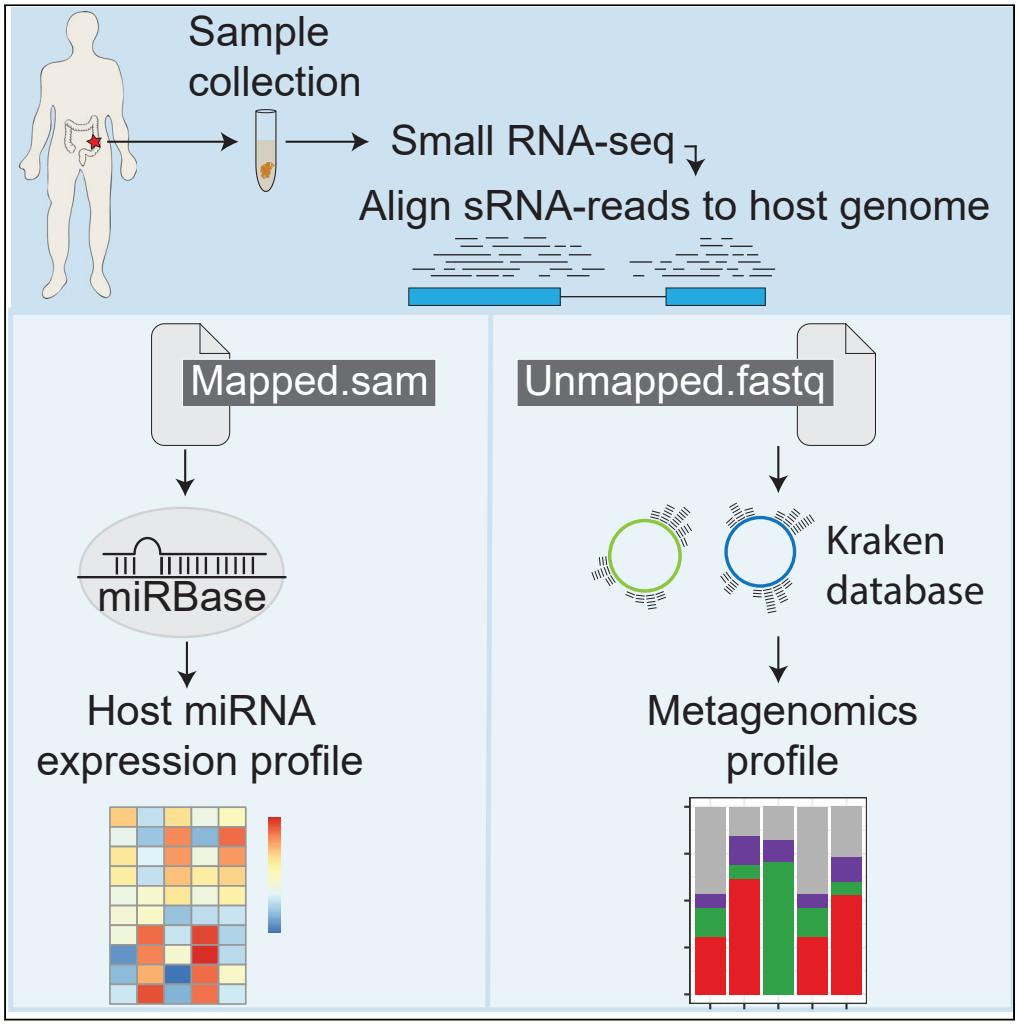
iScience

CellPress
OPEN ACCESS

Article

sMETASeq: Combined Profiling of Microbiota and Host Small RNAs



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

robin.mjelle@ntnu.no

HIGHLIGHTS

Our method "sMETASeq" generates metagenomic profiles from small RNA sequencing data

sMETASeq jointly profiles host and microbe small RNAs in human samples

sMETASeq measures and detects changes in abundance of microbes at species level

sMETASeq is available as open source scripts

Mjelle et al., iScience *23*, 101131
May 22, 2020
https://doi.org/10.1016/j.isci.2020.101131

iScience

CellPress
OPEN ACCESS

Article

sMETASeq: Combined Profiling of Microbiota and Host Small RNAs

Robin Mjelle,[1,6,7,8,*] Kristin Roseth Aass,[1,2] Wenche Sjursen,[1,3] Eva Hofsli,[1,4] and Pål Sætrom[1,5,6,7]

SUMMARY

Understanding microbial communities' roles in human health and disease requires methods that accurately characterize the microbial composition and their activity and effects within human biological samples. We present sMETASeq (small RNA Metagenomics by Sequencing), a novel method that uses sequencing of small RNAs to jointly measure host small RNA expression and create metagenomic profiles and detect small bacterial RNAs. We evaluated the performance of sMETASeq on a mock bacterial community and demonstrated its use on different human samples, including colon cancer, oral leukoplakia, cervix cancer, and a panel of human biofluids. In all datasets, the detected microbes reflected the biology of the different sample types.

INTRODUCTION

Small RNA sequencing (sRNA-seq) has traditionally been a sequencing method for quantifying microRNAs (miRNAs), but increased understanding of small RNAs and improved databases have enabled identification of other small RNA classes such as transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and other small RNAs. Current sRNA-seq protocols require 3′ hydroxyl- and 5′ phosphate groups on the RNA for adapter ligation, whereas subsequent size selection usually enriches for RNAs approximately 22 nucleotides (nts) in length (Pritchard et al., 2012). Several classes of RNA meet these criteria and will therefore be part of the final sequencing library.

Detection and quantification of microbes currently rely on either 16S rDNA-seq, utilizing variable regions within the 16S ribosomal RNA (rRNA) gene (Hamady and Knight, 2009), or shotgun DNA-seq in which the DNA is randomly fragmented and sequenced (Venter et al., 2004). The 16S rDNA-seq method has been the gold-standard for metagenomics owing to its good sensitivity and specificity and relatively low cost. However, 16S rDNA-seq has some limitations including underrepresentation of species owing to primer mismatches (Schulz et al., 2017) and low phylogenetic power due to high DNA sequence similarity of the 16S rRNA genes (Janda and Abbott, 2007). Fungi are usually detected by sequencing the Internal Transcribed Spacer (ITS) (Pankaj, 2013; Schoch et al., 2012) and require a separate primer set than that for 16S rDNA-seq. Viruses are commonly detected using customized oligonucleotide capture probes (O'Flaherty et al., 2018) or ribo-depleted total RNA-seq (Visser et al., 2016), but they can also be detected using small RNA-seq (Massart et al., 2019). Shotgun DNA-seq has the advantage over 16S rDNA-seq in that it can detect other microbes than bacteria, has a higher species specificity than 16S rDNA-seq, and can assemble whole genes and infer gene function (Quince et al., 2017).

Small RNAs have been identified in bacteria and shown to play regulatory roles (Majdalani et al., 2005). Bacterial sRNAs are between 50 and 500 nts long and can be detected using total RNA-seq protocols, which have a bias against RNAs shorter than 50 nts. Bacterial sRNAs resemble eukaryotic miRNAs in their ability to base pair with target RNAs; however, they do not undergo a biogenesis pathway similar to that of miRNAs (Gottesman and Storz, 2011). Moreover, the base pairing usually occurs at the 5′end of the target RNA. The number of bacterial sRNAs varies between species, but the number is likely much smaller than in eukaryotes (Gottesman and Storz, 2011), although the identification has lagged that of eukaryotes because fewer sequencing studies have been performed. Some bacteria express tRNA-derived RNA fragments (tRFs) (Kumar et al., 2014) and yRNAs (Chen et al., 2014), two types of sRNAs frequently found in humans.

Here we present a novel metagenomic method, sMETASeq (small RNA Metagenomics by Sequencing), that can jointly measure host small RNAs and generate a metagenomics profile from the same sample.

[1]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, NTNU, Trondheim 7030, Norway

[2]Centre of Molecular Inflammation Research, Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim 7030, Norway

[3]Department of Medical Genetics, St. Olavs Hospital, Trondheim 7030, Norway

[4]The Cancer Clinic, St. Olavs Hospital, Trondheim University Hospital, Trondheim 7030, Norway

[5]Department of Computer and Information Science, Norwegian University of Science and Technology, NTNU, Trondheim 7491, Norway

[6]Bioinformatics Core Facility-BioCore, Norwegian University of Science and Technology, NTNU, Trondheim 7491, Norway

[7]K.G. Jebsen Center for Genetic Epidemiology, Norwegian University of Science and Technology, NTNU, Trondheim 7491, Norway

[8]Lead Contact

*Correspondence: robin.mjelle@ntnu.no

https://doi.org/10.1016/j.isci.2020.101131

We evaluated the performance of the method together with 16S rDNA-seq on a mock bacterial community and showed that sMETASeq has high sensitivity, specificity, and quantitative performance. We further show that sMETASeq detects differentially expressed microbes in colon cancer and oral leukoplakia and characterizes bacteria and other microbes in human biofluids and cervix samples that reflect the sample type of origin.

RESULTS

Overview of the sMETASeq Pipeline

The method sMETASeq was developed to enable microbiome characterization using sRNA-seq from samples containing both host small RNAs (e.g., microRNAs) and microbes, for instance, human gut biopsies or biofluids. The data for sMETASeq are generated using a standard sRNA-seq wet-lab protocol and can therefore be applied to already generated publicly available sRNA data. First, adapter-trimmed and collapsed sRNA sequencing reads are mapped to the human genome to identify mapped and unmapped human reads. The mapped reads are then compared with available database annotations of human miRNAs, for instance, miRBase (Griffiths-Jones et al., 2006), and used to generate expression profiles for miRNAs, and potentially other small RNAs, by counting the number of uncollapsed reads that map to each gene. The unique unmapped reads are further aligned to the kraken microbiome reference database (Wood and Salzberg, 2014). The microbiome alignment results are then used to generate metagenomics profiles and estimates of relative and absolute expression of microbes and microbial sRNAs.

Experimental Setup and Sequencing Statistics

To assess the performance of sMETASeq in identifying and quantifying microbes, we applied the method on a mock microbial community that had undergone serial dilution and compared the results with the widely used 16S rDNA-seq method. The mock community comprised 20 known bacterial species with a 5% contribution from each species. To better mimic a host-microbe environment, microbial DNA/RNA from the mock community was mixed with DNA/RNA from the human plasma cell line INA-6 at different concentrations (Table S1 and Figure S1). Since the species of the mock community is known, this approach allowed us to evaluate the sensitivity, specificity, and quantitative abilities of sMETASeq across the different dilutions. The experiment consisted of 15 dilutions for which samples D1–D6 contained increasing amounts of human DNA/RNA and samples D8–D15 contained decreasing amounts of bacterial DNA/RNA (Table S1). The amounts of human and bacterial DNA/RNA were fixed for samples D8–D15 and D1–D6, respectively; sample D7 contained equal amounts of human and bacterial DNA/RNA.

For sMETASeq, an average of 20 million reads per sample were generated (Figure S2A) of which on average 8 million reads mapped to the human genome and 10.5 million did not map to the human genome (Figure S2B). For 16S rDNA-seq, 135,980 reads were generated per sample on average, none of which mapped to the human genome (Figure S2C). About 1% of the reads were remove after quality filtering (see Methods) (Figures S2D and S2E). As expected, both methods showed a decrease in the number of non-human reads in samples with decreasing bacterial DNA/RNA (Figures S2B and S2C). The filtered 16S rDNA data were run through kraken to enable a direct comparison between the two methods. We also ran the 16S rDNA data through Qiime2 (Bolyen et al., 2019), to evaluate its performance on a well-established microbiome platform using a different reference database.

Effect of Bacterial RNA on miRNA Expression

We investigated the effect on miRNA expression of having bacterial RNA in the sample. As expected, the number of human miRNA reads increased as the amount of input human RNA increased (Figure S3A). Similarly, the detected number of unique miRNAs increased with increased input human RNA, from 86 unique miRNAs in sample D2 to 535 in sample D14 (Figure S3B). A principal component analysis of the miRNA showed clear separation between the samples with high and low bacterial biomass (Figure S3C). Interestingly, small amounts of bacterial RNA (<10%, sample D8-D15) did not alter the miRNA distribution significantly, indicating the sRNA-seq protocol is robust to low levels of non-human contaminants. Indeed, the miRNA expression profiles of the samples with low bacterial biomass (D8–D15) were highly correlated ($r \geq 0.99$; Figure S3D). In contrast, the samples with high bacterial biomass showed more variation in their miRNA expression profiles and this variation was mainly related to the number of detected miRNAs.

