r = 0.52 , p = 0.00025
g__Parabacteroides
r = 0.72 , p = 1.6e−08
g__Prevotella
r = 0.72 , p = 2.3e−08
g__Roseburia
r = 0.68 , p = 1.6e−07
g__Ruminococcus
r = 0.45 , p = 0.0017
g__Staphylococcus
r = −0.3 , p = 0.045
g__Streptococcus
r = 0.27 , p = 0.066
g__Veillonella
r = 0.55 , p = 8.5e−05
Expression sMETASeq (log2)
Expression 16S rDNA−seq using Qiime2 (log2)
0
3
6
9
r
0.75
0.50
0.25
0.00
−0.25

Figure S6

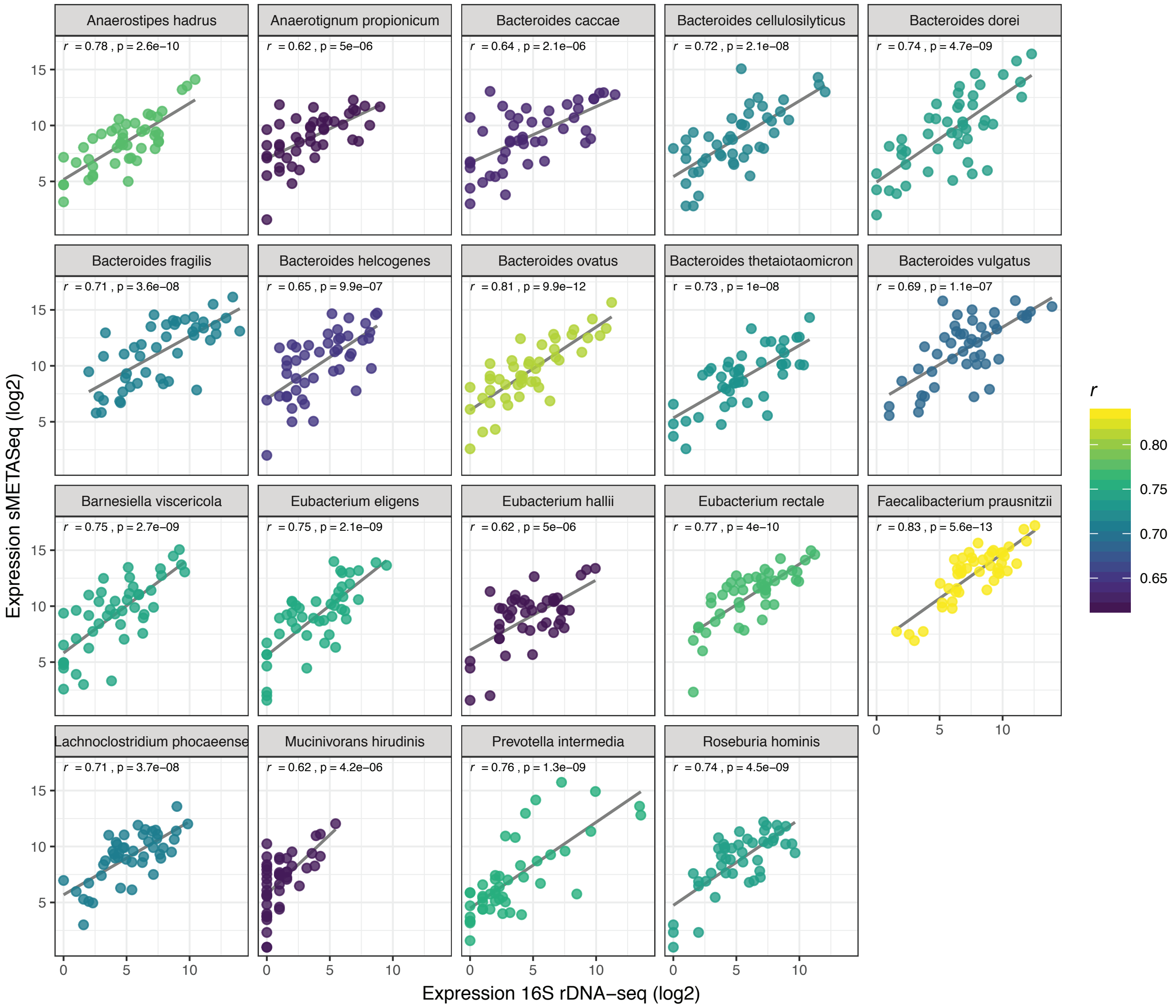

Anaerostipes hadrus
r = 0.78 , p = 2.6e−10
Anaerotignum propionicum
r = 0.62 , p = 5e−06
Bacteroides caccae
r = 0.64 , p = 2.1e−06
Bacteroides cellulosilyticus
r = 0.72 , p = 2.1e−08
Bacteroides dorei
r = 0.74 , p = 4.7e−09
Bacteroides fragilis
r = 0.71 , p = 3.6e−08
Bacteroides helcogenes
r = 0.65 , p = 9.9e−07
Bacteroides ovatus
r = 0.81 , p = 9.9e−12
Bacteroides thetaiotaomicron
r = 0.73 , p = 1e−08
Bacteroides vulgatus
r = 0.69 , p = 1.1e−07
Barnesiella viscericola
r = 0.75 , p = 2.7e−09
Eubacterium eligens
r = 0.75 , p = 2.1e−09
Eubacterium hallii
r = 0.62 , p = 5e−06
Eubacterium rectale
r = 0.77 , p = 4e−10
Faecalibacterium prausnitzii
r = 0.83 , p = 5.6e−13
Lachnoclostridium phocaeense
r = 0.71 , p = 3.7e−08
Mucinivorans hirudinis
r = 0.62 , p = 4.2e−06
Prevotella intermedia
r = 0.76 , p = 1.3e−09
Roseburia hominis
r = 0.74 , p = 4.5e−09
Expression sMETASeq (log2)
Expression 16S rDNA−seq (log2)
5
10
15
0
5
10
r
0.80
0.75
0.70
0.65

