





RESEARCH

Open Access



Rational probe design for efficient rRNA depletion and improved metatranscriptomic analysis of human microbiomes

Asako Tan^{1†}, Senthil Murugapiran^{2†} , Alaya Mikalauskas^{3†}, Jeff Koble⁴, Drew Kennedy⁴, Fred Hyde¹, Victor Ruotti¹, Emily Law², Jordan Jensen² , Gary P. Schroth⁴, Jean M. Macklaim³, Scott Kuersten^{1*}, Brice LeFrançois^{3*}  and Daryl M. Gohl^{2,5,6*} 

Abstract

The microbiota that colonize the human gut and other tissues are dynamic, varying both in composition and functional state between individuals and over time. Gene expression measurements can provide insights into microbiome composition and function. However, efficient and unbiased removal of microbial ribosomal RNA (rRNA) presents a barrier to acquiring metatranscriptomic data. Here we describe a probe set that achieves efficient enzymatic rRNA removal of complex human-associated microbial communities. We demonstrate that the custom probe set can be further refined through an iterative design process to efficiently deplete rRNA from a range of human microbiome samples. Using synthetic nucleic acid spike-ins, we show that the rRNA depletion process does not introduce substantial quantitative error in gene expression profiles. Successful rRNA depletion allows for efficient characterization of taxonomic and functional profiles, including during the development of the human gut microbiome. The pan-human microbiome enzymatic rRNA depletion probes described here provide a powerful tool for studying the transcriptional dynamics and function of the human microbiome.

Keywords Next-generation sequencing, Microbiome, Metatranscriptomics, rRNA depletion

Background

The microbiome plays a critical role in human health and disease [1]. Over the past decade, next-generation sequencing-based analyses have provided insights into the composition of the microbiome across body sites and life stages and have begun to uncover correlations between microbial taxa or microbial functions and disease states [2–4]. Beyond genomic analysis of microbiome composition, multi-omic data incorporate measurements of the microbiota-associated transcriptome, proteome, or metabolome to provide further insights into microbiome activity and function. Although metagenomic and metatranscriptomic profiles tend to be generally consistent, microbial functional profiles derived from DNA sequencing are more conserved across donors than transcriptional profiles, which are highly donor specific [5].

[†]Asako Tan, Senthil Murugapiran and Alaya Mikalauskas contributed equally to this work.

*Correspondence:

Scott Kuersten

SKuersten@illumina.com

Brice LeFrançois

brice.lefrancois@dnagenotek.com

Daryl M. Gohl

dmgohl@umn.edu

¹ Illumina, Inc, Madison, WI 53719, USA

² Diversigen, Inc, New Brighton, MN 55112, USA

³ DNA Genotek, Inc, Ottawa, ON K2V 1C2, Canada

⁴ Illumina, Inc, San Diego, CA 92122, USA

⁵ University of Minnesota Genomics Center, Minneapolis, MN 55455, USA

⁶ Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455, USA



Importantly, many broadly encoded metagenomic pathways are expressed by a small number of organisms, highlighting the utility of metatranscriptomics to identify functional activities [6]. In particular, transcriptomic measurements of the human gut associated microbiome have been used to study microbial carbohydrate metabolism [7] and to provide functional information about intestinal diseases such as IBD [8] as well as mechanisms of drug metabolism [9].

Acquiring metatranscriptomic data is hindered by the fact that the vast majority of microbial-derived RNA molecules correspond to ribosomal RNA (rRNA) [10]. In eukaryotes, non-ribosomal RNA can be easily and efficiently enriched through selective reverse transcription or pull-down approaches that target the poly-A tail or using probes to specifically bind rRNA molecules prior to removal by capture or enzymatic digestion [11, 12]. Although Poly-A polymerase was first isolated from *Escherichia coli* [13, 14], bacterial mRNA transcripts are not, as a rule, poly-adenylated, and when poly-adenylation does occur it is associated with RNA degradation [15, 16]. Thus, for bacterial samples, selective enrichment of mRNA is not easily achievable and the removal of rRNA must be accomplished by other means.

While a large number of studies have developed efficient methods to deplete rRNA in individual bacterial species using probe-based capture [17], enzymatic depletion [18], or CRISPR-based methods [19, 20], depleting the diverse rRNA sequences in complex human microbiome samples that can contain hundreds of species presents a significant technical challenge. In addition, the composition of the microbiota varies substantially across body sites and throughout different life stages, further expanding the taxonomic coverage required for robust depletion of rRNA across human microbiome samples. Probe-based sequence capture methods, such as were employed with Illumina's Ribo-Zero Gold kit can provide strong rRNA depletion across a variety of sample types, including human gut microbiome samples [21]. However, such probes are costly, difficult to manufacture, and tend to perform best with high quality RNA samples. Moreover, capture-based rRNA depletion methods can lead to inconsistent results based on operator skill. These factors led to the discontinuation of Illumina's capture-based bacterial Ribo-Zero Gold (Epidemiology) depletion kit.

Here we describe the development of a pan-human microbiome probe set for efficient and consistent enzymatic (RNase H) microbial rRNA depletion. Through an iterative design process, we developed probes that effectively deplete rRNA found in human oral, vaginal and adult and infant gut microbiome samples, substantially improving mapping rates to coding microbial gene databases. Using defined spike-ins, we demonstrate that the

rRNA depletion process does not introduce substantial bias in the gene expression profiles. In addition, we use the resulting metatranscriptomic data to refine informatic pipelines for rRNA and host mapping and to examine gene expression and functional activity across human microbiome sites. The method described here circumvents the limitations of sequence capture methods and represents a highly effective rRNA depletion option for metatranscriptomic studies of human-associated microbial communities.

Results

Assessment of informatics tools for robust rRNA alignment and filtering

Accurate determination of the fraction of reads in a sample mapping to rRNA is critical to truly assess the rRNA depletion efficacy of existing and emerging methods. Several studies have compared the efficiency of various rRNA depletion methods but each tends to rely on specific bioinformatic tools making comparison across studies difficult [18, 22, 23]. Since no extensive comparison of mapping tool performance with respect to rRNA detection is currently available, we first assessed the ability of five different mapping tools (bowtie2, bbdduk, seal and bbsplit) [24, 25] together with four rRNA databases SILVA [26] SSU/LSU NR99 (ar), SILVA SSU/LSU NR99 + tRNA RFAM clans (art), SortMeRNA [27] (4.3) default database (ds), and SortMeRNA [27] (4.3) sensitive database (ss) to reliably classify rRNA reads in depleted and undepleted samples. We examined data from a set of ten published samples [28], seven of which had high levels of rRNA and three of which had low levels of rRNA (Fig. 1, Figure S2). bbdduk, seal, and bbsplit were the fastest tools in our assessment and considerably faster than bowtie2 and SortMeRNA (data not shown). We detected substantial differences in performance across tools, with several of them (bowtie2, bowtie2 ran in local mode) showing elevated false negative rates and misclassifying rRNA as mRNA in high rRNA samples. Other tools (SortMeRNA) had elevated false positive rates, misclassifying mRNA as rRNA in low rRNA samples (Fig. 1, Figure S2). In addition, bbdduk had better efficiency in classifying mRNA and ncRNA correctly than seal and bbsplit (Fig. 1, Figure S2). Thus, we chose to use bbdduk with SILVA SSU/LSU NR99 version 138.1 + tRNA RFAM clans for assessment of rRNA content in subsequent analyses.

Standard enzymatic Ribo-Zero Plus workflow does not efficiently deplete rRNA from complex human microbiome samples

The discontinued Illumina Ribo-Zero Gold (Epidemiology) kits used a probe-based hybridization approach