Evaluation of Specificity, Sensitivity, and Quantitative Abilities of sMETASeq on a Mock Bacterial Community

We evaluated the ability of sMETASeq to correctly identify the expected species in a mock community of 20 bacteria. First, we investigated how many of the mock species sMETASeq and 16S rDNA-seq were able to identify. We found that sMETASeq identified 19 of the 20 species, whereas 16S rDNA-seq identified 18 of the 20 species (Figures 1A and 1B). Neither sMETASeq nor 16S rDNA-seq was able to identify *Actinomyces odontolyticus* and, additionally, 16S rDNA-seq failed to identify *Lactobacillus gasseri*. For both methods, the abundance of the bacteria decreased with decreasing input bacteria material (Figures 1A and 1B). For sMETASeq, the gram-negative *Rhodobacter sphaeroides* was the most abundant bacteria, and we also observed high expression of the gram-positive *Deinococcus radiodurans*.

Potential contaminant operational taxonomic units (OTUs) were identified using the *decontam* package in R (Davis et al., 2018). Consistent with the decreasing amounts of bacterial relative to human DNA/RNA, reads from the domain Eukaryota and the kingdom Metazoa were identified as the main contaminants in 16S rDNA-seq and sMETASeq, respectively, and the levels correlated well with the dilution series (Figures 1A and 1B). Contaminants from other OTUs were generally lowly expressed and likely represent cross-mapping of the sequencing reads or sequencing errors (Table S2). Importantly, none of the mock species were identified as a contaminant in either of the methods using the default *decontam* threshold parameter of 0.1.

To investigate the ability of sMETASeq to quantify the mock species we compared the normalized number of reads within and between the species. First, when grouping all the dilutions into one average estimate of species abundance we observed high correlation ($r = 0.7$) between sMETASeq and 16S rDNA-seq, indicating that sMETASeq is able to quantify the mock species with good accuracy (Figure 1C). Second, the within-species correlation across the dilutions was also high, indicating that sMETASeq can quantify bacteria at different biomass levels (Figure S4A). When the 16S rDNA data were analyzed using Qiime2 and the GTDB database (Parks et al., 2018), we also observed high correlation between the two methods, although four species were not detected by Qiime2 (Figure S5A). Next, we correlated the abundance of the 20 mock OTUs against the amount of input DNA/RNA to evaluate the ability of sMETASeq to directly quantify bacteria with respect to input bacterial biomass. When performing linear regression against all 15 dilutions, sMETASeq showed higher correlation than 16S rDNA-seq ($p = 0.0001$, Wilcoxon rank-sum test; Figure S4B). However, when limiting the correlation analysis to the most diluted samples (D8–D15), 16S rDNA-seq correlated better with input bacterial biomass than did sMETASeq ($p = 0.0003$, Wilcoxon rank-sum test) (Figure S4C). This indicated that sMETASeq has higher sensitivity at high bacterial biomass, whereas 16S rDNA-seq has higher sensitivity at low bacterial biomass.

The specificity of sMETASeq was investigated by measuring how many of the mock species were correctly assigned. A mock species was defined as correctly assigned if the species had the highest number of assigned reads within its corresponding genus. The correct predictions for sMETASeq ranged between 16 and 19 for the different dilutions, and for 16S rDNA-seq the correct predictions ranged between 13 and 15 (Figure 1D; $p = 6 \times 10^{-6}$, sign test). sMETASeq was particularly good at predicting the correct species for samples with high bacterial biomass. When performing the same analysis using Qiime2 and the GTDB database, we observed on average 13 correct predictions for the highest expressed species (Figure S5B). This slightly reduced performance by Qiime2 is partly due to the fact that four species were not detected by Qiime2. Together, these results indicate that sMETASeq is good at discriminating between species in diverse bacterial communities and that 16S rDNA-seq is more comparable with sMETASeq at low bacterial biomass and that sMETASeq is superior at high bacterial biomass.

Diversity Metrics of sMETASeq

Next, we measured the alpha and beta diversity of sMETASeq across the dilutions. We observed a distinct difference in diversity between the non-diluted and the diluted samples. For samples with high bacterial biomass (D1–D6), 16S rDNA-seq generally overestimate diversity (Shannon and Simpson index) and species richness ($p = 0.007$, $p = 0.01$, and $p = 0.03$, comparing "Simpson," "Shannon," and "Species richness," respectively for D1–D6; paired Wilcoxon rank-sum test) (Figure 1E). When bacterial biomass decreased, the estimated diversity tended to increase more rapidly for sMETASeq followed by and increase for 16S rDNA-seq, but the estimated diversity was comparable for the two most diluted samples with lowest bacterial biomass (D14–D15; "Simpson," "Shannon," and "Species richness"). When comparing the diversity for

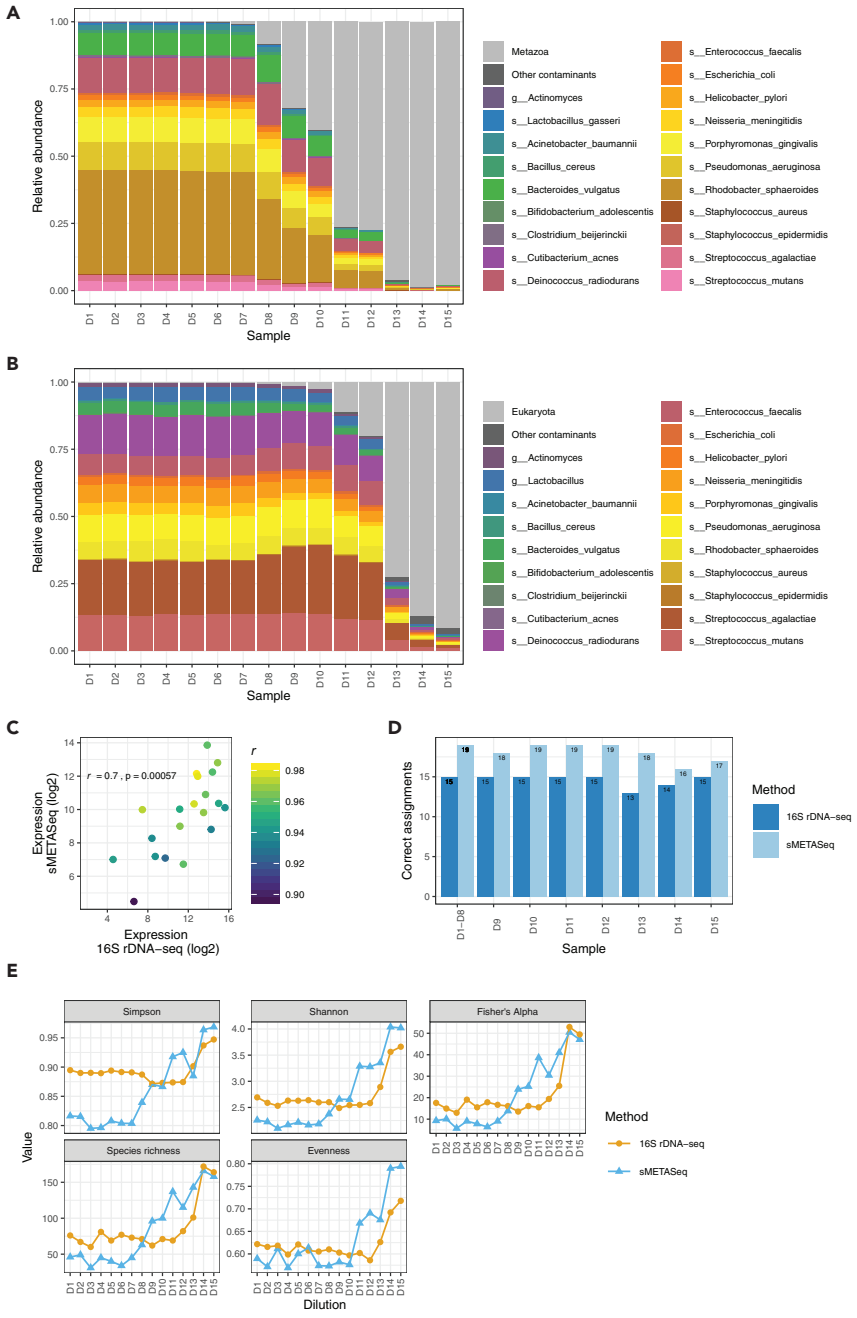


Figure 1. Bacterial Composition in Mock Community

(A) Relative abundance of the 20 mock bacteria across the dilutions as identified by sMETASeq. Contaminants are indicated in light and dark gray.

(B) Similar as in (A) for 16S rDNA-seq.

(C) Correlation of mock bacteria abundance between sMETASeq and 16S rDNA-seq for pooled samples. The Pearson's correlation values for each individual bacterium across dilutions are indicated with color. p = 0.00057. See also Table S2.

(D) Number of correctly assigned mock species for sMETASeq (light blue) and 16S rDNA-seq (dark blue) across dilutions. Dilutions D1–D8 have the same number of correct assignments and are therefore pooled.

(E) Comparison of diversity measurements between sMETASeq and 16S rDNA-seq across the dilutions.

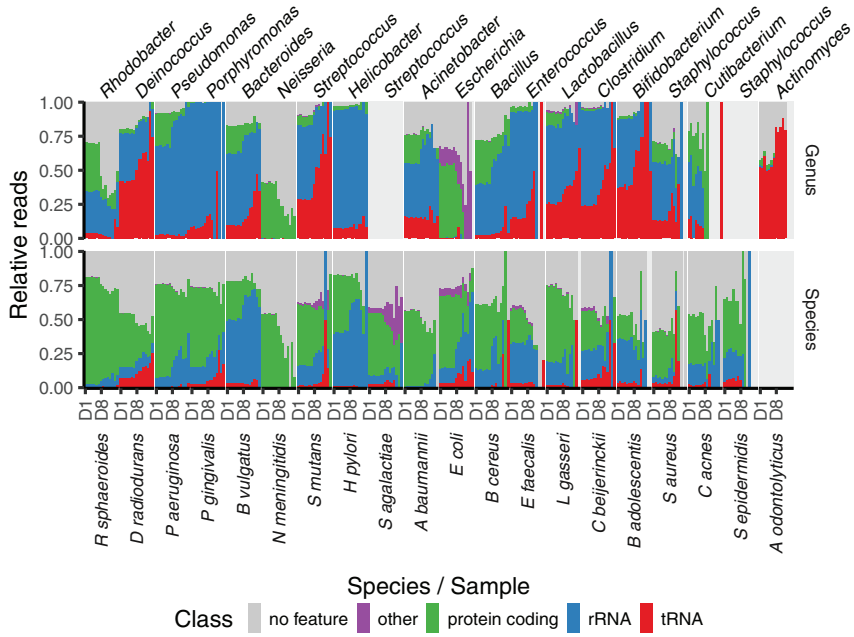


Figure 2. Relative Abundance of RNA Types in Mock Community

Shown are the relative abundances of the different gene classes for the reads assigned to the species (bottom) and genus (top) of the bacteria in the mock community. The dilutions are ordered along the x axis, and the annotation "D8" illustrates the shift from high bacterial biomass samples to low bacterial biomass samples.

all diluted samples where bacteria RNA/DNA was higher than human RNA/DNA (D8-D15), there was a significant difference in Shannon diversity and species richness between sMETASeq and 16S rDNA-seq but not in Simpson diversity ($p = 0.04$, $p = 0.55$, and $p = 0.04$ for "Shannon," "Simpson," and "Species richness," respectively, paired Wilcoxon rank-sum test).