Figure S7

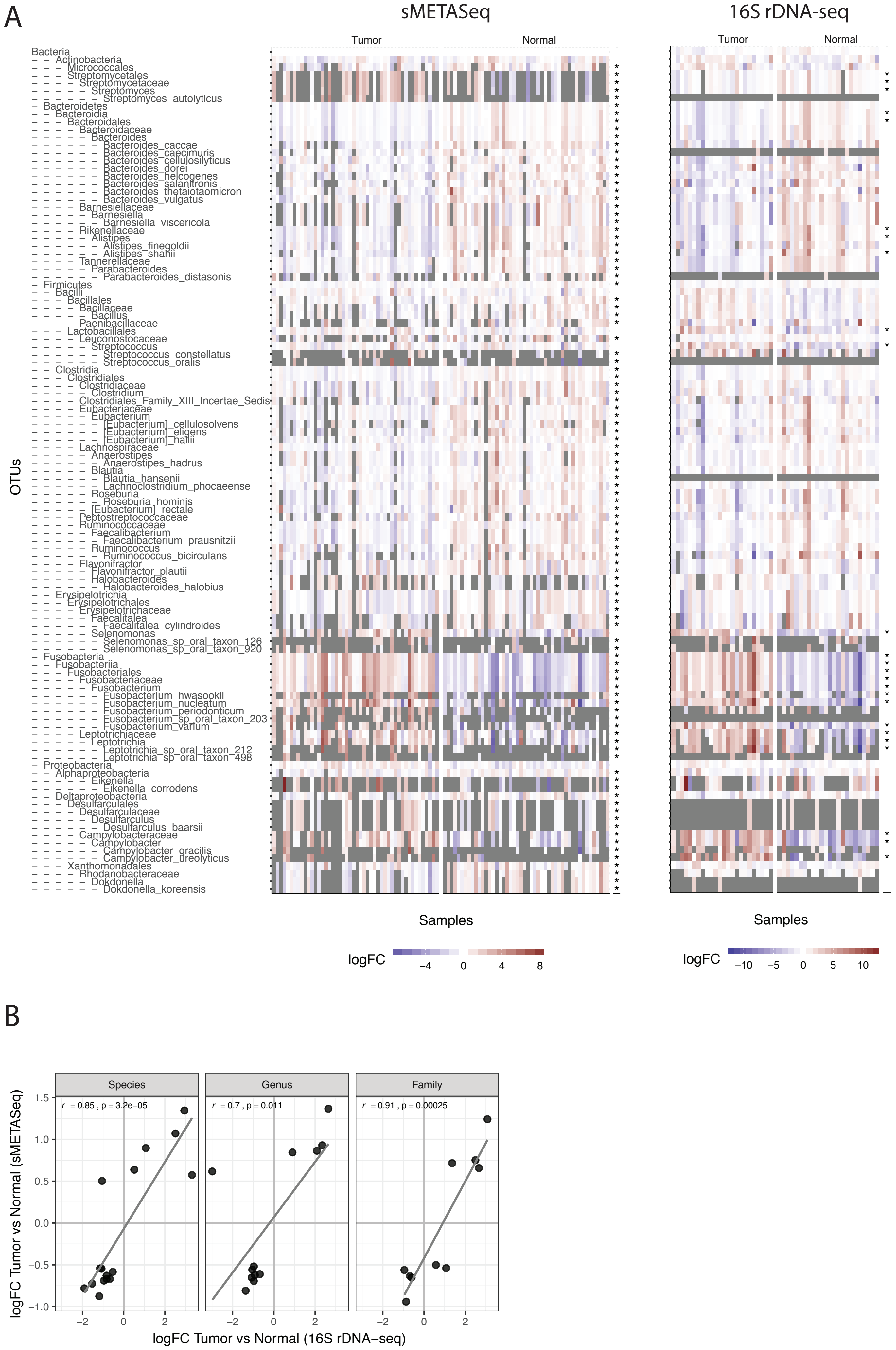