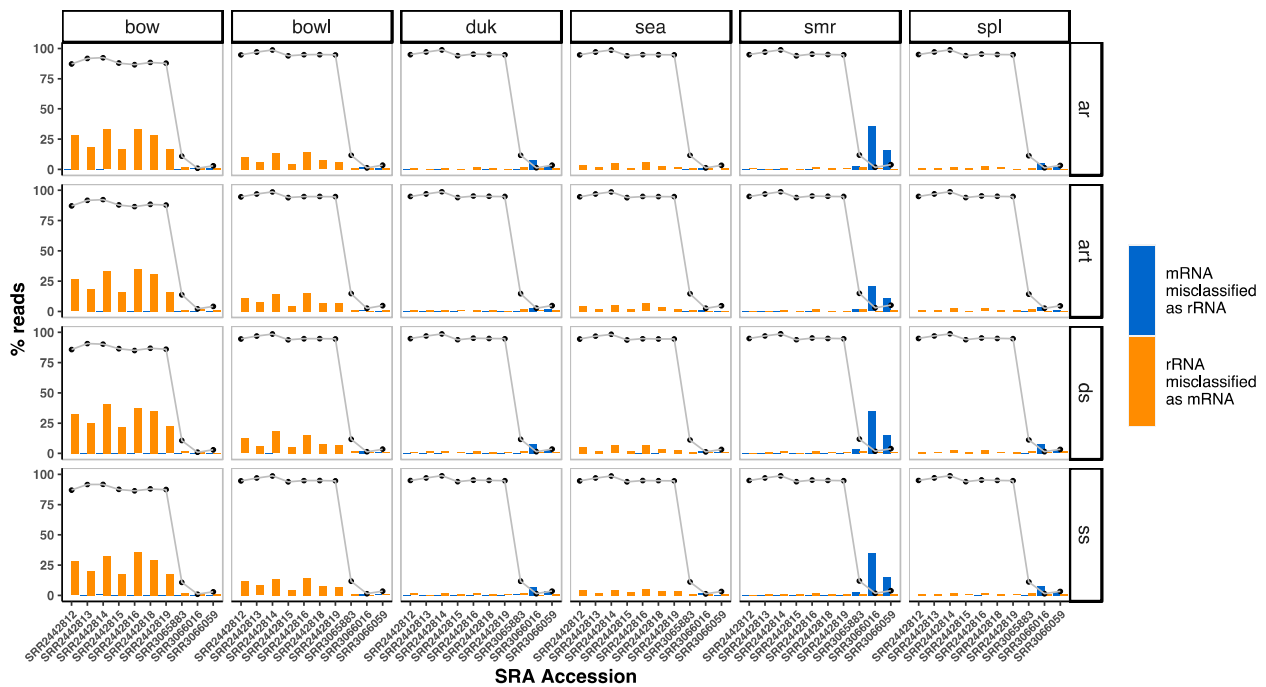


Fig. 1 Comparison of mapping tools for classifying rRNA reads from metatranscriptomes. Bar plots showing fractions classified as either rRNA or ‘clean’ (mRNA) by the various tools (bow: Bowtie2; bowl: Bowtie2 local mode; duk: BBDuk; sea: Seal; smr: SortMeRNA; spl: bbsplit) searched against four different databases (ar: SILVA SSU/LSU NR99; art: ar + tRNA RFAM clans; ds: SortMeRNA (4.3) default database; ss: SortMeRNA (4.3) sensitive database). The line plots represent the percentage of rRNA reads in the sample identified by each method

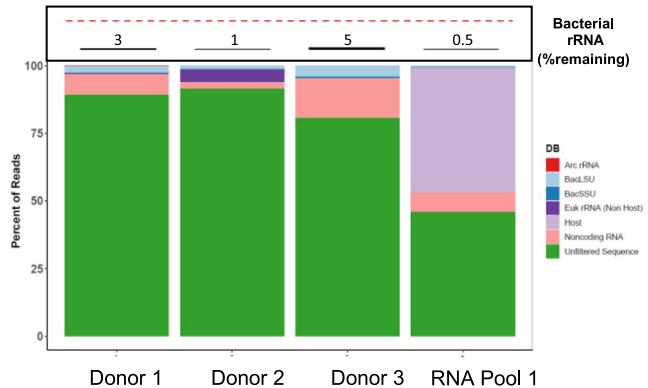
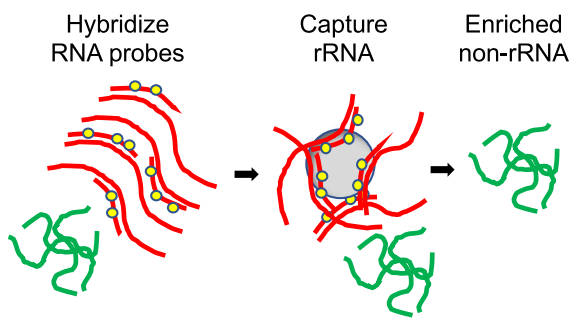
to capture and deplete human and bacterial rRNAs (Fig. 2A). Ribo-Zero Gold (Epidemiology) was able to substantially deplete rRNA from complex human microbiome samples such as stool, as well as from a defined pooled microbial RNA sample made up of a mix of human and bacterial total RNA (Fig. 2B—see Materials and Methods for pool composition). The percentage of bacterial rRNA reads in Ribo-Zero Epidemiology samples was lower than 5% suggesting that most of the reads in these samples corresponded to bacterial and/or human mRNA. In comparison, the Ribo-Zero Plus kit, which relies on probe hybridization followed by RNase H enzymatic depletion (Fig. 2C) performed well with the RNA pool sample but failed to substantially deplete stool samples (Fig. 2D). Between 65–85% of the sequencing reads from Ribo-Zero Plus treated samples corresponded to bacterial rRNA (Fig. 2D), indicating poor overall depletion efficiency for stool samples. Depletion was more successful for the less complex RNA pool sample, where only <7% of the sequencing reads mapped to the SILVA rRNA database [26]. The fact that more complex microbiome samples were less effectively depleted suggested that the standard probe content in Ribo-Zero Plus (probes targeting *E. coli* and *B. subtilis* rRNA) provided insufficient

coverage to target the greater diversity of rRNA found in human gut microbes for enzymatic degradation.

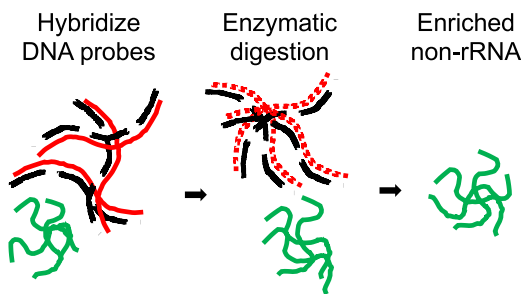
Iterative design of a microbiome depletion oligo pool enables robust rRNA depletion from human stool samples

To improve enzymatic depletion using the Ribo-Zero Plus kit we used an iterative design process to generate additional probes specifically targeting human gut microbiome samples (Fig. 2E, Figure S3). We used sequencing data from stool samples depleted with the standard Ribo-Zero Plus kit and identified the most abundant rRNA sequences that were not effectively depleted across 9 adult healthy stool RNA samples. After eliminating redundancy within these sequences and redundancy with the initial Ribo-Zero Plus probe set (DP1), we designed and synthesized a novel probe set (human microbiome pool, HMv1). Addition of HMv1 probes to Ribo-Zero Plus depletion reactions dramatically improved bacterial rRNA depletion of a group of ten test stool samples that were poorly depleted by DP1 alone (Fig. 3A). The percentage of bacterial rRNA reads was >70% on average for samples depleted with DP1, while the percentage of rRNA reads was <17% on average in samples depleted with the supplemental HMv1 oligo pool (Fig. 3A). In contrast, for undepleted stool samples bacterial rRNA