Reads from Protein Coding Genes Are Enriched at Species Level

Having shown that sRNAs can be used to measure bacteria in a mock community we wanted to investigate the genomic origins of the bacterial sRNAs. All sequencing reads assigned to the bacterial strains in the mock community or to their genus were therefore further mapped to the corresponding strain-specific genomes of the 20 mock species. The bacterial sRNAs overlapped different classes of RNAs, the most common being protein-coding RNAs, followed by rRNAs and tRNAs (Figure 2). Reads assigned to the strains were enriched for protein-coding genes, whereas reads assigned at the genus level were enriched for rRNAs and tRNAs. This difference is consistent with rRNA and tRNAs being more evolutionarily conserved than protein-coding genes. Furthermore, when analyzing the relative abundance of RNA types across dilutions at species level, we observed a relative enrichment of rRNAs and tRNAs in samples with low bacterial biomass (D8–D15) compared with samples with high or equal bacterial biomass (D1–D7) and opposite for protein-coding RNAs (Figure S4D). The relative increase for tRNAs in samples with low bacterial biomass was also present for reads mapping at the genus level. Together, these findings illustrate that species-specific identification of bacteria is a result of reads overlapping the protein coding part of the genome of the bacteria and can explain some of the differences in specificity between 16S rDNA-seq and sMETASeq.

Bacterial Identification by sMETASeq in Colon Tissue Reflects the Gut Microbiota

Our group previously performed sRNA-seq on tumor and adjacent normal samples from 48 patients with colon cancer (96 samples) (Mjelle et al., 2019). Bacteria are known to play important roles in the carcinogenesis of colon cancer (Dahmus et al., 2018), and we therefore wanted to investigate if sMETASeq could be used to detect bacteria in these samples. We detected high amounts of bacteria from the genus *Bacteroides* in most samples, and the most abundant species was *Bacteroides fragilis* and *Bacteroides vulgatus*, both naturally occurring in the colon microbiota (Figure 3A). We observed consistent abundance of other gut-associated bacteria, including the orders *Clostridiales* and *Enterobacteriales*, the species

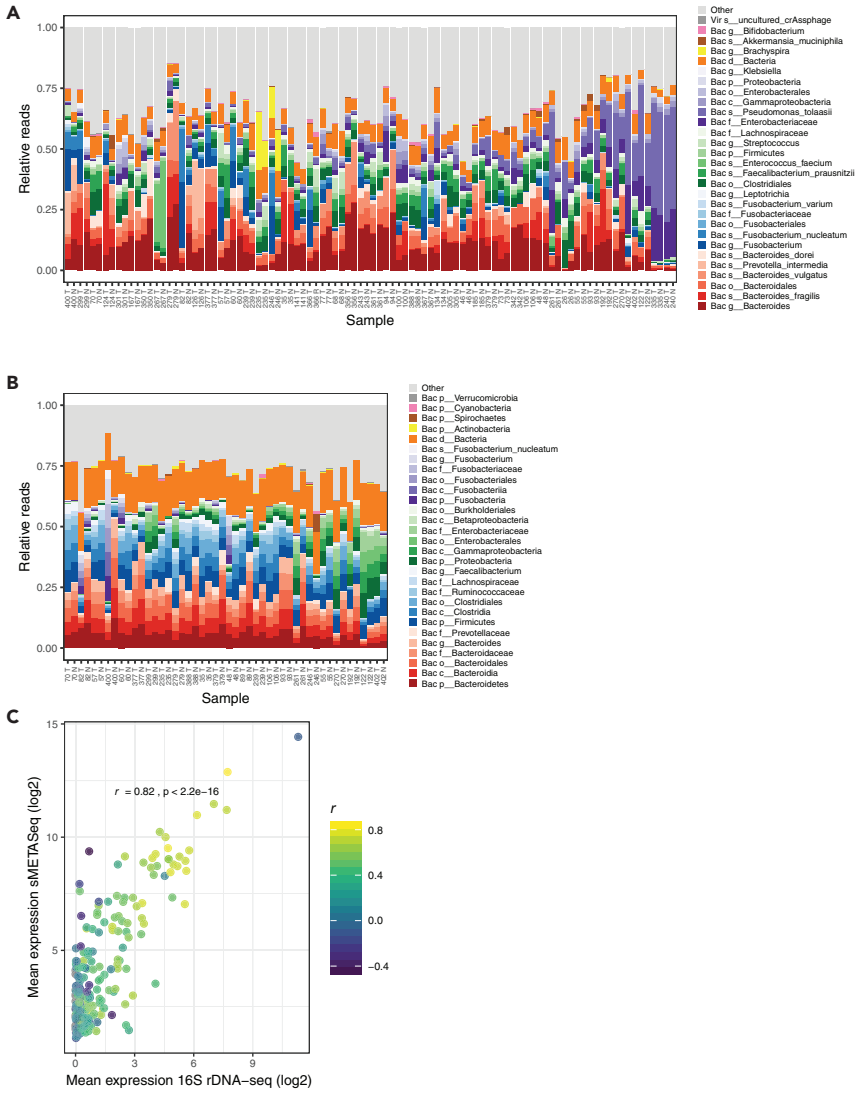


Figure 3. OTUs in Colon Tissue as Identified by sMETASeq and 16S rDNA-seq

(A) Detected OTUs by sMETASeq in 96 colon tissue samples. The samples are ordered from left to right with respect to the number of bacteria reads identified. OTUs within the same taxonomic lineage are indicated with different saturation of the same color. The taxonomic level for each OTU is indicated as "d," domain; "p," phylum; "c," class; "o," order; "g," genus; "s," species, and bacteria are indicated as "Bac" and virus as "Vir." Reads aligned to other OTUs than those listed are indicated as "other." Samples are indicated with patient number and "T" and "N" for tumor and normal, respectively.

(B) Detected OTUs by 16S rDNA-seq in 48 colon tissue samples. See (A) for a description of the plot.

(C) Correlation plot showing the mean expression of OTUs across samples as identified by sMETASeq (y axis) and 16S rDNA-seq (x axis). The color of the points indicates the correlation coefficient (Pearson's) and the axis shows the expression. $p < 2.2e\text{-}16$.

Enterococcus faecium and *Faecalibacterium prausnitzii*, as well as the colon-cancer-associated *Fusobacterium* (Figure 3A).

Bacterial Identification by sMETASeq Correlates with that of 16S rDNA-Seq

Having identified bacteria using sMETASeq, we performed 16S rDNA-seq on a subset of the same samples (48 samples, owing to lack of DNA from all 96) and compared the abundance of the OTUs detected by the two methods. The 16S data also showed high amounts of bacteria from *Bacteroides*, *Clostridiales*, and *Enterobacteriales*, similar to that identified by sMETASeq (Figure 3B). We compared the abundance of the bacterial species that were identified by both 16S rDNA-seq and sMETASeq and observed a high

correlation between the two methods ($r = 0.82$) (Figure 3C). When focusing on the individual species, most species were positively correlated at the sample level between the two methods and more than 60% of the species had a correlation value greater than 0.5. The species *B. ovatus* and *F. prausnitzii* showed the highest correlations between the two methods (Figure S6). When analyzing the 16S rDNA data using Qiime2 and the GTDB database, we observed good correlation between genera that were detected by both methods (Figure S5C).

Bacteria are shown to play important roles in colon cancer, and we therefore wanted to analyze differences in bacterial composition between tumor and normal samples and compare the differences across the two methods. We observed differential expression of several bacteria in tumor and normal samples, the clearest being *Fusobacteria*, *Bacteroidetes*, and *Proteobacteria* (Figure S7A). The differential expression observed by sMETASeq was highly reproducible in the 16S data, although to a lesser extent and partly at different taxonomic levels (Figure S7A, right panel). Furthermore, comparing the logFC values between tumor and normal samples for the two methods showed that most OTUs were changing in the same direction (Figure S7B and Table S3).

sMETASeq Detects Microorganisms in Human Biofluids, Cervicovaginal Self-Samples, and Oral Leukoplakia

Having established that sMETASeq is comparable with 16S rDNA-seq in identifying and quantifying bacteria, we went on applying sMETASeq on panels of different sample types. First, we analyzed a publicly available dataset comprising sRNA-seq data from nine different human body fluids (Seashols-Williams et al., 2016). The highest relative amounts of bacterial reads were found in vaginal secretion, menstrual secretion, feces, and saliva, whereas urine showed low amounts of bacterial reads (Figure S8A). The detected bacteria were representative of the respective biofluids, with *Lactobacillus* being enriched in vaginal and semen samples, *Bacteroides* in feces, *Cutibacterium acnes* in perspiration, and *Prevotella* in saliva (Figure 4A). Samples were generally consistent within the biological replicates but also indicated individual differences, such as *Gardnerella vaginalis* infections in two of the samples (Figure 4A). Next, we analyzed a second publicly available dataset comprising samples from human saliva, urine, serum, plasma, blood, and lymphocytes (El-Mogy et al., 2018). Similar as in the dataset of Seashols-Williams et al., saliva had the highest number of bacterial reads (Figure S8B) and were enriched with *Prevotella* and *Fusobacterium*, both common bacteria of the oral microbiota (Figure 4B).

To further investigate the bacterial composition in vaginal samples, and to show that sMETASeq also detects viruses, fungi, and other eukaryotes, we analyzed a dataset containing 56 HPV-positive cervicovaginal self-samples (Snoek et al., 2018). Twenty-four of the samples were histologically diagnosed with CIN3, characterized by dysplasia in the cervix, and 32 samples were CIN1 and HPV positive. Viral miRNAs can be detected by sMETASeq using miRBase; however, several viruses do not have miRNAs and PCR-based methods are used for detection. We analyzed the HPV-infected cervical samples and detected Alphapapillomavirus in 12 of the 24 CIN3 samples and one of the normal samples ($p = 2.87 \times 10^{-5}$, Chi-square test for the difference between CIN3 and normal), indicating that the level of Alphapapillomavirus increases with increasing dysplasia (Table S4). Next, we searched the literature for known fungi and parasites reported to be present in the female genital tract and focused on *Candida* and *Trichomonas vaginalis* (Bradford and Ravel, 2017). We detected two *Candida* species, *tropicalis* and *albicans*, of which *Candida albicans* was the highest expressed and found in many of the samples (Figure 4C). The protozoan parasite *Trichomonas vaginalis* was also detected in several samples.

Finally, we applied sMETASeq on 20 samples of oral leukoplakia, a potentially malignant disorder affecting the oral mucosa (Philipone et al., 2016). Leukoplakia is clinically important owing to its association with the development of oral squamous cell carcinoma (OSCC), a disease with high morbidity and mortality (Bewley and Farwell, 2017). The dataset consisted of 20 subjects divided into two groups; group 1 ($n = 10$) "progressive group" (patients with leukoplakia that progressed to OSCC within 5 years) and group 2 ($n = 10$) "non-progressive group" (patients with leukoplakia that did not progress to OSCC within 5 years). We compared bacterial expression between the two groups and detected one differentially expressed genus, *Neisseria*, which showed significantly higher expression in the progressive group compared with the non-progressive group (Figure 4D) ($p = 0.005$, Benjamini Hochberg-corrected). When investigating the most abundant microbes across all 20 samples, we detected high abundance of several microbes associated with the oral cavity (Figure S8C). Interestingly, one of the patients in the progressive group showed very high levels