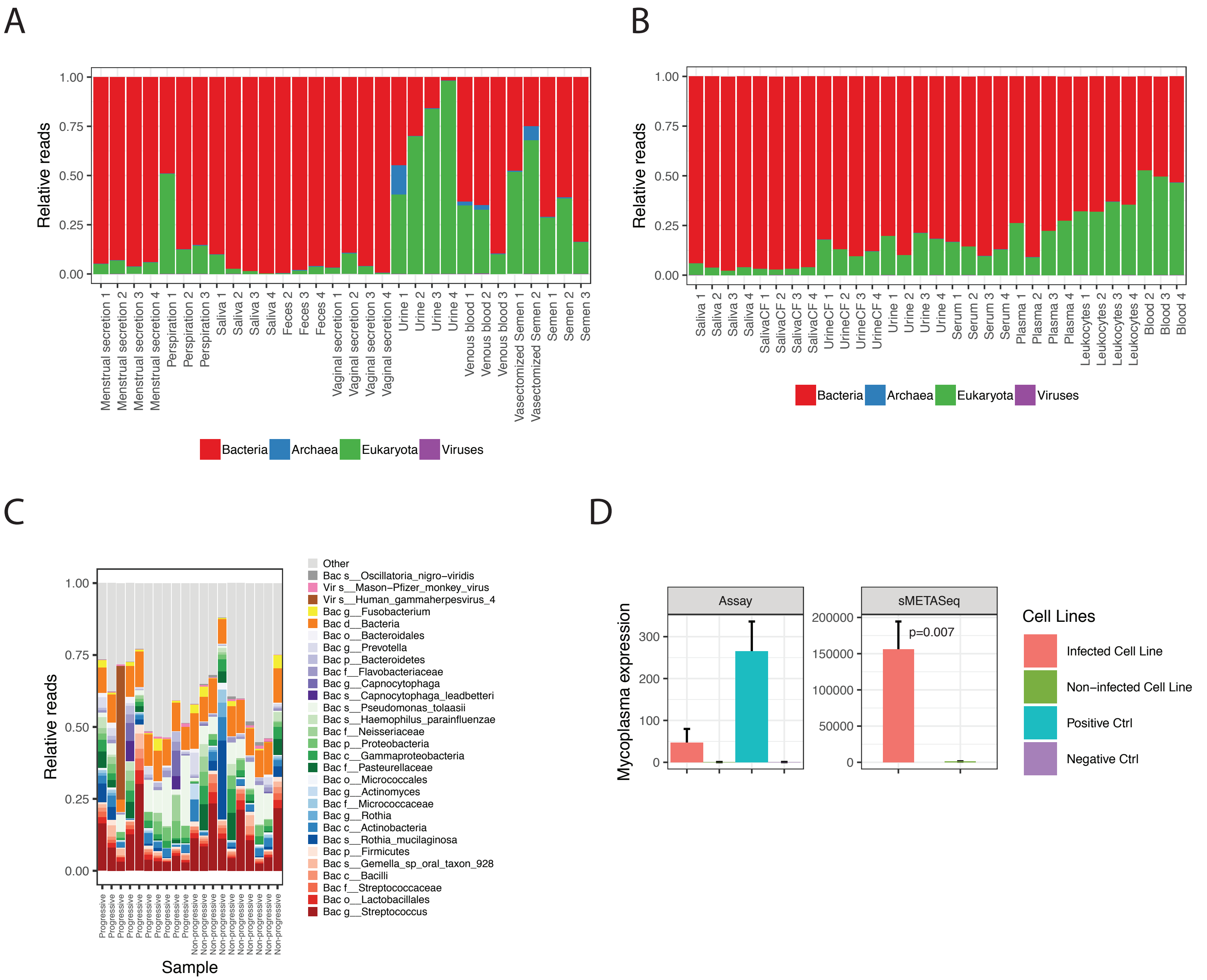