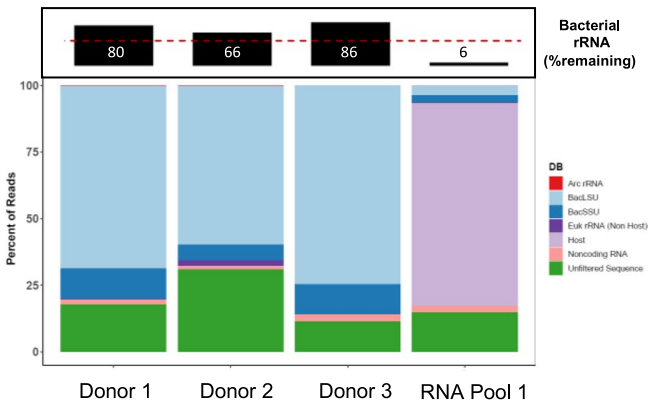
A RiboZero Gold (Epidemiology) B



C RiboZero Plus



D



E

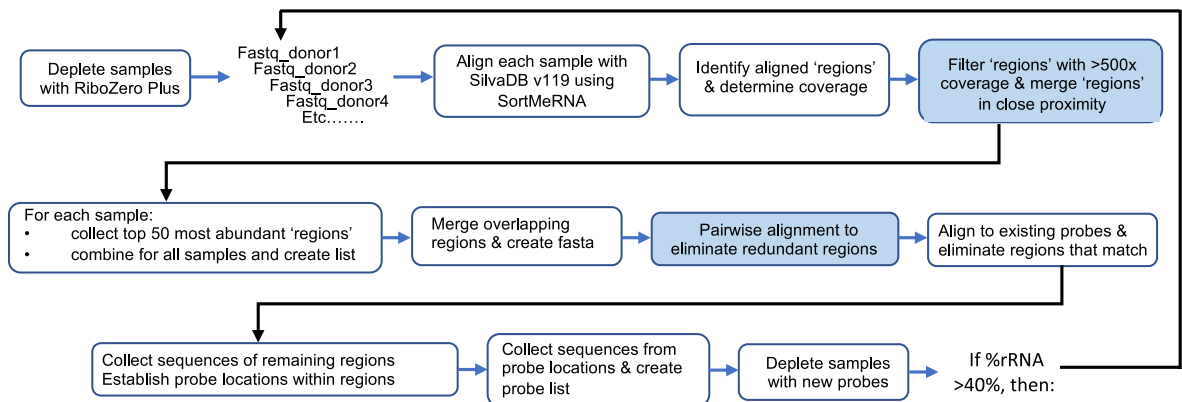


Fig. 2 Comparison of Ribo-Zero Gold (Epidemiology) and Ribo-Zero Plus performance for rRNA removal from total RNA extracted from stool samples versus a synthetic RNA pool containing a mix of bacterial and human RNA. **A** Overview of the Ribo-Zero Gold (Epidemiology) workflow that captures rRNA using biotin labeled anti-sense RNA probes captured by streptavidin magnetic beads for removal of rRNA from the sample. **B** Percentage of reads mapping to eukaryotic and prokaryotic rRNAs vs. coding sequences following depletion of adult stool samples and RNA standard with Ribo-Zero Gold (Epidemiology). Boxed bar plots on top represent remaining bacterial rRNA (LSU & SSU) in each sample with a dashed line at the 50% mark. **C** Overview of the Ribo-Zero Plus method where anti-sense DNA oligonucleotides are hybridized to rRNAs in the sample prior to enzymatic digestion of the rRNA:DNA duplexes with RNase H. **D** Percentage of reads mapping to eukaryotic and prokaryotic rRNAs vs coding sequences following depletion of adult stool samples and RNA standard with Ribo-Zero Plus. Boxed bar plots on top represent remaining bacterial rRNA (LSU & SSU) in each sample with a dashed line at the 50% mark. **E** Diagram of the iterative probe design process, starting from raw sequencing data of samples depleted with the standard Ribo-Zero Plus probe set (DP1). The steps shaded in blue refer to the data shown in Supplemental Figure S3A and B

represented over 98% of all reads. The abundant taxa (Fig. 3B) and functional profiles (Figure S4) detected in the metatranscriptomic data resulting from HMv1 depletion are consistent with published human gut microbiome data, with a high abundance of taxa such as *Faecalibacterium*, *Lachnospiraceae*, and *Clostridium*. Addition of the HMv1 probes improved the number of mapped non-rRNA reads and increased the number of taxa and functional features observed (Fig. 3C).

Pan-human microbiome oligo pool allows robust rRNA depletion of diverse human microbiome samples

Microbial community structure varies widely between donors and across body sites [29]. To assess the performance of the HMv1 probe pool across a variety of human microbiome sample types, we first depleted mock microbial communities using Ribo-Zero Plus. The Ribo-Zero Epidemiology kit was used as a control in these experiments and tested with skin (ATCC MSA-2005) and gut (ATCC MSA-2006) mock communities. These cell-based mock communities were composed of microbes commonly found on human skin (*Corynebacterium*, *Cutibacterium*, *Staphylococcus* and *Streptococcus*) or in the human gut (*Bacteroides*, *Bifidobacterium*, *Clostridioides*, *Enterococcus*). The skin and to a lesser extent the gut mock communities were effectively depleted with the Ribo-Zero Epidemiology kit (Figure S5). However, variability was high between replicates especially for the gut mock. This is a well-known issue with probe capture-based workflows, where operator-dependent variables such as the amount of bead carry-over can greatly impact rRNA depletion. In contrast, Ribo-Zero Plus depletions were more reproducible (Figure S5). Interestingly, the standard Ribo-Zero Plus probe set efficiently depleted rRNA from the gut mock community (~15% of reads mapping to rRNA) but did not perform as well with the skin mock (~50% reads mapping to rRNA). Poor performance of the Ribo-Zero Plus kit with the lower complexity skin mock community was unexpected, but further analysis revealed that it was due to inability to target *Corynebacterium striatum* and *Micrococcus striatum* rRNA sequences. Addition of the HMv1 probe set to the Ribo-Zero Plus reactions greatly improved depletion

efficiency of the skin communities with <2% of reads mapping to rRNA (Figure S5). This demonstrates that the HMv1 custom probe set provides coverage against rRNA sequences from these two species (see Materials and Methods).

To further test the efficacy of the HMv1 probe set on human microbiome samples, we depleted human oral (tongue) and vaginal microbiome samples using the Ribo-Zero plus kit. The standard Ribo-Zero Plus kit was able to efficiently deplete rRNA of half of the vaginal samples V1-V3 but performed sub-optimally for samples V4-V6 (Fig. 4A). These three samples depleted significantly better following addition of HMv1. Moreover, the standard Ribo-Zero Plus kit (without supplemental probes) also performed poorly for higher complexity oral microbiome samples (Fig. 4B, Figure S6). Addition of HMv1 for oral microbiome samples led to a decrease of bacterial rRNA reads from approximately 45% to 5% on average (Fig. 4B and Figure S6). qPCR and taxonomic analysis of the vaginal metatranscriptomic profiles showed that samples efficiently depleted by the standard probe set were dominated by *Lactobacillus* (V2 & V3) or *Corynebacterium* (V1), while samples benefiting from addition of HMv1 had generally a more complex community structure and higher relative abundance of *Gardnerella*, *Bifidobacterium*, and *Olsenella* (Fig. 4C and Figure S6). In comparison, metatranscriptomic profiles and diversity of tongue microbiome samples were highly consistent across donors, with *Veillonella*, *Rothia*, *Streptococcus* and *Prevotella* being the most abundant genera (Fig. 4D). Taken together, our data demonstrates that the custom HMv1 probe pool improved rRNA depletion of complex human microbiome samples through increased coverage of bacterial rRNA sequences found across various sites.

Additional probes improve rRNA depletion of infant stool microbiome samples

Human gut microbiome profiles are known to change rapidly during the first few years of life [30]. In young infants, the gut microbiota is significantly different from adult samples and tends to be dominated by different taxa such as Bifidobacteria [31]. To determine if the Ribo-Zero Plus HMv1 probe set could efficiently remove rRNA

(See figure on next page.)

Fig. 3 The Pan-Human Microbiome probe pool, HMv1, effectively depletes rRNA in stool samples from healthy adult donors to generate metatranscriptomic profiles. **A** Percentage of reads mapping to eukaryotic and prokaryotic rRNAs vs. coding sequences in samples depleted with Ribo-Zero Plus standard probes (DP1) vs combination of DP1 and HMv1 probes (stool RNA samples from 10 healthy adult donors). Boxed bar plot on top represents remaining bacterial rRNA (LSU & SSU) for either DP1-depleted samples (black bars) or HMv1-depleted samples (grey bars) with a dashed line at the 50% mark. **B** Taxonomic heatmap showing the 20 most abundant taxa for each sample based on metatranscriptomic analysis of adult stool samples. **C** Number of reads, taxonomic and functional features detected in undepleted, DP1-depleted, or HMv1-depleted stool samples

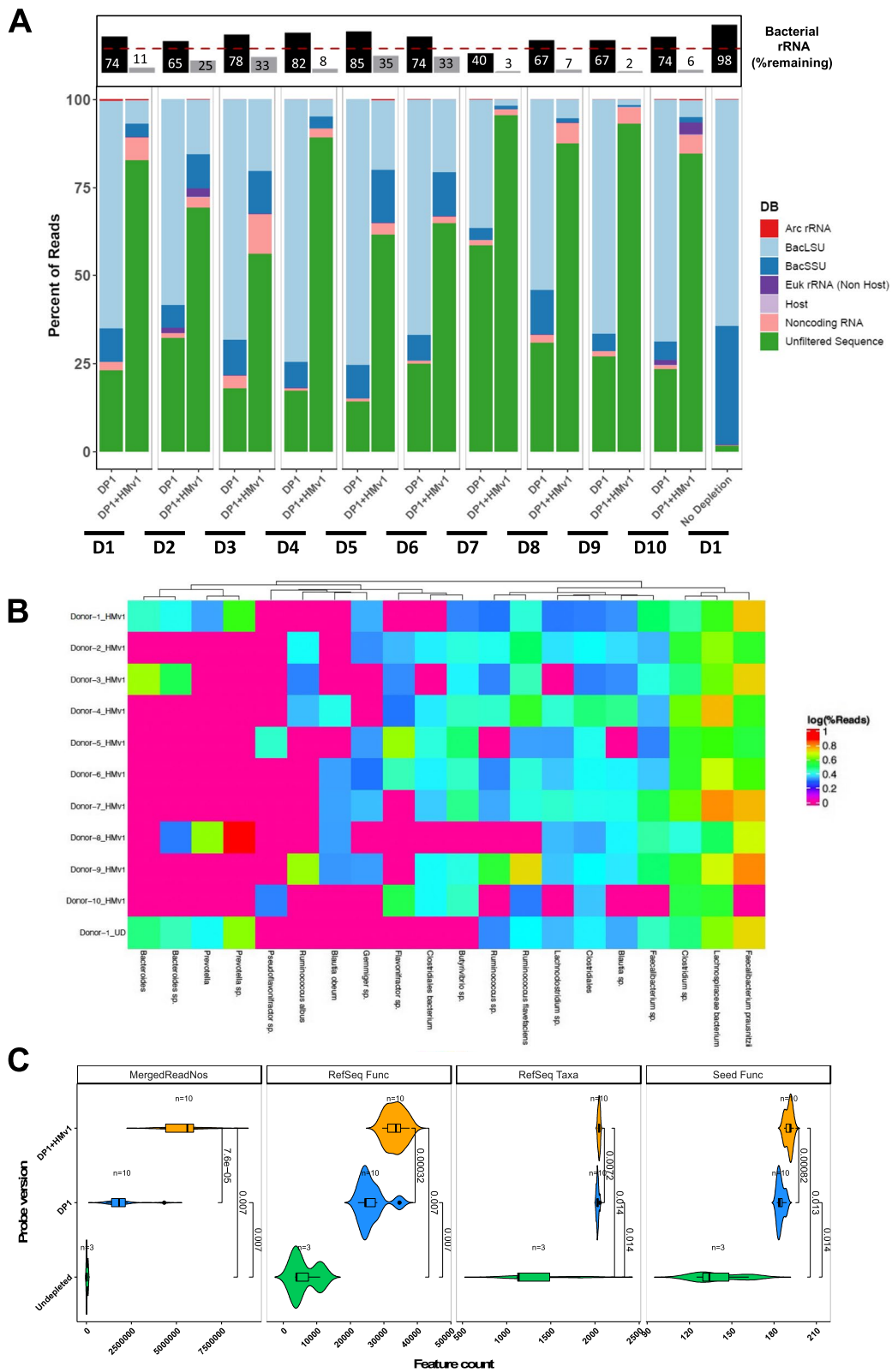


Fig. 3 (See legend on previous page.)