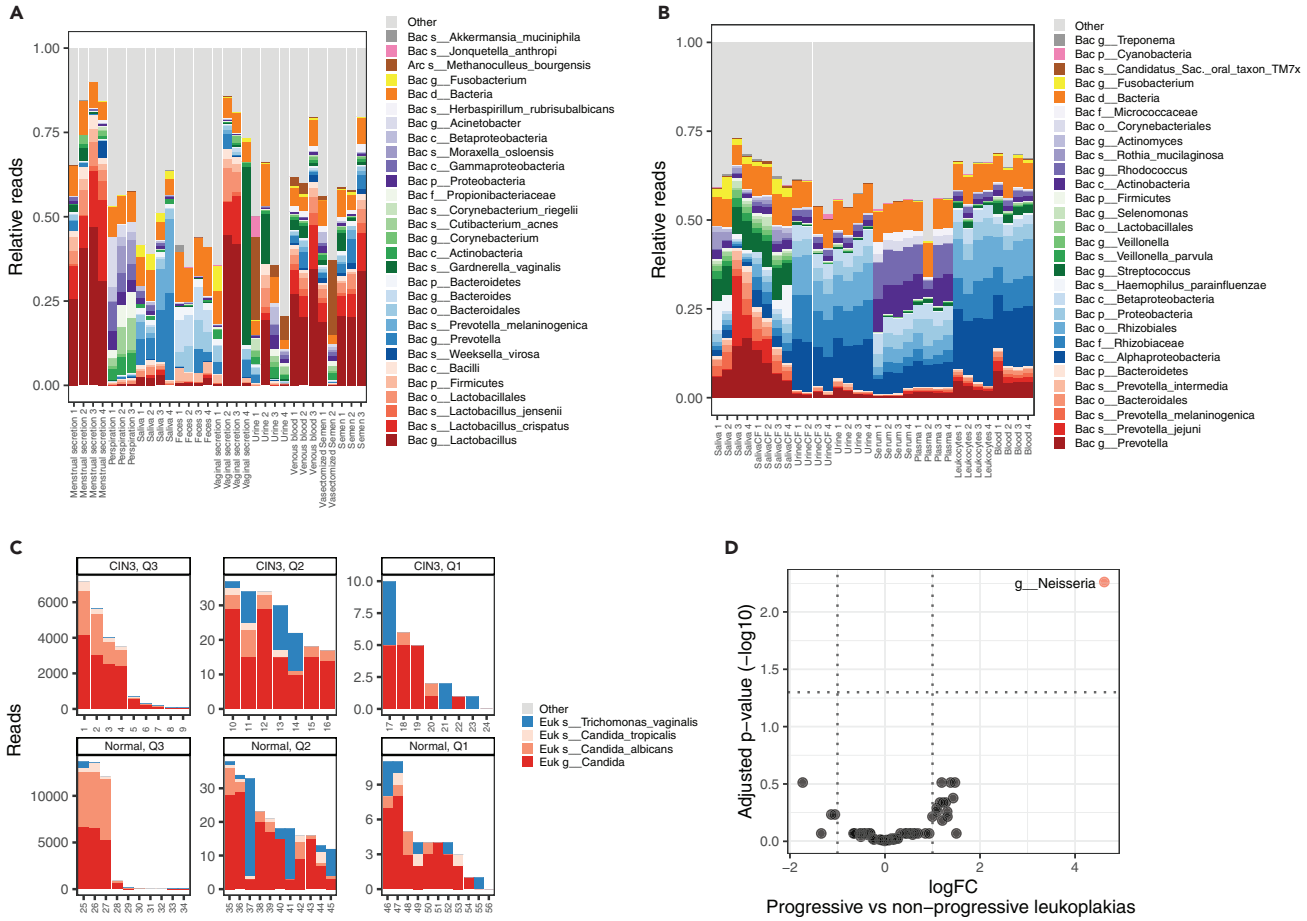


Figure 4. Microbes in Human Biofluids, Cervix and Oral Leukoplakia Samples

(A) Distribution of bacteria in nine different human biofluids (Seashols-Williams et al., 2016) as identified by sMETASeq. See Figure 3A for a description of the plot.

(B) Distribution of bacteria from eight different human biofluids (Seashols-Williams et al., 2016) as identified by sMETASeq. See Figure 3A for a description of the plot.

(C) Expression of fungi and parasites in cervix samples from persons with CIN1 (normal) cervix samples and persons with CIN3 (severe dysplasia) (Snoek et al., 2018). The samples are grouped into three quantiles (Q1–3) based on the number of reads aligning to the specific OTUs. The plots marked with Q3 show the samples with the highest number of reads for the specific OTU, and the plots marked with Q1 show the samples with the lowest number of reads for the specific OTU.

(D) Differentially expressed bacteria in progressive oral leukoplakia versus non-progressive oral leukoplakia. The x axis shows the logFC between progressive and non-progressive oral leukoplakia. The y axis shows the inverse (-log10) Benjamini-Hochberg-adjusted p values. Significant bacteria are indicated in red. The dotted gray lines indicated absolute logFC of 1 and the gray horizontal line indicates the significant threshold of 0.05.

of Epstein-Barr virus (EBV, human gammaherpesvirus 4) (Figure S8C). EBV has been linked to both oral carcinomas and oral leukoplakia, although the sample size is too small to draw conclusions from this dataset (Guidry et al., 2018). Together, these results indicate that bacterial small RNAs can be used to differentiate progressive and non-progressive oral leukoplakia samples and potentially serve as biomarkers for OSCC development.

Bacterial Detection in Infected Cell Lines

To further validate that sMETASeq can indeed detect bacteria known to be present in a sample, we performed sRNA-seq of a mycoplasma-infected JJN-3 myeloma cell line and a matched non-infected cell line and compared the level of mycoplasma with a luciferase-based mycoplasma test assay. In the sRNA-seq data, the infected cell line showed between 100 and 200 times more mycoplasma than the non-infected cell line across replicates (Figure S8D). The same cell lines were tested using the MycoAlert Mycoplasma Detection Kit (Lonza), which showed no detectable levels of mycoplasma in the non-infected

cell line (readout of 0.6 and 0.4) and high levels in the infected cell line (readout of 70 and 24) (Figure S8D). A readout above 1.2 indicates a positive test.

DISCUSSION

In this article, we present sMETASeq for combined metagenomics and host small RNA profiling. The method provides high-quality metagenomic profiling, is an alternative to current DNA-based methods, and can be applied to various sample material from tissue biopsies to biofluids. The method is particularly suited for research questions where both host small RNAs, for instance, human miRNAs, and the host microbiome are of interest. The method can, for instance, be used to study associations between human miRNAs and microbial composition in one single experiment. sMETASeq displays the versatility of small RNA-based sequencing and shows that bacterial sRNAs are more widespread and consistently expressed than previously anticipated. This could indicate that bacterial sRNAs are protected from degradation in many sample types, either by binding to proteins or being contained in vesicles. Indeed, studies have shown that prokaryotic vesicles contain different RNA types that could be delivered and interact with eukaryotic cells Dauros-Singorenko et al. (2018). Given the stability of bacterial sRNA, they could also function as biomarker for disease.

In addition to describing sMETASeq, this is the first study to perform a comprehensive analysis of small RNA metagenomics and to compare it with the widely used 16S rDNA method. It was recently shown that a modified sRNA-seq protocol focusing on bacterial tRNAs (tRNA-seq) can be used to characterize the microbiome (Schwartz et al., 2018). In contrast to sMETASeq, tRNA-seq will not provide information on other RNA types, for instance, miRNAs, and is a more specific protocol to study tRNA modifications. The strengths of sMETASeq lie in its ability to investigate multiple types of sRNAs and that the data can be used to study both metagenomics and other biological questions. As shown in this paper, sMETASeq tends to perform better at higher bacterial concentration and to some degree lack the sensitivity of 16S rDNA-seq at very low bacterial concentrations. For instance, sMETASeq tended to slightly overestimate sample diversity at low concentrations, and fewer reads mapped to microbes when bacteria concentration decreased. Our results show that, for sample types with high microbial biomass such as fecal, colon, oral, and vaginal samples, sMETASeq would be a good method for identification and quantification of the microbiome. However, in samples with low microbial biomass, such as blood or different human tissues, sMETASeq would likely lack the sensitivity to detect lowly expressed microbes. However, future developments of sMETASeq, with, for instance, a microbial enrichment step, would make the method more sensitive toward microbes; however, this has to be optimized not to compromise the host sRNA profiles.

RNA is generally more prone to degradation compared with DNA, which makes sMETASeq more sensitive to samples with degraded RNA. However, since sMETASeq aims at small RNAs, even degraded RNA molecules contain valuable information that can be used to identify microbes in the sample. Using a *k*-mer-based analysis method as applied by sMETASeq in the kraken pipeline, short fragments will be suited for taxonomic classification even if they are degradation products.

To compare the sensitivity and the specificity of sMETASeq and 16S rDNA-seq we chose to perform an RNA- and DNA-based sequencing experiment of a mock bacterial community comprising 20 known bacterial species. In terms of specificity, sMETASeq showed good performance, both in samples with high and low bacterial biomass, although the specificity was slightly compromised in samples with low bacterial biomass. These differences in specificity between sMETASeq and 16S rDNA-seq could be attributed to the fact that sMETASeq is not limited to the 16S region for bacteria identification but utilizes reads mapping to the whole bacterial genome. For species with highly similar genomes, sMETASeq is likely to improve the discrimination. Indeed, when investigating the abundance of reads mapping to different RNA types for sMETASeq, we observed that species-specific identification is a result of reads mapping to protein-coding RNAs. The enrichment of protein-coding RNAs for reads that map to the specific bacterial strains shows that the protein-coding region contains valuable strain-specific information that can be utilized when discriminating between closely related species within the same genus. The reduced specificity and increased diversity measures in samples with low bacterial biomass could be explained by the reduction in detected protein-coding RNAs for these samples. Regarding sensitivity and quantitative abilities, the two methods correlated well with each other in measuring the abundance of the mock species across the dilution. All correlation values were above 0.9, and several correlation values were close to 1. When focusing on the samples with low bacterial biomass, 16S rDNA-seq correlated better with input

bacterial biomass than did sMETASeq. In contrast, when including the high bacterial biomass samples, sMETASeq correlated better with biomass than did 16S rDNA-seq.

We applied sMETASeq on a previously published in-house sRNA-seq dataset from colon tissue and publicly available data from human biofluids and showed that the bacteria identified largely reflected the sample of origin, supporting that the findings are biologically relevant and not a result of sample contamination. Analyzing the colon dataset, we observed high levels of bacteria commonly found as part of the gut microbiota, including *Faecalibacterium, Enterobacteriaceae, Bacteroides*, and *Fusobacterium* (Garrett et al., 2010; Miquel et al., 2013). The latter has been identified as a potential player in colon cancer (Shang and Liu, 2018). Two patients showed high levels of the genus *Brachyspira*, both in the normal and tumor samples. *Brachyspira* has been associated with diarrhea and colitis in several animals and is the cause of spirochetosis in human, an infection of the colonic mucosa (Amat Villegas et al., 2004).

Human biofluids were analyzed using two publicly available datasets and were shown to contain a wide range of bacteria down to genus and species level. The bacteria identified largely reflected the known microbiota of the different biofluids. The saliva samples showed high levels of the oral bacterium *Fusobacterium*. *Fusobacterium* can also be isolated from the vaginal microbiome (Hillier et al., 1993) and one of the vaginal secretion sample showed very high levels of *Fusobacterium,* indicating potential vaginosis. The saliva and vaginal secretion samples also showed high levels of the bacteria *Prevotella*, which have previously been associated to the oral, vaginal, and gut microbiota (Gholizadeh et al., 2016; Ley, 2016; Si et al., 2017). *Prevotella* was highly expressed in the saliva samples in both datasets. The oral bacterium *Veillonella parvula* was detected in the saliva samples in the dataset of El-Mogy et al., and in the dataset of Seashols-Williams et al. we detected the phylum Firmicutes to which *Veillonella parvula* belongs. Stool samples are known to express high levels of bacteria, and sMETASeq showed that the feces samples had the highest proportion of bacterial reads. We found that these bacteria are mainly within the order *Bacteroides*, as expected from previous 16S studies (Eggerth and Gagnon, 1933). The perspiration samples showed high amounts of the skin-specific bacterium *C. acnes* and *Moraxella osloensis,* which is also frequently found in skin (Alkhatib et al., 2017; Dreno et al., 2018). In one of the vaginal secretion samples high amounts of the bacterium *Gardernella vaginalis* was detected. This bacterium is involved in bacterial vaginosis, a vaginal condition caused by abnormal bacterial composition in the vagina (Schwebke et al., 2014). In the cervix samples we detected many highly expressed bacteria at the species levels, all previously associated with the female genital tract. Interestingly, these samples also contained high levels of other eukaryotes, in particular two fungi species within the Candida genus, *Candida albicans* and *tropicalis*, and overgrowth of these fungi is shown to cause vaginal candidiasis, an infection in the vagina. Another interesting finding is the parasite *Trichomonas vaginalis*, which was detected in a subset of the samples. *T. vaginalis* has been isolated from samples of the vagina and has been associated with bacterial vaginosis (Franklin and Monif, 2000; Moodley et al., 2002). In the same dataset, we detected viral RNA fragments from the HPV virus Alphapapillomavirus 9. This HPV species has several subtypes, including HPV-16, which is one of the subtypes that can lead to cervix cancer (Holl et al., 2015).