A
sMETASeq
16S rDNA-seq
Tumor
Normal
Tumor
Normal
Bacteria
– – Actinobacteria
– – – Micrococcales
– – – Streptomycetales
– – – – Streptomycetaceae
– – – – – Streptomyces
– – – – – – Streptomyces_autolyticus
– Bacteroidetes
– – Bacteroidia
– – – Bacteroidales
– – – – Bacteroidaceae
– – – – – Bacteroides
– – – – – – Bacteroides_caccae
– – – – – – Bacteroides_caecimuris
– – – – – – Bacteroides_cellulosilyticus
– – – – – – Bacteroides_dorei
– – – – – – Bacteroides_helcogenes
– – – – – – Bacteroides_salanitronis
– – – – – – Bacteroides_thetaiotaomicron
– – – – – – Bacteroides_vulgatus
– – – – Barnesiellaceae
– – – – – Barnesiella
– – – – – – Barnesiella_viscericola
– – – – Rikenellaceae
– – – – – Alistipes
– – – – – – Alistipes_finegoldii
– – – – – – Alistipes_shahii
– – – – Tannerellaceae
– – – – – Parabacteroides
– – – – – – Parabacteroides_distasonis
– Firmicutes
– – Bacilli
– – – Bacillales
– – – – Bacillaceae
– – – – – Bacillus
– – – – Paenibacillaceae
– – – Lactobacillales
– – – – Leuconostocaceae
– – – – – Streptococcus
– – – – – – Streptococcus_constellatus
– – – – – – Streptococcus_oralis
– – Clostridia
– – – Clostridiales
– – – – Clostridiaceae
– – – – – Clostridium
– – – – Clostridiales_Family_XIII_Incertae_Sedis
– – – – Eubacteriaceae
– – – – – Eubacterium
– – – – – – [Eubacterium]_cellulosolvens
– – – – – – [Eubacterium]_eligens
– – – – – – [Eubacterium]_hallii
– – – – Lachnospiraceae
– – – – – Anaerostipes
– – – – – – Anaerostipes_hadrus
– – – – – Blautia
– – – – – – Blautia_hansenii
– – – – – – Lachnoclostridium_phocaeense
– – – – – Roseburia
– – – – – – Roseburia_hominis
– – – – – [Eubacterium]_rectale
– – – – Peptostreptococcaceae
– – – – Ruminococcaceae
– – – – – Faecalibacterium
– – – – – – Faecalibacterium_prausnitzii
– – – – – Ruminococcus
– – – – – – Ruminococcus_bicirculans
– – – – Flavonifractor
– – – – – Flavonifractor_plautii
– – – – – Halobacteroides
– – – – – – Halobacteroides_halobius
– – Erysipelotrichia
– – – Erysipelotrichales
– – – – Erysipelotrichaceae
– – – – – Faecalitalea
– – – – – – Faecalitalea_cylindroides
– – – – – Selenomonas
– – – – – – Selenomonas_sp_oral_taxon_126
– – – – – – Selenomonas_sp_oral_taxon_920
– Fusobacteria
– – Fusobacteriia
– – – Fusobacteriales
– – – – Fusobacteriaceae
– – – – – Fusobacterium
– – – – – – Fusobacterium_hwasookii
– – – – – – Fusobacterium_nucleatum
– – – – – – Fusobacterium_periodonticum
– – – – – – Fusobacterium_sp_oral_taxon_203
– – – – – – Fusobacterium_varium
– – – – Leptotrichiaceae
– – – – – Leptotrichia
– – – – – – Leptotrichia_sp_oral_taxon_212
– – – – – – Leptotrichia_sp_oral_taxon_498
– Proteobacteria
– – Alphaproteobacteria
– – – – – Eikenella
– – – – – – Eikenella_corrodens
– – Deltaproteobacteria
– – – Desulfarculales
– – – – Desulfarculaceae
– – – – – Desulfarculus
– – – – – – Desulfarculus_baarsii
– – – – Campylobacteraceae
– – – – – Campylobacter
– – – – – – Campylobacter_gracilis
– – – – – – Campylobacter_ureolyticus
– – – Xanthomonadales
– – – – Rhodanobacteraceae
– – – – – Dokdonella
– – – – – – Dokdonella_koreensis
OTUs
Samples
Samples
logFC
−4
0
4
8
logFC
−10
−5
0
5
10
B
Species
Genus
Family
r = 0.85 , p = 3.2e−05
r = 0.7 , p = 0.011
r = 0.91 , p = 0.00025
logFC Tumor vs Normal (sMETASeq)
1.5
1.0
0.5
0.0
−0.5
−1.0
−2
0
2
−2
0
2
−2
0
2
logFC Tumor vs Normal (16S rDNA−seq)