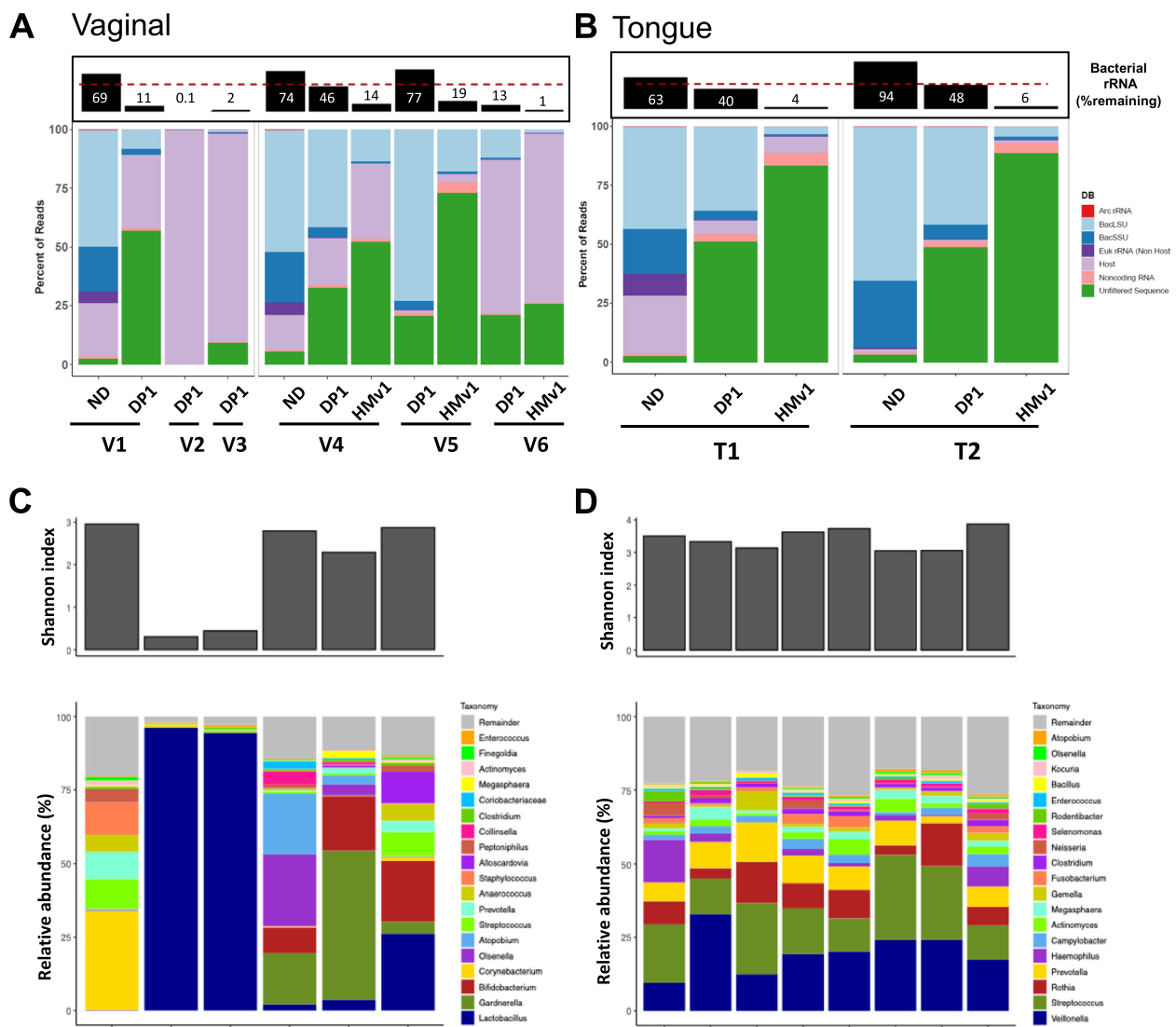


Fig. 4 The Human microbiome pool probe set can effectively deplete rRNAs found in vaginal and oral microbiome samples to generate metatranscriptomic profiles. Percentage of reads mapping to eukaryotic and prokaryotic rRNAs vs. coding sequences following depletion of vaginal and tongue samples using Ribo-Zero Plus ± pan-microbiome HMv1 probes. **A** rRNA depletion efficiency of vaginal samples using the Ribo-Zero standard probe set (DP1) ± pan-microbial HMv1 probe set. Non-rRNA host RNA content can be prominent in vaginal samples. ND=Non depleted controls. **B** rRNA depletion efficiency of representative tongue microbiome samples (T1 & T2) using the Ribo-Zero Plus standard probe set (DP1) ± pan-microbiome HMv1 probe set. Boxed bar plots on top in panels **A** and **B** represent remaining bacterial rRNA (LSU & SSU) for indicated samples with a dashed line at the 50% mark. **C** Alpha diversity (Shannon index) and metatranscriptomic taxonomic profiles for Ribo-Zero Plus (DP1 or DP1 + HMv1) depleted vaginal microbiome samples. **D** Alpha diversity (Shannon index) and metatranscriptomic taxonomic profiles for Ribo-Zero Plus (DP1 + HMv1) depleted oral microbiome samples

in infant gut microbiome samples, we extracted total stool RNA collected from infants and young children aged 4 to 33-months. The HMv1 probe pool led to efficient and consistent rRNA depletion of most infant stool samples (Fig. 5A and Figure S7) with < 26% of reads mapping to bacterial rRNA on average. Interestingly, rRNA depletion was less efficient for a subset of donors (infants E, I, J and L) primarily in the 9 to 14-months old group.

Taxonomic analysis revealed that these samples had high levels of *Bifidobacterium bifidum*. Lack of depletion suggests that the HMv1 probe set poorly targets rRNA from this species. Additional probes targeting the *Bifidobacterium bifidum* rRNA sequence were designed using our iterative process and added to HMv1 probe to create a second human microbiome pool (HMv2). To test the effectiveness of the new probe set, RNA samples from

infants E, I, J and L were depleted with the Ribo-Zero Plus kit using HMv1 or HMv2. The HMv2 probe set was able to improve rRNA depletion in all 4 samples compared to HMv1 (inset Fig. 5A, Figure S7), with average bacterial rRNA content decreasing from 35 to 16%. The HMv2 pool was also tested on adult samples and compared to HMv1, leading to slightly better depletion efficiency in some samples (Figure S8). Unlike infant stool samples, better depletion in adult samples was not attributable to improved depletion of *Bifidobacteria*, but rather to improved rRNA coverage and depletion of unrelated taxa such as *Blautia*. This demonstrates that additional bacterial rRNA sequences can be targeted by the iterative probe design process to further improve the performance of the human microbiome probe pool. Despite marginally better rRNA depletion, HMv2 did not significantly impact the number of features detected relative to HMv1, however, the total numbers of raw reads were different in these experiments making a direct comparison difficult (Figure S9).