Oral leukoplakia are white patches on the oral mucosa and has malignant potential as these patches can be precursor lesions of OSCC. The likelihood of malignant transformation of oral leukoplakia is observed to lie between 0.2% and 3% and varies between studies and population groups (Bewley and Farwell, 2017). Early detection of OSCC increases the survival rate from 20% to 30% to approximately 80%, which highlights the importance of developing new good diagnostic biomarkers (Noone et al., 2018). Different microorganisms have been linked to oral cancer; however, there are still discussions regarding the role and importance of these microorganisms (Gholizadeh et al., 2016; Healy and Moran, 2019; Pushalkar et al., 2011). Using sMETASeq we were able to find high abundance of *Neisseria*, a genus of gram-negative species belonging to the phylum Proteobacteria. Most *Neisseria* species are non-pathogenic; however, at least two species are regarded as pathogenic, *Neisseria meningitidis* and *Neisseria gonorrhoeae*. Neisseria species are highly abundant in the human oral cavity and have been detected in, for instance, saliva, plaque, mucosal surfaces in the mouth, and teeth, and Neisseria has been regarded as part of the "core microbiome" of the healthy human oral cavity (Keijser et al., 2008; Zaura et al., 2009). It has previously been shown that *Neisseria* is able to produce the carcinogenic organic compound acetaldehyde; however, it is not clear if *in vivo* production of acetaldehyde by *Neisseria* is related to carcinogenesis (Muto et al., 2000; Yokoyama et al., 2018). The other microorganisms detected by sMETASeq in the oral leukoplakia samples generally reflected the oral sample type. For instance, *Capnocytophaga*, *Fusobacterium*, *Pasteurellaceae*, *Rothia*, *Gemella*, and

gammaherpesvirus are all related to the oral cavity, and some are implicated in oral cancer or oral leukoplakia.

In summary, applying sMETASeq to different human biofluids, cervical self-samples, and oral leukoplakia, we showed that bacteria, fungi, parasites, and viruses can be identified and quantified between groups. We show that the identification is comparable for similar sample types across datasets and that the microorganisms reflect the biology of the sample in which they are detected.

The establishment of sMETASeq for bacterial identification enables researchers to analyze publicly available datasets and to plan new experiments in which both human small RNAs and other organisms can be identified. sMETASeq also identifies viruses, independent of whether they encode viral miRNAs or not, as well as fungi and other eukaryotes and parasites, and is therefore one of the most versatile protocols for metagenomics. Moreover, by adjusting the gel-purification step during sRNA library preparation the ratio between long or short fragments can be changed to favor different RNA species. Another advantage of using sRNAs in metagenomics is that sRNAs provide information about transcription. If bacterial sRNAs are detected in the samples, it is possible that the bacterium has active transcription since sRNAs from latent or dead bacteria would be rapidly degraded; however, RNA from latent or dead bacteria could also be present in samples if the RNA is protected by proteins, such as for Hfq-associated sRNAs (De Lay et al., 2013). 16S DNA, on the other hand, would be stable for a longer period of time and using 16S metagenomics would be better in case of dormant cells and non-dividing bacteria.

Thousands of sRNA-seq datasets have been submitted to the NCBI sequence read archive. These datasets might contain valuable information on sample microbiomes that researchers could access and analyze through sMETASeq. Several large consortium projects include sRNA as part of the pipeline. For instance, the FANTOM consortium (de Rie et al., 2017) contains sRNA-seq from every major human organ as well as primary cell lines, and The Cancer Genome Atlas project (TCGA) (Chu et al., 2016) contains sRNA-seq from the most common human cancers and includes both cancer and normal samples. We expect that researchers with an interest in metagenomics and microbes will apply sMETASeq to gain new insights into the role of microorganisms in human health.

Limitations of the Study

We here present a method for metagenomics profiling using small RNAs. Although the method performs well at characterizing and quantifying microbes in samples with high bacterial biomass, the sensitivity might be a limiting factor in samples with low bacterial microbes. Furthermore, in samples with low sequencing depth, the number of bacterial reads will often be low, compromising both the sensitivity and the specificity of the method. Further developments of sMETASeq addressing both the library preparation and the data analysis could potentially improve the method.

Resource Availability

Lead Contact

Further information and requests for resources, code, and scripts should be directed to and will be fulfilled by the Lead Contact, Robin Mjelle (robin.mjelle@ntnu.no).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

sMETASeq is available through github (https://github.com/MjelleLab/sMETASeq).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101131.

ACKNOWLEDGMENTS

We thank St. Olavs Hospital for providing colon tissue material for sRNA-seq and 16 rDNA-seq and for funding the sequencing. We thank the Genomics Core Facility at The Norwegian University of Science and Technology for performing sequencing.

AUTHORS CONTRIBUTIONS

R.M. performed 16S rDNA-seq, sRNA-seq, and data analysis and prepared the manuscript; K.R.A. provided cell lines and performed mycoplasma tests; E.H. and W.S. provided biobank material for colon samples and E.H. was responsible for acquiring funding; P.S. wrote the bioinformatics pipeline, generated figures, and helped preparing the manuscript. All authors commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 13, 2020
Revised: April 20, 2020
Accepted: April 28, 2020
Published: May 22, 2020

REFERENCES

Alkhatib, N.J., Younis, M.H., Alobaidi, A.S., and Shaath, N.M. (2017). An unusual osteomyelitis caused by *Moraxella osloensis*: a case report. Int. J. Surg. Case Rep. *41*, 146–149.

Amat Villegas, I., Borobio Aguilar, E., Beloqui Perez, R., de Llano Varela, P., Oquinena Legaz, S., and Martinez-Penuela Virseda, J.M. (2004). [Colonic spirochetes: an infrequent cause of adult diarrhea]. Gastroenterol. Hepatol. *27*, 21–23.

Bewley, A.F., and Farwell, D.G. (2017). Oral leukoplakia and oral cavity squamous cell carcinoma. Clin. Dermatol. *35*, 461–467.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857.

Bradford, L.L., and Ravel, J. (2017). The vaginal mycobiome: a contemporary perspective on fungi in women's health and diseases. Virulence *8*, 342–351.

Chen, X., Sim, S., Wurtmann, E.J., Feke, A., and Wolin, S.L. (2014). Bacterial noncoding Y RNAs are widespread and mimic tRNAs. RNA *20*, 1715–1724.

Chu, A., Robertson, G., Brooks, D., Mungall, A.J., Birol, I., Coope, R., Ma, Y., Jones, S., and Marra, M.A. (2016). Large-scale profiling of microRNAs for the cancer genome atlas. Nucleic Acids Res. *44*, e3.

Dahmus, J.D., Kotler, D.L., Kastenberg, D.M., and Kistler, C.A. (2018). The gut microbiome and colorectal cancer: a review of bacterial pathogenesis. J. Gastrointest. Oncol. *9*, 769–777.

Dauros-Singorenko, P., Blenkiron, C., Phillips, A., and Swift, S. (2018). The functional RNA cargo of bacterial membrane vesicles. FEMS Microbiol. Lett. *365*.

Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A., and Callahan, B.J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome *6*, 226.

De Lay, N., Schu, D.J., and Gottesman, S. (2013). Bacterial small RNA-based negative regulation: Hfq and its accomplices. J. Biol. Chem. *288*, 7996–8003.

de Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Astrom, G., Babina, M., Bertin, N., Burroughs, A.M., et al. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat. Biotechnol. *35*, 872–878.

Dreno, B., Pecastaings, S., Corvec, S., Veraldi, S., Khammari, A., and Roques, C. (2018). *Cutibacterium acnes* (*Propionibacterium acnes*) and acne vulgaris: a brief look at the latest updates. J. Eur. Acad. Dermatol. Venereol. *32* (*Suppl 2*), 5–14.

Eggerth, A.H., and Gagnon, B.H. (1933). The Bacteroides of human feces. J. Bacteriol. *25*, 389–413.

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.

Franklin, T.L., and Monif, G.R. (2000). *Trichomonas vaginalis* and bacterial vaginosis. Coexistence in vaginal wet mount preparations from pregnant women. J. Reprod. Med. *45*, 131–134.

Garrett, W.S., Gallini, C.A., Yatsunenko, T., Michaud, M., DuBois, A., Delaney, M.L., Punit, S., Karlsson, M., Bry, L., Glickman, J.N., et al. (2010). Enterobacteriaceae act in concert with the gut microbiota to induce spontaneous and maternally transmitted colitis. Cell Host Microbe *8*, 292–300.

Gholizadeh, P., Eslami, H., Yousefi, M., Asgharzadeh, M., Aghazadeh, M., and Kafil, H.S. (2016). Role of oral microbiome on oral cancers, a review. Biomed. Pharmacother. *84*, 552–558.

Gottesman, S., and Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb. Perspect. Biol. *3*.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. *34*, D140–D144.

Guidry, J.T., Birdwell, C.E., and Scott, R.S. (2018). Epstein-Barr virus in the pathogenesis of oral cancers. Oral Dis. *24*, 497–508.

Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res. *19*, 1141–1152.

Healy, C.M., and Moran, G.P. (2019). The microbiome and oral cancer: more questions than answers. Oral Oncol. *89*, 30–33.

Hillier, S.L., Krohn, M.A., Rabe, L.K., Klebanoff, S.J., and Eschenbach, D.A. (1993). The normal vaginal flora, H2O2-producing lactobacilli, and bacterial vaginosis in pregnant women. Clin. Infect. Dis. *16* (*Suppl 4*), S273–S281.

Holl, K., Nowakowski, A.M., Powell, N., McCluggage, W.G., Pirog, E.C., Collas De Souza, S., Tjalma, W.A., Rosenlund, M., Fiander, A., Castro Sanchez, M., et al. (2015). Human papillomavirus prevalence and type-distribution in cervical glandular neoplasias: results from a European multinational epidemiological study. Int. J. Cancer *137*, 2858–2868.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J. Clin. Microbiol. *45*, 2761–2764.

Keijser, B.J., Zaura, E., Huse, S.M., van der Vossen, J.M., Schuren, F.H., Montijn, R.C., ten Cate, J.M., and Crielaard, W. (2008). Pyrosequencing analysis of the oral microflora of healthy adults. J. Dent. Res. *87*, 1016–1020.

Kumar, P., Anaya, J., Mudunuri, S.B., and Dutta, A. (2014). Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. *12*, 78.

Ley, R.E. (2016). Gut microbiota in 2015: Prevotella in the gut: choose carefully. Nat. Rev. Gastroenterol. Hepatol. *13*, 69–70.

Majdalani, N., Vanderpool, C.K., and Gottesman, S. (2005). Bacterial small RNA regulators. Crit. Rev. Biochem. Mol. Biol. *40*, 93–113.

Massart, S., Chiumenti, M., De Jonghe, K., Glover, R., Haegeman, A., Koloniuk, I., Kominek, P., Kreuze, J., Kutnjak, D., Lotos, L., et al. (2019). Virus detection by high-throughput sequencing of small RNAs: large-scale performance testing of sequence analysis strategies. Phytopathology *109*, 488–497.

Miquel, S., Martin, R., Rossi, O., Bermudez-Humaran, L.G., Chatel, J.M., Sokol, H., Thomas, M., Wells, J.M., and Langella, P. (2013). *Faecalibacterium prausnitzii* and human intestinal health. Curr. Opin. Microbiol. *16*, 255–261.

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.

Moodley, P., Wilkinson, D., Connolly, C., Moodley, J., and Sturm, A.W. (2002). *Trichomonas vaginalis* is associated with pelvic inflammatory disease in women infected with human immunodeficiency virus. Clin. Infect. Dis. *34*, 519–522.

Muto, M., Hitomi, Y., Ohtsu, A., Shimada, H., Kashiwase, Y., Sasaki, H., Yoshida, S., and Esumi, H. (2000). Acetaldehyde production by non-pathogenic Neisseria in human oral microflora: implications for carcinogenesis in upper aerodigestive tract. Int. J. Cancer *88*, 342–350.

A.M. Noone, N. Howlader, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, and D.R. Lewis, et al., eds. (2018). SEER Cancer Statistics Review, 1975-2015 (National Cancer Institute). https://seer.cancer.gov/csr/1975_2015/.

O'Flaherty, B.M., Li, Y., Tao, Y., Paden, C.R., Queen, K., Zhang, J., Dinwiddie, D.L., Gross, S.M., Schroth, G.P., and Tong, S. (2018). Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. Genome Res. *28*, 869–877.