Figure S8


A
Relative reads
0.00
0.25
0.50
0.75
1.00
Menstrual secretion 1
Menstrual secretion 2
Menstrual secretion 3
Menstrual secretion 4
Perspiration 1
Perspiration 2
Perspiration 3
Saliva 1
Saliva 2
Saliva 3
Saliva 4
Feces 2
Feces 3
Feces 4
Vaginal secretion 1
Vaginal secretion 2
Vaginal secretion 3
Vaginal secretion 4
Urine 1
Urine 2
Urine 3
Urine 4
Venous blood 1
Venous blood 2
Venous blood 3
Vasectomized Semen 1
Vasectomized Semen 2
Semen 1
Semen 2
Semen 3
Bacteria
Archaea
Eukaryota
Viruses
B
Relative reads
0.00
0.25
0.50
0.75
1.00
Saliva 1
Saliva 2
Saliva 3
Saliva 4
SalivaCF 1
SalivaCF 2
SalivaCF 3
SalivaCF 4
UrineCF 1
UrineCF 2
UrineCF 3
UrineCF 4
Urine 1
Urine 2
Urine 3
Urine 4
Serum 1
Serum 2
Serum 3
Serum 4
Plasma 1
Plasma 2
Plasma 3
Plasma 4
Leukocytes 1
Leukocytes 2
Leukocytes 3
Leukocytes 4
Blood 2
Blood 3
Blood 4
Bacteria
Archaea
Eukaryota
Viruses
C
Relative reads
0.00
0.25
0.50
0.75
1.00
Progressive
Non-progressive
Sample
Other
Bac s__Oscillatoria_nigro−viridis
Vir s__Mason−Pfizer_monkey_virus
Vir s__Human_gammaherpesvirus_4
Bac g__Fusobacterium
Bac d__Bacteria
Bac o__Bacteroidales
Bac g__Prevotella
Bac p__Bacteroidetes
Bac f__Flavobacteriaceae
Bac g__Capnocytophaga
Bac s__Capnocytophaga_leadbetteri
Bac s__Pseudomonas_tolaasii
Bac s__Haemophilus_parainfluenzae
Bac f__Neisseriaceae
Bac p__Proteobacteria
Bac c__Gammaproteobacteria
Bac f__Pasteurellaceae
Bac o__Micrococcales
Bac g__Actinomyces
Bac f__Micrococcaceae
Bac g__Rothia
Bac c__Actinobacteria
Bac s__Rothia_mucilaginosa
Bac p__Firmicutes
Bac s__Gemella_sp_oral_taxon_928
Bac c__Bacilli
Bac f__Streptococcaceae
Bac o__Lactobacillales
Bac g__Streptococcus
D
Mycoplasma expression
Assay
0
100
200
300
sMETASeq
0
50000
100000
150000
200000
p=0.007
Cell Lines
Infected Cell Line
Non−infected Cell Line
Positive Ctrl
Negative Ctrl

Supplemental information

Figure S1: Bioanalyzer trace of the isolated DNA and RNA from the mock community. Related to Figure 1. The length of the fragments is shown on the x-axis and the fluorescence unit (FU) are shown on the y-axis. For the RNA, the ribosomal genes are depicted if they are detected.

Figure S2: Sequencing statistics for the mock community. Related to Figure 1. A) Number of raw reads identified by sMETASeq. **B)** Number of unique non-human reads identified by sMETASeq after first aligning to the human genome. **C)** Number of raw reads identified by 16S rDNA-Seq. **D)** Quality profiles for the 16S rDNA-Seq data. Shown is the frequency of each quality score at each base position for the forward reads. The median quality score at each position is shown by the green line, and the quartiles of the quality score distribution by the orange lines. The red line shows the scaled proportion of reads that extend to at least that position (since we filtered the same number of bases for all samples, the line is flat for all samples). The x-axis shows the length of the reads. The number of reads post-filtering is shown in red within each sample. **E)** Same is in D for the reverse-reads.

Figure S3: Overview of miRNA data from the mock community. Related to Figure 1. A) Shown is the number of miRNA-reads detected by sMETASeq across all dilutions. **B)** Shown is the number of unique miRNAs detected across all dilutions. **C)** Principal component analysis plot of normalized (cpm, log2) miRNA count. The samples are colored based on the dilution for which samples with high bacterial biomass (D1-D6) are in green, samples with equal human/bacteria (D7) are in red, and low samples with low bacterial biomass (D8-D15) are in blue. The percentage variation explained by the two first components are indicated on the corresponding axes. **D)** Pearson correlation of miRNA expression between samples. We calculated the correlation of the log2-normalized miRNA count matrix which was first filtered to contain only miRNAs that was expressed with at least 1 cpm in all samples. The correlation values were calculated in R using the function *cor*.

Figure S4: Correlation between sMETASeq and 16S DNA-seq in mock community. Related to Figure 1. A) Comparison of expression values for the 20 mock species across all dilutions between sMETASeq and 16S DNA-seq. The correlation values are Pearson's

correlation. **B)** Density plot of the correlation values (Spearman's) between the 20 mock species and the input bacterial biomass (ng) for sMETASeq and 16S DNA-seq across all 15 dilutions. The correlations are calculated by the linear model lm() in R. The p-value indicated the difference in correlation values for sMETASeq and 16S DNA-seq and is calculated using Wilcoxon rank sum test in R. **C)** Similar as in B) for samples D8-D15. **D)** Shown is the relative abundance of protein coding RNAs, rRNAs and tRNAs across dilution for reads assigning to specific species strains and for reads assigning to the genus level. The boxplots comprise the reads for the mock species (red) and the genera for the mock species (blue).

Figure S5: Correlation between sMETASeq and 16S rDNA-seq using Qiime2. Related to Figure 1. A) Comparison of expression values across all dilutions for the species detected by both sMETASeq and Qiime2 using the GTDB database. The correlation values are Pearson's correlation. Some OTUs were not detected at the species level using Qiime2 (*Pseudomonas aeruginosa*; *Staphylococcus epidermidis*; *Bacillus_cereus; Rhodobacter sphaeroides*) or sMETASeq (*Actinomyces odontolyticus*), and are therefore not shown. **B)** Number of correctly assigned mock species across dilutions for sMETASeq and 16S rDNA-seq run through kraken and Qiime2 using the GTDB database. Dilutions D1-D8 have the same number of correct assignments and are therefore pooled. **C)** Shown is the most highly correlated bacteria genera in colon tissue between sMETASeq and 16S rDNA-seq run through Qiime2 using the GTDB database.