Stool metatranscriptomic analyses reveal significant differences in functional profiles across age groups

The effective depletion of rRNA from both infant and adult stool samples provided an opportunity to compare the taxonomic and functional profiles of these stool samples across age groups. Infants < 6 months (A, B, C and D), generally had a lower Shannon index than older infants (Fig. 5B – top panel) and their metatranscriptomic taxonomic profiles were largely dominated by *Bifidobacteria* (Fig. 5B – bottom panel). *Bifidobacteria* relative abundance started decreasing in young children > 6 months old, concomitantly with the appearance of canonical gut commensal genera such as *Faecalibacterium* and *Bacteroides* and an increase in alpha diversity (Fig. 5B). Interestingly, differential abundance analysis revealed higher prevalence of *Escherichia*, *Veillonella*, *Klebsiella*, and *Shigella* species in younger infants, taxa not typically associated with healthy gut samples, as well as > 20 species of the *Enterococcus* genus (Figure S10). In contrast, older infants displayed higher prevalence of several species of the *Alistipes* and *Akkermansia* genera, as well as

Mucinivorans hirsutinis and *Clostridium viride*. Comparison of the genes differentially expressed in infants and children of varying ages also showed significant differences, spanning many major enzymatic groups (Figure S11–12). To further understand how gut metatranscriptomic profiles evolve with age, we compared gene expression in infants (< 6 months) and young children (> 20 months). Infants displayed higher levels of enzymes involved in amino sugar metabolism, glycolysis, pyruvate and succinate metabolism, protein translation and cell growth (Fig. 5C, D, Supplemental Data File 2). Samples from > 20 month children displayed different metabolic profiles, with a large representation of genes involved in the degradation of amino acids such as glycine and lysine, and butanoate fermentation (Fig. 5D, Supplemental Data File 2). Genes involved in sporulation and motility were also significantly increased in > 20 month old children, correlating with the appearance of spore-forming and/or motile bacteria (Fig. 5B). Interestingly, both young infants and > 20 month old children displayed strong upregulation of stress response pathways (cold shock proteins, chaperones, universal stress protein, two-component system), yet specific gene expression appeared to remain quite distinct in both groups (Supplemental Data File 2). We also compared gene expression in adult versus infant stool samples and observed marked differences in metabolic activity. Many enzymes of the glycolysis and the pentose phosphate pathways were increased in infants as well as downstream metabolic enzymes (such as pyruvate oxidase, pyruvate dehydrogenase, and lactate dehydrogenase) (Supplemental Data File 3). Other upregulated genes included sugar transporters (galacto-N-biose-/lacto-N-biose ABC transporter) as well as several genes associated with cell division. Stress response pathways were heavily represented in infants with > 10 chaperone/cold shock proteins showing increased expression (Supplemental Data File 3). In contrast, adult samples showed little to no stress response but expressed high levels of a wide variety of sugar transporters and glycosylases (Supplemental Data File 3). Additionally, adult samples also showed high expression of genes involved in epithelium invasion/adherence, quorum sensing and

(See figure on next page.)

Fig. 5 Iteration on the Human Microbiome probe pool to achieve optimal depletion of infant stool samples for metatranscriptomic analysis.

A Percentage of reads mapping to eukaryotic and prokaryotic rRNAs vs coding sequences following depletion of infant stool RNA samples 4–33 months of age using Ribo-Zero Plus and HMv1 probes (Left). Bacterial rRNA depletion (LSU & SSU) efficiency using HMv1 vs HMv2, a modified probe set supplemented with 42 additional probes targeting *Bifidobacterium bifidum* rRNA for three infant samples which had suboptimal depletion with the HMv1 probes is shown in the inset on lower right. **B** Alpha diversity (Shannon index) and metatranscriptomic taxonomic profiles for HMv1 depleted stool samples from infants and children aged 4 to 33 months. **C** Differential gene expression in young (< 6 month of age) vs older infants (> 20 months) across major metabolic groups (carbohydrate metabolism, cell wall and motility). **D** Table showing carbohydrate and amino acid metabolism genes with large changes in relative expression in infants (< 6 months) versus young children (> 20 months). Positive effect size values correlate with increased expression in infants while negative values correlate with increased expression in > 20 month children

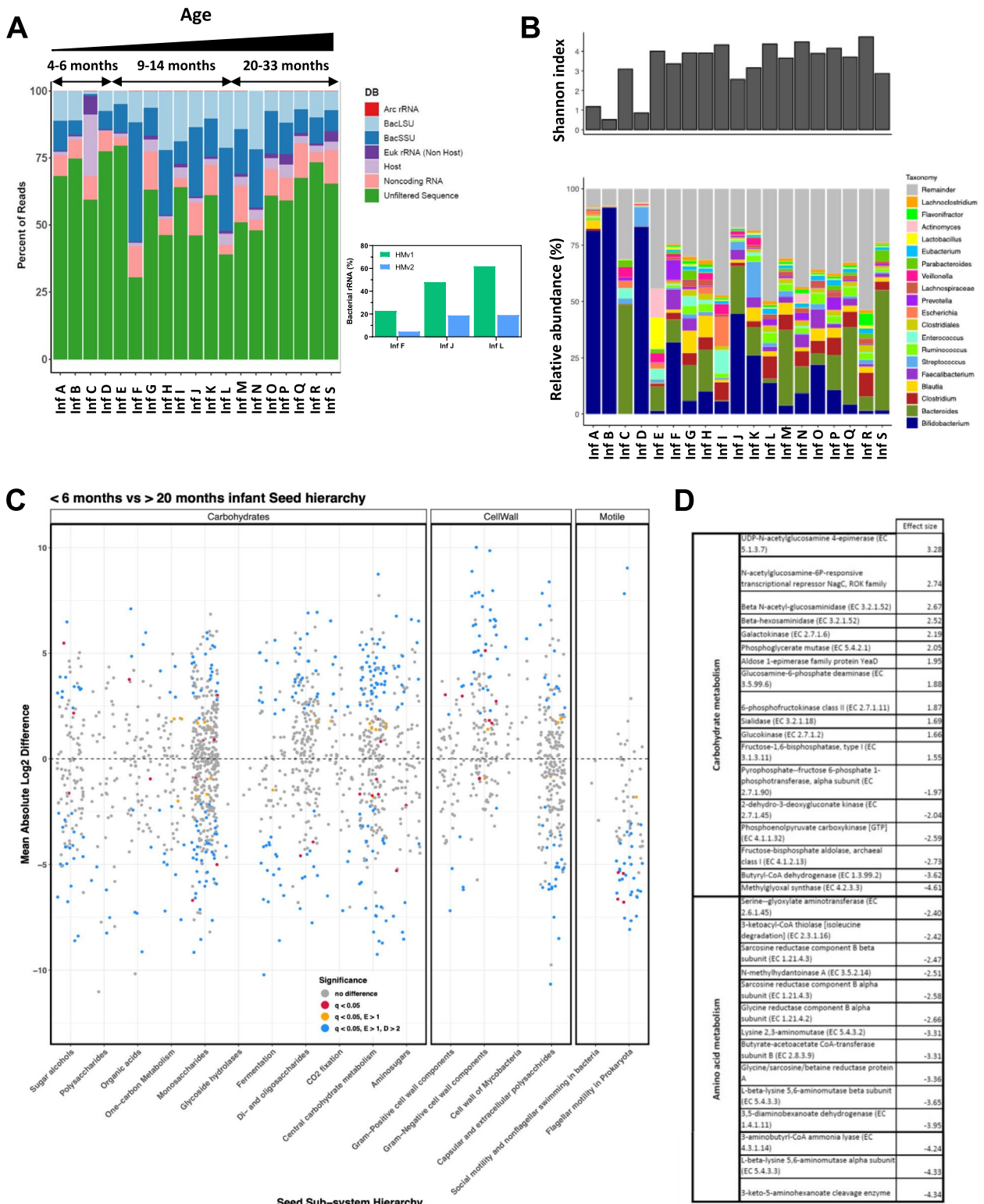


Fig. 5 (See legend on previous page.)

biofilm formation as well as increased expression of several enzymes of the benzoate degradation and acetogenic pathways (Supplemental Data File 3).

Assessing the accuracy and reproducibility of rRNA depletion using Ribo-Zero Plus with the supplemental pan-human microbiome oligo pool

In order to assess the potential impact of the rRNA depletion process on the quantitative accuracy of the resulting metatranscriptomic measurements, we carried out an experiment where synthetic External RNA Controls Consortium (ERCC) spike-ins were added to samples pre- or post-depletion with Ribo-Zero Plus (Fig. 6A). Fecal samples from 3 healthy adult donors as well as an RNA pool were tested in triplicate to assess reproducibility and accuracy of the rRNA depletion workflow. As expected, efficient rRNA depletion was seen for all samples and ERCC spike-in mRNAs represented between 5–40% of total reads (Figure S13). The higher proportion of ERCC reads recovered in samples spiked post-depletion is due to underestimation of the absolute rRNA depletion that was achieved (Figure S13). Both within and between donors, the correlation between samples spiked pre-depletion and spiked post-depletion were universally high (Fig. 6B). When replicates were averaged, all combinations of samples spiked pre-depletion had Spearman correlation coefficients higher than 0.95 (Fig. 6B, Figure S14). ERCC spike-in abundances for a representative pair of pre-depletion samples are shown in Fig. 6C. Likewise, samples that were spiked post-depletion also had Spearman correlation coefficients greater than 0.95 (Fig. 6B, D). Comparison of pre- and post-depleted samples allow assessment of the amount of quantitative error introduced by the depletion process. Both within donor (Fig. 6E) and between donor (Fig. 6F) comparisons of samples spiked with ERCC spike-ins pre- and post-depletion had lower correlation coefficients than comparisons within either the pre- or post-depleted sample sets (Fig. 6B). However, the correlation of ERCC spike-in abundances remained high for pre- vs. post-depletion samples. Thus, while the process of rRNA depletion introduces a detectable shift in the relative abundance of the ERCC standards, the quantitative accuracy and reproducibility of the resulting measurements remains high.