Pankaj, K. (2013). Methods for rapid virus identification and quantification. Mater. Methods *3*.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004.

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.

Pritchard, C.C., Cheng, H.H., and Tewari, M. (2012). MicroRNA profiling: approaches and considerations. Nat. Rev. Genet. *13*, 358–369.

Pushalkar, S., Mane, S.P., Ji, X., Li, Y., Evans, C., Crasta, O.R., Morse, D., Meagher, R., Singh, A., and Saxena, D. (2011). Microbial diversity in saliva of oral squamous cell carcinoma. FEMS Immunol. Med. Microbiol. *61*, 269–277.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Corrigendum: shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 1211.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., and Chen, W.; Fungal Barcoding Consortium; Fungal Barcoding Consortium Author List (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. U S A *109*, 6241–6246.

Schulz, F., Eloe-Fadrosh, E.A., Bowers, R.M., Jarett, J., Nielsen, T., Ivanova, N.N., Kyrpides, N.C., and Woyke, T. (2017). Towards a balanced view of the bacterial tree of life. Microbiome *5*, 140.

Schwartz, M.H., Wang, H., Pan, J.N., Clark, W.C., Cui, S., Eckwahl, M.J., Pan, D.W., Parisien, M., Owens, S.M., Cheng, B.L., et al. (2018). Microbiome characterization by high-throughput transfer RNA sequencing and modification analysis. Nat. Commun. *9*, 5353.

Schwebke, J.R., Muzny, C.A., and Josey, W.E. (2014). Role of *Gardnerella vaginalis* in the pathogenesis of bacterial vaginosis: a conceptual model. J. Infect. Dis. *210*, 338–343.

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.

Shang, F.M., and Liu, H.L. (2018). *Fusobacterium nucleatum* and colorectal cancer: a review. World J. Gastrointest. Oncol. *10*, 71–81.

Si, J., You, H.J., Yu, J., Sung, J., and Ko, G. (2017). Prevotella as a hub for vaginal microbiota under the influence of host genetics and their association with obesity. Cell Host Microbe *21*, 97–105.

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 *144*, 372–379.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66–74.

Visser, M., Bester, R., Burger, J.T., and Maree, H.J. (2016). Next-generation sequencing for virus detection: covering all the bases. Virol. J. *13*, 85.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Yokoyama, S., Takeuchi, K., Shibata, Y., Kageyama, S., Matsumi, R., Takeshita, T., and Yamashita, Y. (2018). Characterization of oral microbiota and acetaldehyde production. J. Oral Microbiol. *10*, 1492316.

Zaura, E., Keijser, B.J., Huse, S.M., and Crielaard, W. (2009). Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol. *9*, 259.

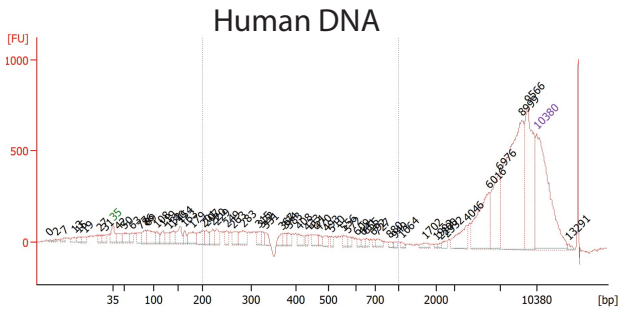
iScience, Volume *23*

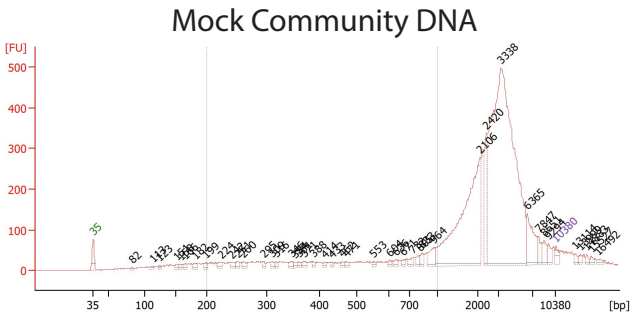
Supplemental Information

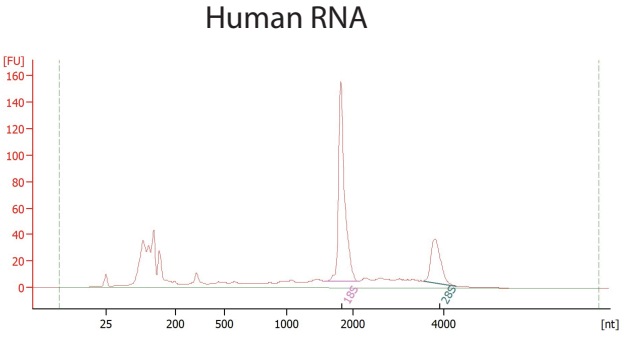
sMETASeq: Combined Profiling of Microbiota and Host Small RNAs

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

Figure S1


Human DNA
[FU]
1000
500
0
35
100
200
300
400
500
700
2000
10380
[bp]


Mock Community DNA
[FU]
500
400
300
200
100
0
35
100
200
300
400
500
700
2000
10380
[bp]


Human RNA
[FU]
160
140
120
100
80
60
40
20
0
25
200
500
1000
2000
4000
[nt]

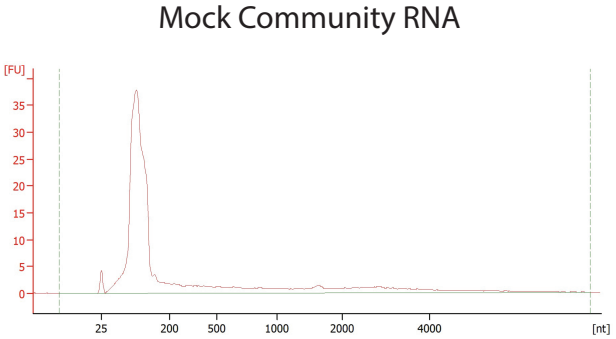

Mock Community RNA
[FU]
35
30
25
20
15
10
5
0
25
200
500
1000
2000
4000
[nt]

Figure S2

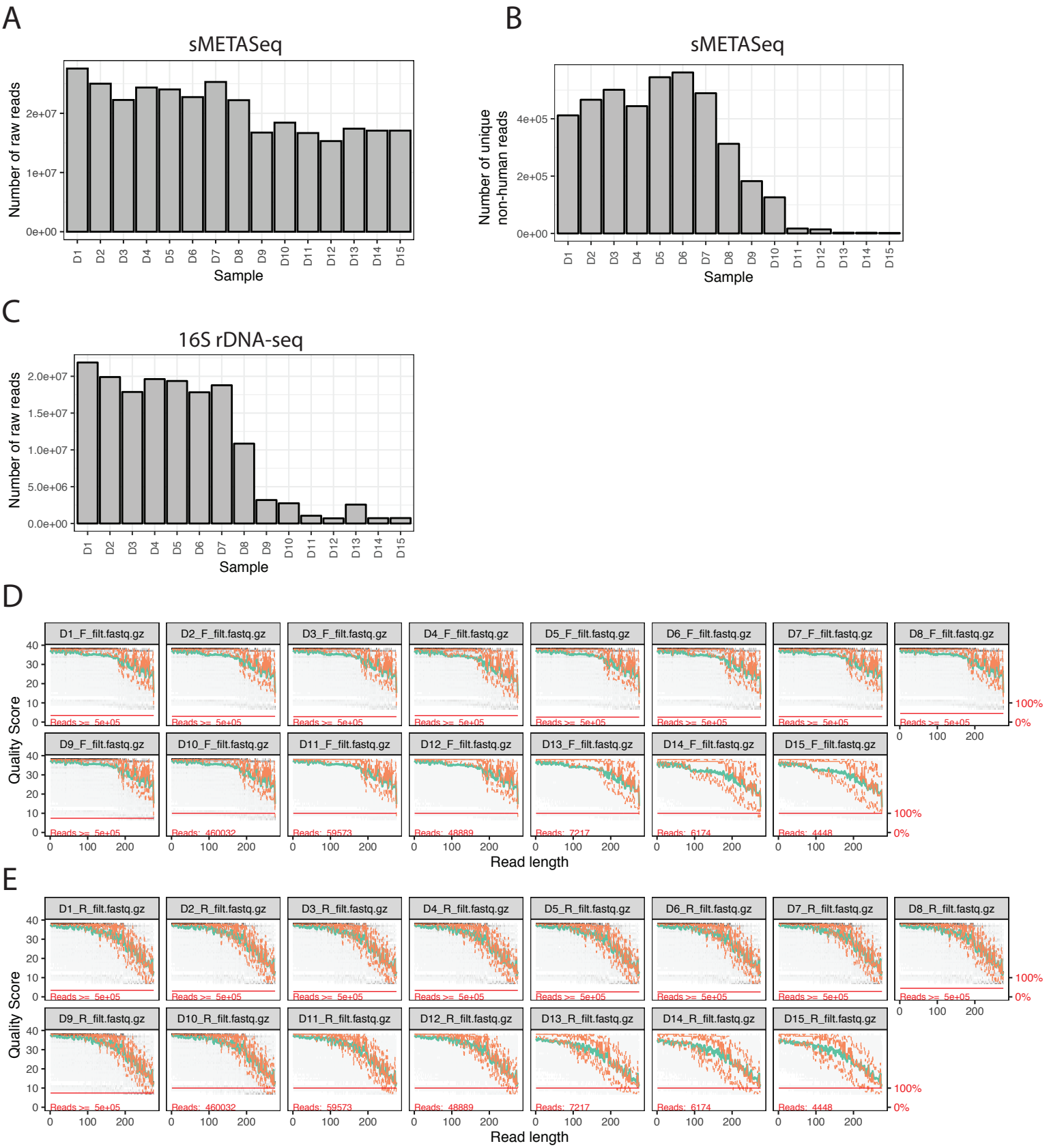

A
sMETASeq
Number of raw reads
0e+00
1e+07
2e+07
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
B
sMETASeq
Number of unique non-human reads
0e+00
2e+05
4e+05
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
C
16S rDNA-seq
Number of raw reads
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
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
Reads >= 5e+05
Reads: 460032
Reads: 59573
Reads: 48889
Reads: 7217
Reads: 6174
Reads: 4448
100%
0%
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
Reads >= 5e+05
Reads: 460032
Reads: 59573
Reads: 48889
Reads: 7217
Reads: 6174
Reads: 4448
100%
0%
Read length