Figure S6: Correlation of bacteria species in colon tissue between sMETASeq and 16S rDNA-seq. Related to Figure 3. Shown is the most highly correlated bacteria species (R>0.6, Pearson's correlation) between sMETASeq and 16S DNA-seq in colon tissue.

Figure S7: Comparison of sMETASeq and 16S DNA-seq in colon tissue. Related to Figure 3. A) Heatmap showing expression of OTUs as identified by sMETASeq (left panel) and 16S DNA-seq (right panel) in tumor and normal colon tissue. The y-axis shows OTUs at different levels and the x-axis shows samples indicated with "T" for tumor samples and "N" for normal samples. The comparison is tumor vs normal such that red indicates higher levels of bacteria in tumor compared to normal and blue indicates lower levels in tumor compared to normal. Asterisk indicate that the OUT is significantly differentially expressed between tumor and normal samples. **B)** Comparison of logFC values for the difference between tumor and

normal samples between sMETASeq and 16S DNA-seq as determined by *limma*. Shown is OTUs with absolute logFC values above 0.5. Shown is OTUs at species, genus and family level. The correlation values are Pearson's correlation calculated in R. See supplementary tables for complete list of differentially expressed OTUs.

Figure S8: Alignment results for human biofluids. Related to Figure 4. A) Shown is the relative abundance of reads aligning to the three domains of life in addition to viruses in the dataset of Seashols-Williams et al. **B)** Similar as in A) for the dataset of El-Mogy et al. **C)** Distribution of bacteria in samples from oral leukoplakia (Philipone et al.) as identified by sMETASeq. See Figure 4A for a description of the plot. **D)** Quantification of Mycoplasma bacterium in sMETASeq and MycoAlert Mycoplasma Detection Kit (Lonza). "Positive Ctrl" and "Negative Ctrl" are the controls in the Lonza kit; "Infected" is a mycoplasma-infected cell-line; "Non-infected" is a non-infected cell line. The y-axis for the mycoplasma assay is the readout from the MycoAlert test. Mycoplasma contamination is indicated if the readout has a value > 1.2. The sequencing reads are shown as raw reads and the p-value was calculated using a one-tailed Student's t-test on cpm-log2-normalized values. The standard deviation is calculated from two biological replicates.

Transparent Methods

Analysis pipeline for 16S data using Qiime2

A detailed procedure on how Qiime2 was run can be found below. In short, the data were filtered using dada2 with the parameters --p-trunc-len-f 290 and --p-trunc-len-r 290. Next, we used the “bac_120” fasta file from GTDB to generate a feature classifier using the primers CCTACGGGNGGCWGCAG and GACTACHVGGGTATCTAATCC, which corresponds to the region amplified for our data. Using this classifier, the data was analyzed using the function *feature-classifier classify-sklearn* with parameter --p-confidence 0.1, otherwise default parameters. The 16S data were analyzed by Qiime2 using the following scripts and parameters:

```
qiime tools import
--type 'SampleData[PairedEndSequencesWithQuality]'
--input-format PairedEndFastqManifestPhred33V2
--input-path ./manifestPE2.tsv
--output-path ./demux_seqsPE.qza

qiime dada2 denoise-paired \
  --i-demultiplexed-seqs ./demux_seqsPE.qza \
  --p-trunc-len-f 290 \
  --p-trunc-len-r 290 \
  --p-n-threads 20 \
  --o-table ./dada2_tablePE.qza \
  --o-representative-sequences ./dada2_rep_set_PE.qza \
  --o-denoising-stats ./dada2_stats_PE.qza

qiime tools import \
 --input-path ./bac120_ssu.fna \
 --output-path ./bac120_ssu.qza \
 --type 'FeatureData[Sequence]'

qiime tools import \
 --input-path bac120_taxonomy.tsv \
 --output-path ./bac120_taxonomy.tsv.qza \
 --type 'FeatureData[Taxonomy]' \
 --input-format HeaderlessTSVTaxonomyFormat
```

```
qiime feature-classifier extract-reads \
 --i-sequences ./bac120_ssu.qza \
 --p-f-primer CCTACGGGNGGCWGCAG \
 --p-r-primer GACTACHVGGGTATCTAATCC \
 --p-trunc-len 450 \
 --p-min-length 100 \
 --p-max-length 600 \
 --o-reads ./bac120_ssu_341_805.qza

qiime feature-classifier fit-classifier-naive-bayes \
 --i-reference-reads ./bac120_ssu_341_805.qza \
 --i-reference-taxonomy ./bac120_taxonomy.tsv.qza \
 --o-classifier ./bac120_ssu_341_805_classifer.qza

qiime feature-classifier classify-sklearn \
  --i-classifier bac120_ssu_341_805_classifer.qza \
  --i-reads dada2_rep_set_PE.qza \
  --p-confidence 0.1 \
  --o-classification bac120_ssu_341_805_confidence0.1_PE.qza

qiime metadata tabulate \
  --m-input-file bac120_ssu_341_805_confidence0.1_PE.qza \
  --o-visualization bac120_ssu_341_805_confidence0.1_PE.qzv

qiime feature-table filter-samples \
  --i-table ./dada2_tablePE.qza \
  --p-min-frequency 100 \
  --o-filtered-table ./table_2k_PE.qza

qiime taxa barplot \
  --i-table ./table_2k_PE.qza \
  --i-taxonomy ./bac120_ssu_341_805_confidence0.1_PE.qza \
  --m-metadata-file ./metadata.tsv \
  --o-visualization ./barplot_bac120_ssu_341_805_confidence0.1_PE.qzv
```