Discussion

Here we describe the development of probes for enzymatic rRNA depletion of human-associated microbiomes to enable metatranscriptomic analysis. First, in order to generate accurate measures of rRNA depletion of human microbiome samples, we assessed the ability of different sequence alignment algorithms to accurately classify microbial rRNA and mRNA reads across depleted and undepleted stool samples available in public databases. Bbduk had minimal false negative and false positive mapping rates across a broad range of rRNA content and was therefore used in subsequent analyses (Figs. 1, S2). Next, we used an iterative design process to design probes that effectively deplete rRNAs found in commonly studied human microbiomes (Figs. 2, S3) for the Ribo-Zero Plus workflow. This enzymatic rRNA depletion approach overcomes issues associated with the cost of synthesizing sequence capture-based rRNA depletion probes and the variability and dependence of capture-based rRNA depletion on both operator skill and RNA quality. Furthermore, this design process is based upon abundant rRNA sequences present in the samples and is not taxa-based. Ribo-Zero Plus is an enzymatic method to remove rRNA sequences and is limited by the amount of DNA probe that can be added to a reaction while still maintaining optimal performance. Attempts at a taxa-based probe design required far too many probes for this assay to both perform optimally and be cost-effective (data not shown).

Using the newly designed custom depletion probe set, we demonstrated robust rRNA depletion of human-associated microbiota across body sites and developmental stages, including adult and infant gut samples (Figs. 3, 5, S7-S8) as well as human oral and vaginal samples (Fig. 4). The human microbiome probe set was also effective at depleting rRNA from a skin mock community (Figure S5). Of all the human microbiome sites included in our study, vaginal samples and mock communities were the only ones that did not always require addition of the custom human microbiome probe set. *Lactobacillus*-dominated samples were well depleted by the standard Ribo-Zero Plus rRNA depletion probes while more complex samples with high levels of *Gardnerella*, *Corynebacterium*, or *Bifidobacterium* typically associated with vaginosis [32, 33] required additional probes for

(See figure on next page.)

Fig. 6 Addition of ERCC spike in controls prior to or after rRNA depletion to measure the accuracy of rRNA depletion and potential library prep bias caused by the DP1 and HMV2 probes. **A** ERCC controls were spiked into the library prep either immediately before or after rRNA depletion. **B** Spearman correlation coefficients between different groups tested. **C-F** Representative pairwise correlations between ERCC spike-in abundance for sample pairs with the following characteristics: **C**) pre-depletion, different donors, **D**) post-depletion, different donors, **E**) pre- vs. post-depletion, same donor, **F**) pre- vs. post-depletion, different donors

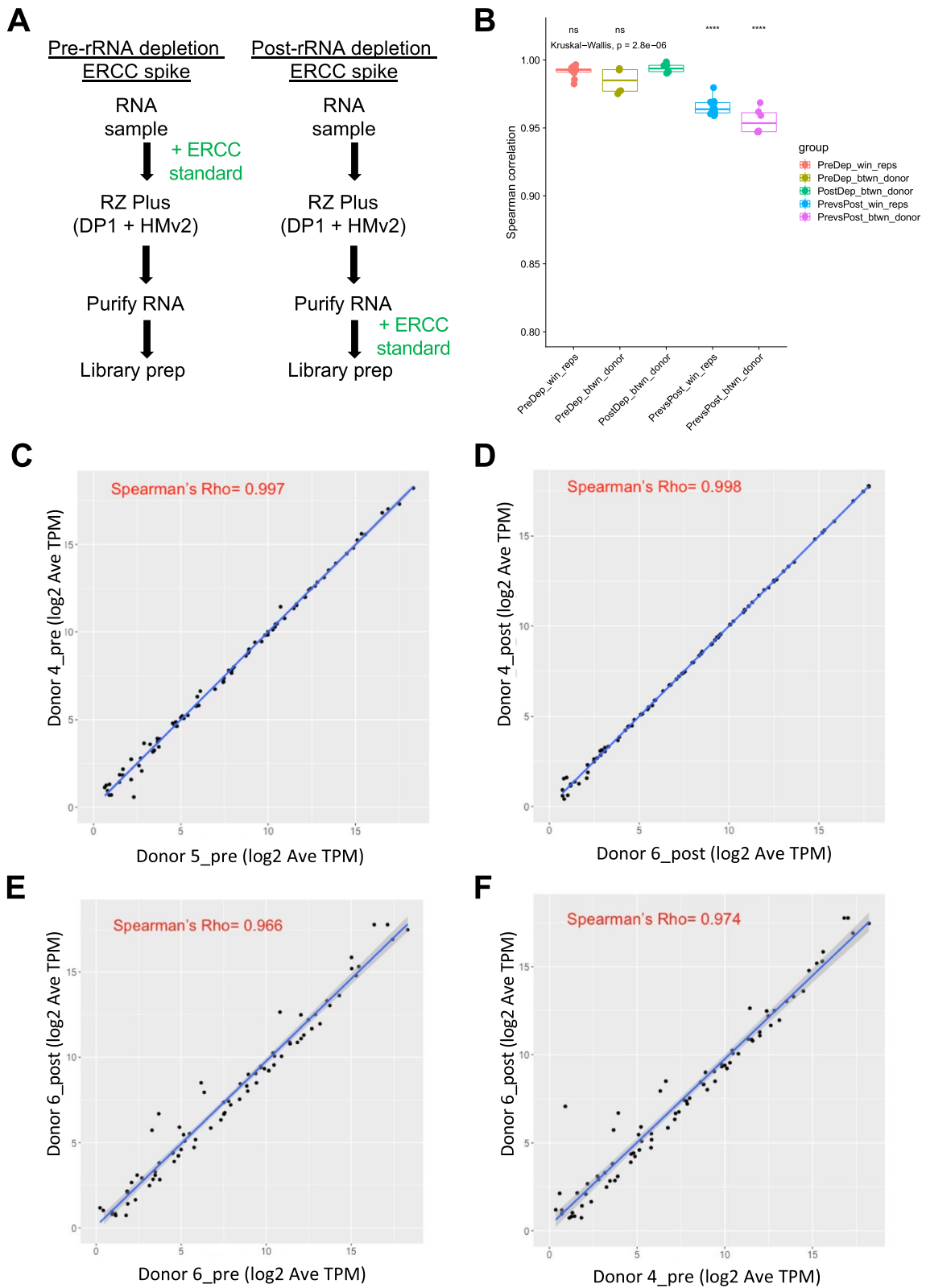


Fig. 6 (See legend on previous page.)

optimal depletion (Fig. 4). This suggests that in general increased bacterial diversity correlates with increased presence of unique rRNA sequences in the sample, which in turn require additional coverage for successful depletion using RNase H-based depletion strategies. However, there were two notable exceptions to this rule: V1, which was dominated by *Corynebacterium*, *Streptococcus*, and *Prevotella* depleted well with the standard probes, while the ATCC-2005 skin mock required HMv1 for optimal results. Surprisingly, both samples had high levels of *Corynebacterium*, albeit not the same species or strains. Nevertheless, we also showed that the probe set can be further refined if needed to add sequence coverage for additional targets in cases where depletion is sub-optimal. For certain infant samples, addition of probes targeting *Bifidobacteria* improved rRNA depletion (Figs. 5, S7). Interestingly, the same was true for adult samples that lacked *Bifidobacteria* (Figure S8). Both HMv1 and HMv2 were designed to target abundant stool rRNA sequences, yet unexpectedly these probe sets show efficient depletion of rRNA in microbiome samples with completely different community structure (oral and vaginal). This demonstrates that pan microbiome rRNA depletion can be achieved through broad and strategic coverage of slightly divergent rRNA sequences. One limitation of this study is that it is currently focused on Western microbiome samples and there is a chance that non-Western samples will have distinct microbes that have rRNA that is not well depleted by the current probe set. The ability to quickly refine, iterate, and supplement the base rRNA depletion probe set may be important for microbiome studies focusing on non-western or non-industrialized populations where microbiome structure and diversity may vary significantly [34]. A similar iterative probe design approach can be used to extend this depletion method to other sample types (microbiomes of other species, soil, or other environmental samples).

This enzymatic-based rRNA depletion method relies on hundreds of DNA probes targeting the diversity of rRNA sequences found in complex microbial community, an approach that has the potential to introduce bias and off-target effects. Therefore, we used synthetic spike-ins to assess the impact of the rRNA depletion process on the quantitative accuracy and reproducibility of gene expression measurements (Fig. 6). All ERCC transcripts were detected and their relative abundances were highly reproducible across all of the conditions and samples tested. Comparison of samples spiked pre- and post-depletion indicated that there was a small, but detectable effect of rRNA depletion on ERCC transcript abundance. However, the Spearman correlations of ERCC transcript abundance between pre- and post-depleted samples were >0.95 among replicates from the same sample as

well as across samples, thus the quantitative bias introduced by the rRNA depletion process is minimal and depletion is not expected to introduce significant error in the resulting microbial gene expression measurements.