Figure S3

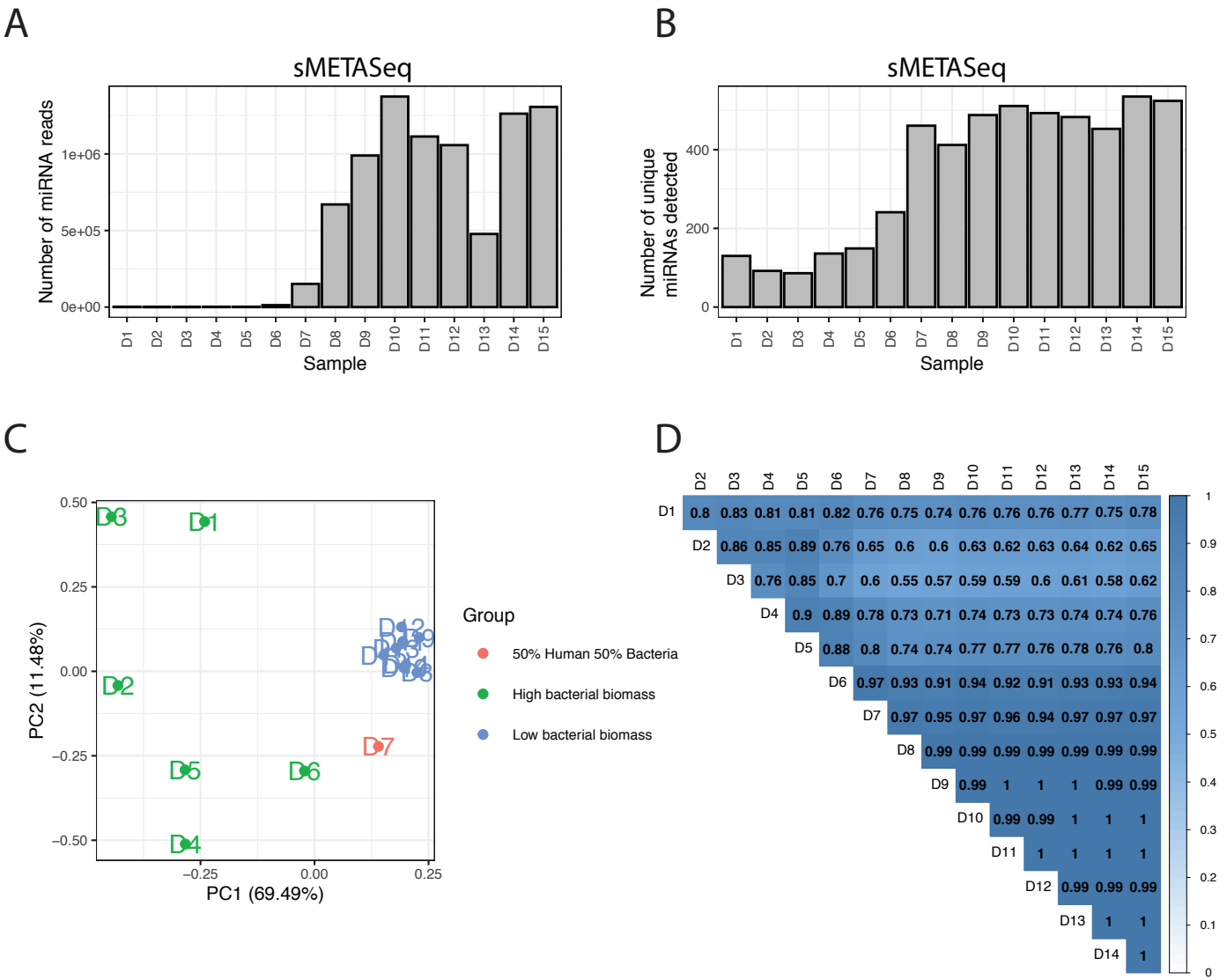

A
sMETASeq
Number of miRNA reads
1e+06
5e+05
0e+00
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
B
sMETASeq
Number of unique miRNAs detected
400
200
0
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10 D11 D12 D13 D14 D15
Sample
C
PC2 (11.48%)
0.50
0.25
0.00
−0.25
−0.50
−0.25
0.00
0.25
PC1 (69.49%)
Group
50% 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

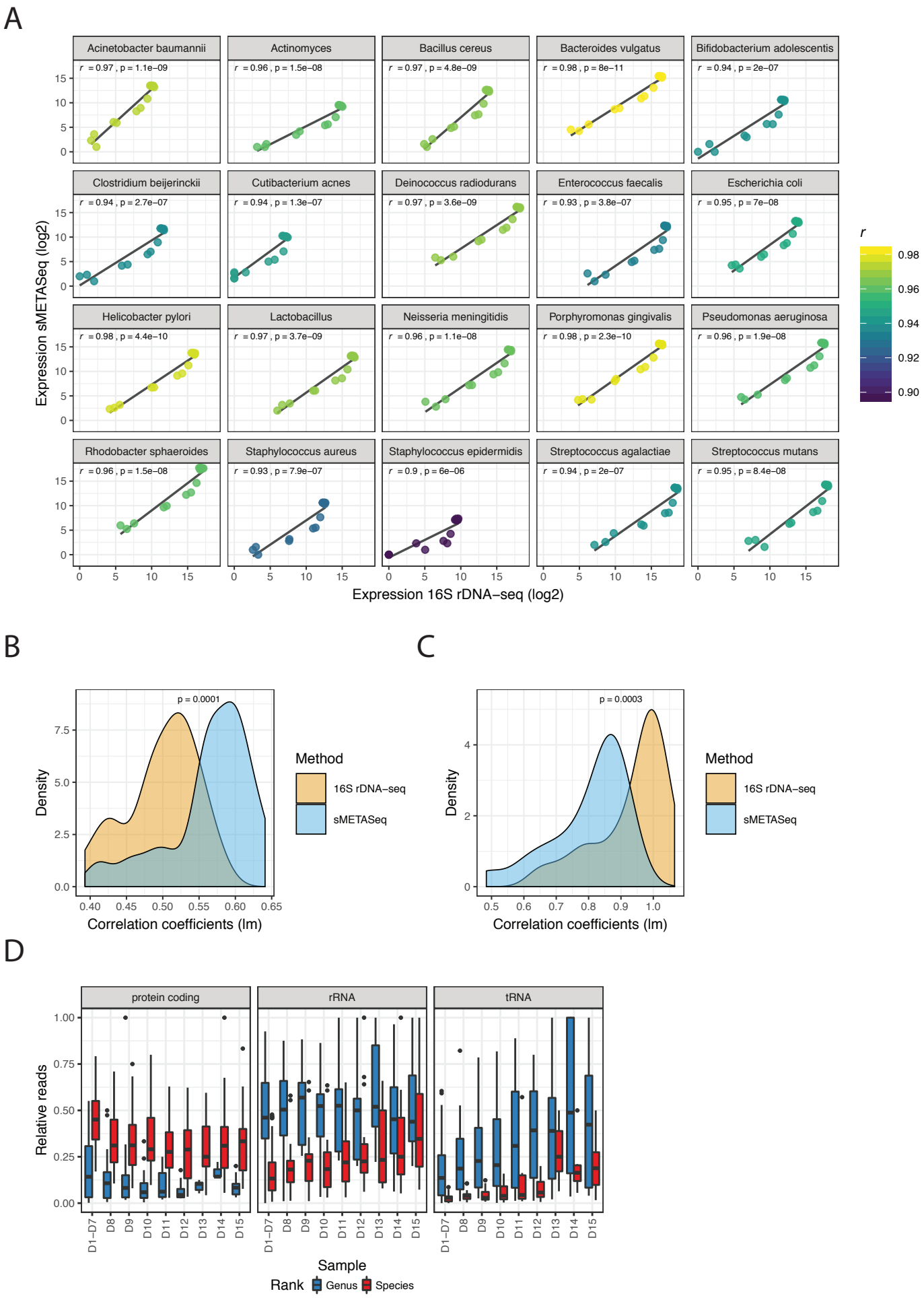

A
Acinetobacter baumannii
r = 0.97 , p = 1.1e−09
Actinomyces
r = 0.96 , p = 1.5e−08
Bacillus cereus
r = 0.97 , p = 4.8e−09
Bacteroides vulgatus
r = 0.98 , p = 8e−11
Bifidobacterium adolescentis
r = 0.94 , p = 2e−07
Clostridium beijerinckii
r = 0.94 , p = 2.7e−07
Cutibacterium acnes
r = 0.94 , p = 1.3e−07
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 mutants
r = 0.95 , p = 8.4e−08
Expression sMETASeq (log2)
Expression 16S rDNA−seq (log2)
r
0.98
0.96
0.94
0.92
0.90
B
p = 0.0001
Density
Correlation coefficients (lm)
Method
16S rDNA−seq
sMETASeq
C
p = 0.0003
Density
Correlation coefficients (lm)
Method
16S rDNA−seq
sMETASeq
D
protein coding
rRNA
tRNA
Relative reads
D1−D7
D8
D9
D10
D11
D12
D13
D14
D15
Sample
Rank
Genus
Species

Figure S5

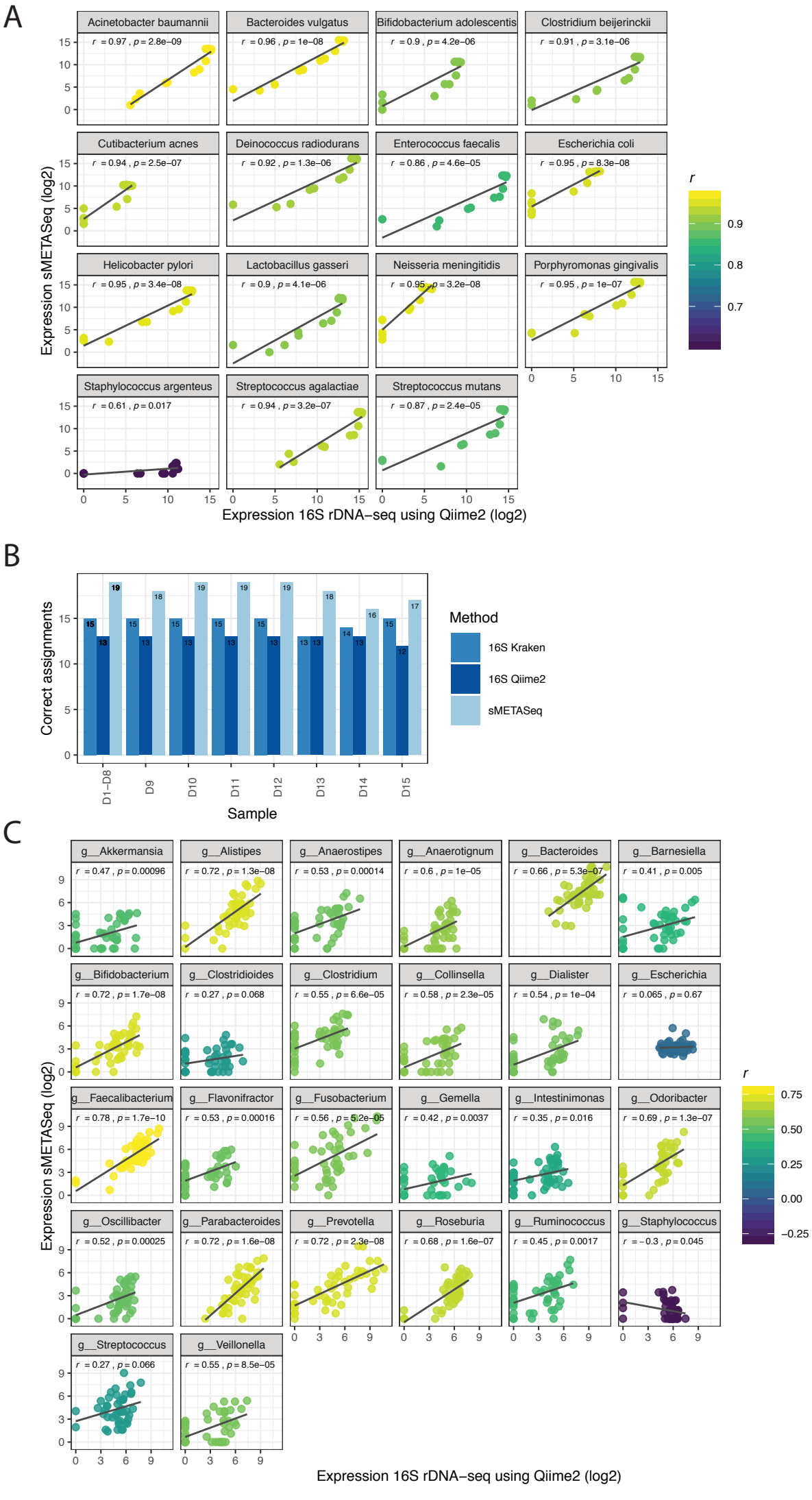

A
Acinetobacter baumannii
r = 0.97 , p = 2.8e−09
Bacteroides vulgatus
r = 0.96 , p = 1e−08
Bifidobacterium adolescentis
r = 0.9 , p = 4.2e−06
Clostridium beijerinckii
r = 0.91 , p = 3.1e−06
Cutibacterium acnes
r = 0.94 , p = 2.5e−07
Deinococcus radiodurans
r = 0.92 , p = 1.3e−06
Enterococcus faecalis
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)
r
0.9
0.8
0.7
B
Correct assignments
Sample
D1−D8
D9
D10
D11
D12
D13
D14
D15
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)
r
0.75
0.50
0.25
0.00
−0.25