Overview of public datasets

The sRNA-seq dataset from colon cancer tissue is described in (Mjelle et al., 2019). The human biofluid datasets are described in (El-Mogy et al., 2018; Seashols-Williams et al., 2016). The cervix dataset is described in (Snoek et al., 2018). The oral leukoplakia dataset is described in (Philipone et al., 2016).

DNA isolation and 16S rDNA-seq on colon tissue

16S rDNA-seq was performed on 48 samples from 24 colon cancer patients all of which were included in the sRNA-seq. DNA was isolated from frozen tissue samples using the DNeasy Blood & Tissue Kits from Qiagen (Cat No./ID: 69504). The DNA was normalized to equal concentration and used as input in the 16S Ribosomal RNA Gene Amplicons and sequenced on the Illumina MiSeq System using 300bp paired end reads. PCR primers (5µl (1 µM) pr. sample) was ordered from Invitrogen based on the 16S rDNA-seq protocol from Illumina. We used the following primers:

"16S Amplicon PCR Forward Primer:
“5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG”

16S Amplicon PCR Reverse Primer: 5'
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC
"

For indexing of the samples, we used the Nextera XT Index Kit v2, set D, FC-131-2004. The gene specific sequences used in this protocol target the 16S V3 and V4 region. They are selected from the Klindworth et al. publication (Klindworth et al., 2013).

16S rDNA-seq analysis using kraken

Quality analysis and filtering of the raw reads were performed using DADA2 (Callahan et al., 2016). The reads were filtered in DADA2 using the function *filterAndTrim* with the parameters: trimRight=c(0,0),trimLeft=c(25,25), maxN=0, maxEE=Inf, truncQ=1, rm.phix=TRUE. The DADA2-filtered reads were used as input to kraken1.0 for taxonomic classification using these two commands: *Kraken --db database Sample_forward. fastq.gz*

Sample_reverse.fastq.gz > *Sample.kraken.stderr* and *kraken-mpa-report --db database Sample.kraken.stderr* > *Sample.kraken.report*. Kraken is previously shown to perform good on long 16S reads (Valenzuela-Gonzalez et al., 2016).

DNA and RNA isolation from mock community

We used the "20 Strain Even Mix Whole Cell Material (ATCC® MSA-2002™)" from ATCC as mock community. Bacterial DNA was isolated using the DNeasy Blood & Tissue Kits from Qiagen, following the protocol of the kit (Cat No./ID: 69504). Bacterial RNA was isolated using miRVana RNA isolation (Cat No. AM1560).

16S rDNA-seq and sRNA-seq of mock community.

The 16S rDNA-seq of the mock community was performed as for the colon tissue samples described above. The sRNA-seq was performed using "NEXTFLEX® Small RNA-Seq Kit v3 for Illumina" using 16 PCR cycles. The input material for the NEXTFLEX protocol was the output from miRVana without further size selection. The finished libraries were gel-purified using automated gel purification aiming for RNA fragments of approximately 15-200nts in length. The sRNA libraries were sequenced on a HiSeq 4000 from Illumina using 75bp single reads.

Overview of the sMETASeq pipeline

We here describe how sRNA-seq data can be used to identify non-human RNA species. Sequencing adapters were removed from the raw fastq files by cutadapt (v2.7) (Martin) using the parameters *cutadapt -f fastq -a*. The cut reads were collapsed into unique reads using *fastx_collapser (FASTX-Toolkit)* and aligned to the human genome (GRCh38) using bowtie2 (Langmead and Salzberg, 2012) with the parameters *bowtie2 -p20 -k10* and the file with the mapped reads was saved as *Mapped.sam.* Human microRNAs were identified using htseq-count (v0.11.1) (Anders et al., 2015) with the miRbase (v21) reference GFF file using the parameters *htseq-count -a 0 -s yes -i Name -t miRNA*. The reads from bowtie2 that did not align to the human genome were saved in a separated file called *Unmapped.fastq*. The files containing unique unaligned reads (*Unmapped.fastq*) were used as input in the metagenomic pipeline Kraken (v1.0) using the 50gb pre-build index using the following two kraken-scripts: *kraken --db database Sample.fasta* > *Sample.kraken.stderr* and *kraken-mpa-report --db database Sample.kraken.stderr* > *Sample.kraken.report*. Each sequence is assigned an appropriate label based on the lowest common ancestor from the Kraken k-mer database and a classification tree is generated which can be used as input in a statistical software like R for statistical analyses.