To date, the microbiome field has only limited insights on gene expression patterns in the gut microbiome [6, 35], and relatively few comparative metatranscriptomic analyses have been performed. Abu-Ali et al. looked at the transcriptomics profiles of 372 healthy adult men and identified glycolysis, carbohydrate metabolism pathways as well as the pentose phosphate pathway as part of the core gut transcriptome [6]. Analysis of the transcriptomic profiles of stool samples depleted with the pan-human microbiome probe set confirmed the central role of these pathways in gut metabolic activity. In addition, by comparing the metatranscriptomes of infants and adults we were able to reveal additional insights on the dynamic regulation of the developing microbiome. In younger infants (<6 months of age), we saw evidence of glycolysis, but also higher expression of genes involved in amino sugar metabolism (Supplemental Data File 2). The gut microbiome of these infants was largely dominated by *Bifidobacteria*, a genus well known for its ability to process human milk oligosaccharides that contains high levels of N-acetylglucosamine [36]. Moreover, infant stool samples displayed higher levels of many genes involved in nucleotide metabolism, protein synthesis and cell division, indicative of an evolving environment. In contrast, stool samples from >20 month old children showed strikingly different transcriptomic profiles and higher expression of genes involved in the catabolism of amino acids such as glycine and lysine as well as butanoate fermentation (Supplemental Data File 2). This shift paralleled a significant increase in the diversity of >20 month old children's gut microbial profiles and acquisition of new taxa with additional metabolic capabilities (Fig. 5B) that likely correlate with the introduction of solid food to the diet [37]. Amino acids are poorly absorbed in the distal colon, and hence they become an abundant and significant energy source for the gut microbiota [38, 39]. Young children (>20 months) also showed higher levels of genes associated with sporulation, likely driven by the establishment of Firmicutes such as *Clostridiales*, *Lachnospiraceae* and *Ruminococcus* [40] as well as motility (pili, flagella) which are found in many bacterial species and play multiple functions including promoting adhesion to the intestinal lumen (Supplemental Data File 2). Differential taxonomic expression also revealed that beyond *Bifidobacteria* younger infants had higher levels of *Escherichia*, *Shigella*, *Veillonella*, as well as >20 species of the *Enterococcus* genera. Interestingly, *Veillonella* and *Enterococcus* have been described as core components of the human milk microbiome [41, 42] while

Escherichia/Shigella have been shown to be present at higher relative abundance in vaginally-born infants [43, 44]. Although we don't have metadata beyond age for these infants, our findings highlight potential features of their early life environment.

Comparison of the transcriptomic profiles of infant and adult stool samples also identified numerous enzymes of the core gut transcriptome [6]. However, we noticed a striking increase in the relative expression of enzymes of the glycolysis and pentose phosphate pathways in infant samples compared to adults, suggesting a critical role for these two highly intertwined pathways in the infant gut [45]. This is in agreement with a previous study that showed the prevalence of glycolysis during the first year of life [45]. In contrast, adult samples showed increased expression of several glycosylases and sugar transporters, presumably to deal with the greater variety of carbohydrates that are part of the adult diet. Enzymes of the benzoyl-CoA pathway also showed increased expression in adult samples, which suggests that aromatic compounds are processed as carbon sources in the adult gut [46]. Interestingly, sodium benzoate is also one of the most commonly used food preservatives [47]. Additionally, adult samples expressed a number of genes involved in adhesion, biofilm formation motility, competence and quorum sensing, indicative of cells adapting and competing or cooperating to survive in an established community [48, 49].

One of the most intriguing differences we observed in metatranscriptomes was centered around the stress response. Virtually all samples showed activation of stress response pathways (heat shock, cold shock, and oxidative stress), but these pathways differed significantly across age groups. Younger infants showed increased expression of many universal stress response proteins, well known for their ability to respond to a number of environmental stressors [50]. When compared to adults, infants unexpectedly showed upregulation of several cold shock proteins. We believe this difference may stem from the study design itself, as all infant stool samples were frozen and stored at -80°C prior to RNA extraction while adult samples were either extracted immediately or stabilized in DNA Genotek's OMNIgene●GUT DNA/RNA devices. Our data highlights the importance of experimental design in metatranscriptomic analyses, since unlike DNA, gene expression profiles change quickly in response to environmental factors [51]. In summary, the enzymatic rRNA depletion method reported here will enable robust and accurate metatranscriptomic studies of human-associated microbial communities, allowing for detailed studies of microbiome functional activity to complement DNA-based assessments of microbial community composition.

Methods

Samples

All samples were collected under DNA Genotek's IRB protocol (RD-PR-00087). Human stool samples were either collected in specimen cups or in prototypes of DNA Genotek's OMNIgene●GUT DNA & RNA (OMR-205). Unstabilized samples were returned to the lab on ice packs and either extracted within 2–3 h or stored at -80°C until extraction. OMR-205 stabilized samples were stored at room temperature and RNA was extracted within 10–14 days of sample collection. Vaginal and oral microbiome samples were collected in DNA Genotek's OMNIgene●VAGINAL (OMR-130, vaginal microbiome sampling kit) and OMNIgene●ORAL (OMR-120, tongue microbiome sampling kit) following the device IFU. Samples were stored at room temperature until further processing. Prior to extraction, collected samples were treated with Proteinase K for 1 h at 50°C as per DNA Genotek's instructions and concentrated down to 200–250 μl using Eppendorf's Vacufuge Plus (centrifugation was performed at 30°C for 45–60 min) for optimal extraction yields.

The pooled RNA samples were generated by mixing total RNA extracted from human cells and bacterial cultures. *Francisella philomiragia* (ATCC 25017), *Escherichia coli* (ATCC 13706), *Pseudomonas aeruginosa* (ATCC 10145), *Staphylococcus aureus* (ATCC 25923), *Moraxella catarrhalis* (ATCC 25238), *Klebsiella pneumonia* (ATCC 13883), *Micrococcus luteus* (ATCC 4698), *Yersinia enterocolitica* (ATCC 9610), *Bacillus subtilis* and *Lactobacillus crispatus* (ATCC 33820) were grown overnight in their recommended culture media. Cells were pelleted, washed once with ddH₂O and frozen until extraction. RNA pool 1 consists of 30% human THP-1 RNA (ATCC[®] TIB-202[™]) and 10% total RNA from each the following bacterial species: *B. subtilis*, *E. coli*, *F. philomiragia*, *L. crispatus*, *P. aeruginosa*, *S. aureus* and *Y. enterocolitica*. RNA pool 2 consists of equal amounts (9.09%, by ng of RNA) of total THP-1 human RNA, *F. philomiragia*, *E. coli*, *P. aeruginosa*, *S. aureus*, *M. catarrhalis*, *K. pneumonia*, *M. luteus*, *Y. enterocolitica*, *B. subtilis* and *L. crispatus* total RNA. Gut and skin intact cell mock communities were purchased from ATCC (cat# MSA-2006, and MSA-2005).

RNA extractions, RNA QC and RT-qPCR

Human total RNA was extracted from THP-1 human cells using TRIzol (ThermoFisher, cat# 15,596,026) as per manufacturer instructions. Bacterial RNA was extracted from intact cell mock communities (ATCC, MSA-2006, and MSA-2005), bacterial cultures as well as stool, vaginal and oral microbiome samples

using Qiagen's RNeasy® PowerMicrobiome® Kit (cat# 26,000–50), according to the manufacturer's instructions. For raw stool samples, 50–100 mg was used as input, while for oral and vaginal samples 200–250 µl of pre-concentrated OMNIgene® sample was used as input. Bead beating was performed in the presence of phenol–chloroform–isoamyl alcohol (Sigma-Aldrich, cat# 77,617) and 2-mercaptoethanol (Sigma-Aldrich, cat# M6250). DNase treatment was performed on-column and total RNA samples were eluted in 100 µl nuclease-free water, then quantified using the Quant-iT™ RiboGreen™ RNA Assay Kit (ThermoFisher, cat# R11490). RNA quality and integrity of samples was checked on the Agilent 4200 TapeStation System using RNA ScreenTape or High Sensitivity RNA ScreenTape (Agilent, cat# 5067–5576 and 5067–5579). RNA integrity numbers (RINs) were highly variable across body sites/sample types and ranged from 2.0 to 9.0. Representative traces for each sample type are shown in Figure S1.