Figure S6

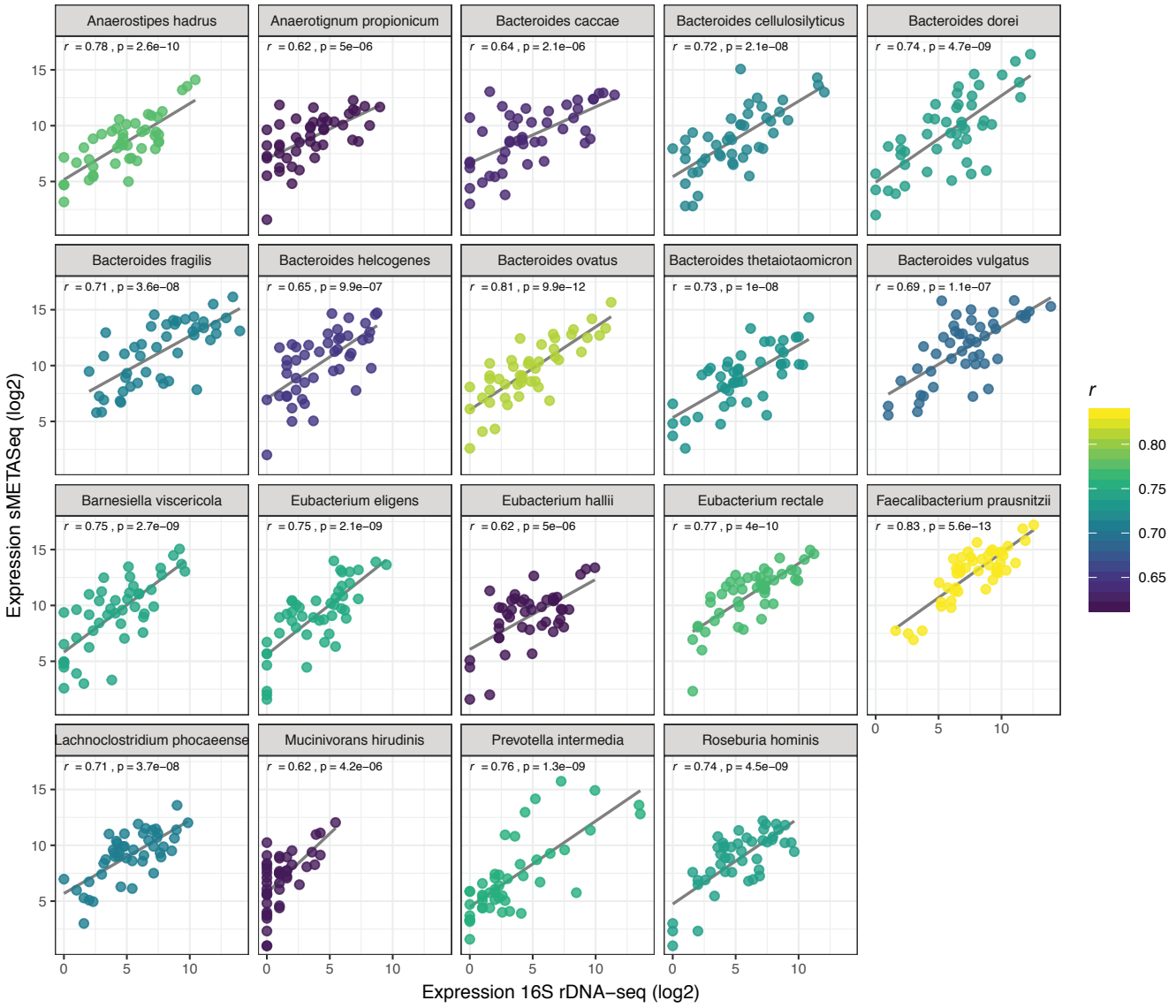

Anaerostipes hadrus
r = 0.78 , p = 2.6e−10
Anaerotignum propionicum
r = 0.62 , p = 5e−06
Bacteroides caccae
r = 0.64 , p = 2.1e−06
Bacteroides cellulosilyticus
r = 0.72 , p = 2.1e−08
Bacteroides dorei
r = 0.74 , p = 4.7e−09
Bacteroides fragilis
r = 0.71 , p = 3.6e−08
Bacteroides helcogenes
r = 0.65 , 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)
r
0.80
0.75
0.70
0.65

Figure S7

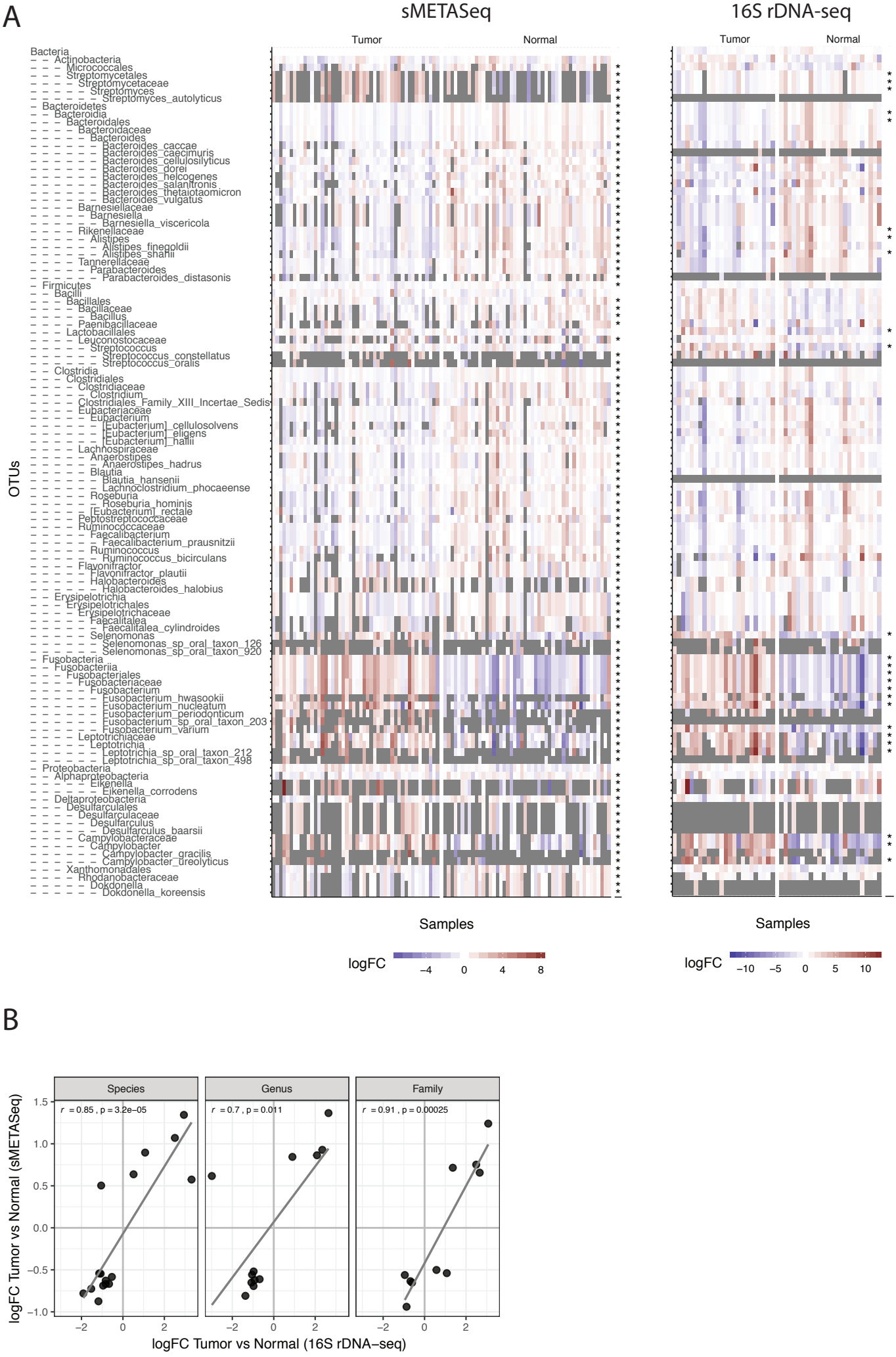
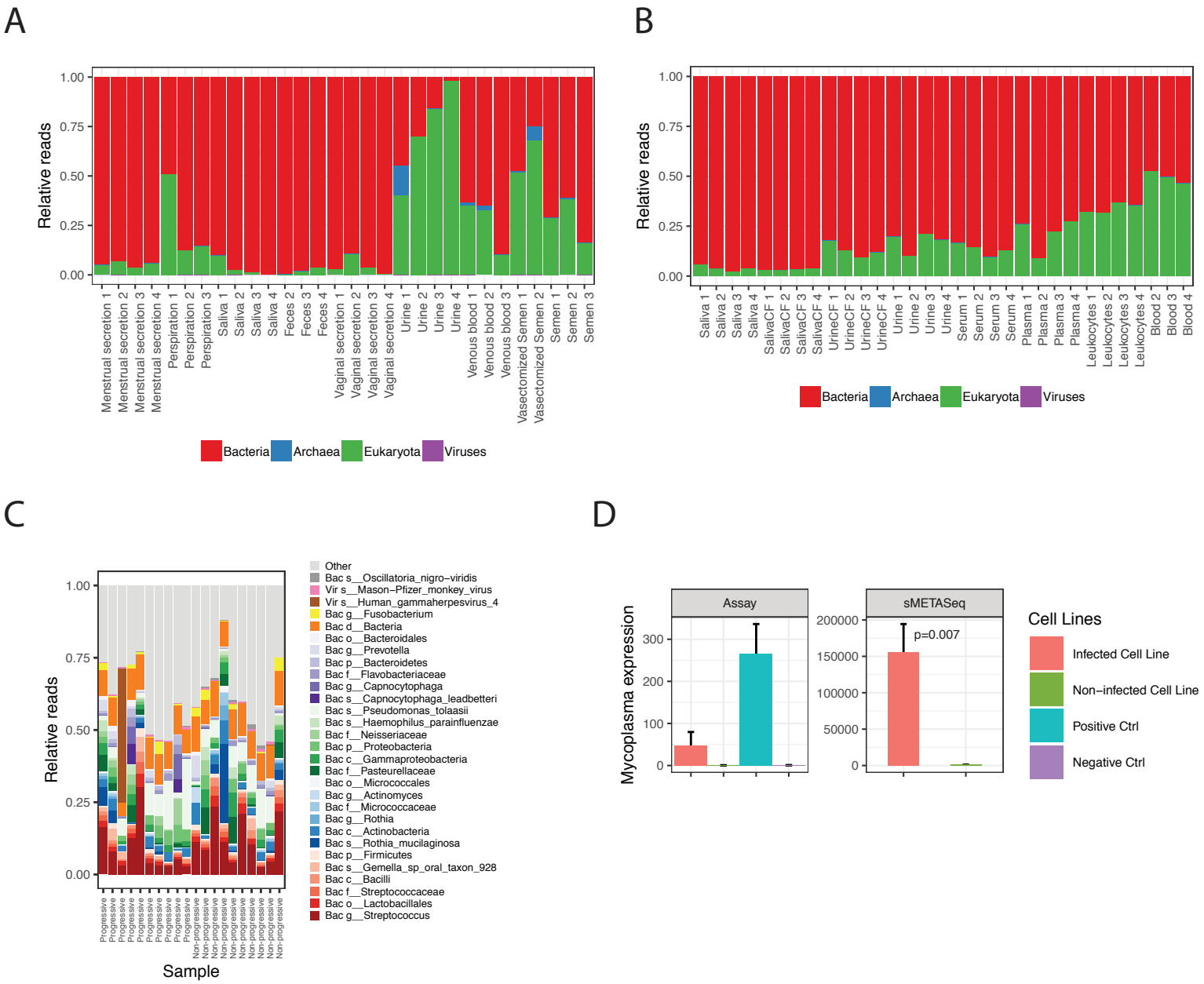

A
sMETASeq
16S rDNA-seq
Tumor
Normal
OTUs
Samples
logFC
B
Species
Genus
Family
r = 0.85 , p = 3.2e−05
r = 0.7 , p = 0.011
r = 0.91 , p = 0.00025
logFC Tumor vs Normal (sMETASeq)
logFC Tumor vs Normal (16S rDNA−seq)

Figure S8


A
Relative reads
1.00
0.75
0.50
0.25
0.00
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
1.00
0.75
0.50
0.25
0.00
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
1.00
0.75
0.50
0.25
0.00
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
sMETASeq
300
200
100
0
200000
150000
100000
50000
0
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.