Mycoplasma testing

Mycoplasma infection was tested using the MycoAlert Mycoplasma Detection Kit (Lonza, Cat: LT07-218) with three replicates. The MycoAlert ratio was calculated by dividing Read B by Read A. Cells which are infected with mycoplasma will produce ratios greater than 1.

Cell culturing, RNA isolation and sequencing of mycoplasma infected cells

The JJN-3 myeloma cell line was cultured at 37℃ in a humidified atmosphere containing 5 % CO_2, in RPMI 1640 medium (Sigma Aldrich) supplemented with glutamine (100 µg/ml, Sigma Aldrich), gentamicin/gensumycin (20 µg/mL, Sanofi) and 10% fetal calf serum (Gibco/Invitrogen). The cells were split twice a week. RNA was isolated using miRVana (Thermo Fisher, cat: AM1560). Small RNA-seq was performed using the NEXTflex small

RNA library preparation kit (Bio-Scientific, Cat: NOVA-5132-05), following the manufacturer’s protocol, and sequenced using on a HiSeq 4000 Flowcell from Ilumina using 75bp single reads. The data was processed as described above.

Statistics and diversity measurements

To correlate expression values against bacteria concentrations we used the lm() function in R and extracted the estimated coefficients from the result summary. The statistical differences in estimated coefficients was evaluated using Wilcoxon rank sum test in R. Pearson's correlation coefficients were used when comparing expression values between 16S rDNA-seq and sMETASeq. Differentially expressed bacteria between tumor and normal samples were detected using *limma-voom* in R (v3.6.1), and p-values were adjusted using Benjamini-Hochberg. For both sMETASeq and 16S rDNA-seq, diversity and richness in the mock community experiment was calculated using the *vegan* (v2.5-6) R package using the *rrarefy* function using taxonomical counts from the kraken alignments. The following functions were used: Diversity was calculated using the function *diversity* with the parameters “simpson” or “shannon”; Fisher’s alpha was calculated using the function *fisher.alpha;* Species richness was calculated using the function *specnumber;* Pielou's evenness was calculated by H/log(S) where H' is Shannon diversity and S is the total number of species in a the sample (*specnumber*). The kraken output files (.report) containing the taxonomical counts were used as input for the diversity analyses. Contaminant reads were identified using the *decontam* (v1.4.0*)* package in R with the “*frequency*” method.

References

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J., and Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods *13*, 581-583.

El-Mogy, M., Lam, B., Haj-Ahmad, T.A., McGowan, S., Yu, D., Nosal, L., Rghei, N., Roberts, P., and Haj-Ahmad, Y. (2018). Diversity and signature of small RNA in different bodily fluids using next generation sequencing. BMC Genomics *19*, 408.

FASTX-Toolkit.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glockner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res *41*, e1.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

Mjelle, R., Sjursen, W., Thommesen, L., Saetrom, P., and Hofsli, E. (2019). Small RNA expression from viruses, bacteria and human miRNAs in colon cancer tissue and its association with microsatellite instability and tumor location. BMC Cancer *19*, 161.

Philipone, E., Yoon, A.J., Wang, S., Shen, J., Ko, Y.C., Sink, J.M., Rockafellow, A., Shammay, N.A., and Santella, R.M. (2016). MicroRNAs-208b-3p, 204-5p, 129-2-3p and 3065-5p as predictive markers of oral leukoplakia that progress to cancer. Am J Cancer Res *6*, 1537-1546.

Seashols-Williams, S., Lewis, C., Calloway, C., Peace, N., Harrison, A., Hayes-Nash, C., Fleming, S., Wu, Q., and Zehner, Z.E. (2016). High-throughput miRNA sequencing and identification of biomarkers for forensically relevant biological fluids. Electrophoresis *37*, 2780-2788.

Snoek, B.C., Verlaat, W., Babion, I., Novianti, P.W., van de Wiel, M.A., Wilting, S.M., van Trommel, N.E., Bleeker, M.C.G., Massuger, L., Melchers, W.J.G., et al. (2018). Genome-wide microRNA analysis of HPV-positive self-samples yields novel triage markers for early detection of cervical cancer. Int J Cancer.

Valenzuela-Gonzalez, F., Martinez-Porchas, M., Villalpando-Canchola, E., and Vargas-Albores, F. (2016). Studying long 16S rDNA sequences with ultrafast-metagenomic sequence classification using exact alignments (Kraken). J Microbiol Methods *122*, 38-42.