An RT-qPCR approach was used to screen the relative abundance of *Lactobacillus* in vaginal samples and control samples (pure *Lactobacillus crispatus* RNA and RNA pool 2). 50–80 ng of control or RNA extracted from OMNIgene●VAGINAL collected samples was reverse transcribed using Superscript II (ThermoFisher, cat# 18,064,022) or Superscript III (ThermoFisher, cat# 28,025,013) and random hexamers (ThermoFisher, cat# N8080127). Total and *Lactobacillus* 16S rRNA levels were quantified by qPCR using universal bacterial 16S primers (Fwd 5'-ATTACCGCGGCTGCTGG-3'; Rev 5'-CCTACG GGAGGCAGCAG-3') and *Lactobacillus*-genus primers (Fwd 5'-ATGGAAGAACCAGTGGCG-3'; Rev 5'-CAGCACTGAGAGGCGGAAAC-3'). 5 µl of diluted cDNA was then used as a template in qPCR reactions containing 1 µM Syto 9, 1.5 mM MgCl₂ and 0.1 µg/µl BSA. Real-time PCR amplification was performed on the Corbett Rotor-Gene 6000 (discontinued) using the following conditions: 95 °C for 2 min followed by 35 cycles consisting of 95 °C for 30 s, 50 °C or 55 °C for 20 s and 72 °C for 20 s for the *Lactobacillus*-specific primers and universal 16S primers respectively. PCR amplification was followed by an incubation at 72 °C for 1 min 30 s and melting (72 °C to 95 °C in 1 °C increments). Relative abundance of *Lactobacillus* was determined by calculating ΔCt between *Lactobacillus* 16S and total 16S.

Human microbiome probe pool design

Total RNA from gut microbiome samples of 9 donors [52] was processed in triplicate with the Ribo-Zero Plus rRNA Depletion Kit (using DP1 probes), converted into RNAseq libraries using the TruSeq Stranded Total RNAseq kit and sequenced on a NextSeq (PE 76), producing between 11 to 36 million reads per sample.

The FASTQ [53] files from each donor were then aligned to the SILVA (v119) [26] database using SortMeRNA [27] to identify the regions of rRNA to target for depletion. Any sequence regions that align in close proximity (1–3 nt) were merged and sorted by coverage depth and then filtered to remove any with less than 500× coverage. These regions were typically less than 200 nt in length representing rRNA segments that are not depleted by DP1 probes (Figure S3A). The top 50 most abundant regions were collected from each sample (donor) and combined to create a list of abundant regions. Any regions that overlapped were then merged and the list converted into a FASTA file. To identify and remove redundancies, a pairwise alignment of each region was performed and any regions that demonstrate equal to or greater than 80% identity were flagged and only one region was chosen for probe design (Figure S3B). The Ribo-Zero Plus probes (DP1) were then aligned to the selected, non-redundant regions and any regions where the probes were aligned at equal to or greater than 80% identity were eliminated. The remaining regions were collected, probe locations were established and antisense probe sequences were created. Each probe is 50 nt in length composed only of standard DNA bases; no modified bases or 5'/3' end modifications are included. In addition, both HMv1 and HmV2 also include probes that were designed directly against the known rRNA sequences from all 38 species present in the ATCC mock community samples (MSA-2002, MSA-2005 & MSA-2006) as well as *E. coli* and *B. subtilis*. These probes were designed to specifically target both the large and small subunits of each species present in the MSA samples and were not designed using the method described in Fig. 2E. The commercial product that contains HMv2 is called the Illumina® Stranded Total RNA Prep with Ligation, Ribo-Zero Plus Microbiome and the probe pool equivalent to HMv2 is now referred to as DPM.

RNAseq library preparation and sequencing

80 to 500 ng total RNA was used as input for rRNA depletion using either Illumina's Ribo-Zero Gold rRNA Removal Epidemiology kit (discontinued) or Illumina's Ribo-Zero Plus rRNA depletion kit (Illumina, cat# 20,037,135). For Ribo-Zero Plus depletion reactions, total RNA was mixed with 1 µl of DP1 (standard probe set) in the presence or absence of 1 to 1.5 µl of human microbiome probe pool (HMv1 or HMv2). Probes were hybridized and rRNA depleted as per manufacturer's instructions. For undepleted samples, 10–20 ng total RNA was used as template for library preparation. RNAseq libraries were prepared using Illumina's TruSeq Stranded mRNA Library Prep Kit (cat# 20,020,595) or Illumina's Stranded Total RNA Prep

Ligation Kit (cat# 20,040,529). Final libraries were quantified with the Quant-iT™ PicoGreen™ dsDNA Assay (ThermoFisher, cat# P7589) or the Qubit dsDNA HS Assay (ThermoFisher, cat# Q32851). Final library size was assessed by running libraries on the Agilent 4200 TapeStation System using D1000 ScreenTapes (Agilent, cat# 5067–5582) or alternatively, on Bioanalyzer High Sensitivity Chip (cat# 5067–4626). Individual libraries were then normalized, pooled, and sequenced on Illumina platforms (MiSeq, NextSeq or NovaSeq, see Supplemental Data File 1).

ERCC spike-in experiment

To assess the performance and consistency of rRNA depletion/library preparation in complex microbiome samples, we designed an experiment where the ERCC RNA Spike-In Mix (ThermoFisher, cat# 4,456,740) was added to RNA samples (total stool RNA from 3 donors and RNA pool 2) either before or after rRNA depletion with Ribo-Zero Plus. The ERCC Mix contains 92 artificial transcripts of varying sizes that can be used to assess sensitivity and potential bias introduced in NGS workflows. Briefly, 1 µl of a 1/200 dilution of the ERCC RNA Spike-In Mix was mixed with 250 ng of total RNA and used as input for Ribo-Zero Plus depletion (containing the human microbiome probe pool HMv2) or added to matching rRNA-depleted samples prior to the fragmentation step of library preparation. In order to assess reproducibility, triplicates were processed for each sample and condition using Illumina's Stranded Total RNA Prep Ligation Kit (cat# 20,040,529).

Data analysis

Detailed methods for data analysis are described in the subsections below. Across all the analysis processes, the statistical analyses and plotting were carried out using R [54] (version 4.0.5) and ggplot2 [55] or Graph-Pad Prism 9.

Testing methods for filtering rRNA content

Validation of rRNA removal methods made use of a data set from BioProject PRJNA295252 [28]. Reads matching human genome were removed from the FASTQ files related to BioProject PRJNA295252 using bowtie2 [24] and hg37 [56]. To test the efficacy of various rRNA removal tools and databases, we tested bowtie2 [24] (version 2.4.2), SortMeRNA [27] (version 4.2.0) and three tools from the BBTools [25] (version 38.90) package (bbduk, bbsplit and seal), combined with four databases: (ar: SILVA SSU/LSU NR99 [26]; art: ar+tRNA RFAM clans; ds: SortMeRNA [27] (v4.3) default database; ss:

SortMeRNA [27] (v4.3) sensitive database). In order to conserve time and memory requirements, the data were subsampled for bowtie2 (2.5% or 25%), bbsplit, SortMeRNA [27] (2.5%), seal (25%) and bbduk (25% or 100%). For cmscan [57] (Infernal v1.1.4) and UProC [58] (v1.2.0), reads were merged using bbmerge-auto.sh from BBTools [25] (version 38.90), converted to FASTA using Seqtk (<https://github.com/lh3/seqtk>) and dereplicated using vsearch, then 1000 sequences were randomly subsampled using Seqtk and analyzed using UProC [58] by searching against the KEGG Orthologs database [59] (kegg_20140317.uprocdg.gz) and cmscan [57] by searching against the RFAM14.4 database [60].

Host and rRNA removal

Adapter sequences were removed from the paired-end reads using Trimmomatic-0.38 [61] with a sliding window of '4:20' and minimum length cut-off of 50 bases. Host and rRNA sequences were filtered using BBduk [25] (BBTools version 38.90) using GRCh38 UCSC human genome and 5S Ribosomal RNA reference and the following categories subset from the art database: BacSSU, BacLSU, ArcSSU, ArcLSU, EukSSU, EukLSU and RFAM, for easier identification of various non-coding RNA targets. Reads that did not map to any of these databases were then processed for downstream analyses.

Annotation of RNASeq reads

The filtered (non-rRNA and non-host) reads were annotated for taxonomy either using Kaiju [62] (version 1.7.0) by searching against the included proGenomes database or using SAMSA2 [63] (version 2) diamond search against the included RefSeq database. Functional annotation was carried out using SAMSA2 [62] (version 2) diamond search against the RefSeq functional genes and the SEED (2017) diamond database.

Assessing the accuracy and reproducibility of rRNA depletion using spike-in ERCC Samples

Gene counts were used to assess the accuracy and reproducibility of RiboZeroPlus method using ERCC spike-ins. Once rRNA reads were removed using bbduk as described in the "Host and rRNA removal" section above, the remaining reads were aligned to ERCC gene sequences using TopHat2 [64] to calculate counts per gene. Gene counts were imported to R, converted to transcript per million (TPM), then the Spearman correlations were calculated between various conditions to demonstrate the accuracy and reproducibility of rRNA depletion with RiboZeroPlus (Figure S14).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12866-023-03037-y>.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Acknowledgements

We thank Andrew Cross and Shane Sontag for help with DNA sequencing and Evgueni Doukhanine for helpful comments on the manuscript. We kindly thank Dr. George Weinstock for RNA samples used to aid in the design of DPM.

Authors' contributions

A.T., S.K., B.L.F., and D.M.G. conceived and designed experiments. A.T. designed rRNA depletion probes. A.M., J.K., D.K., F.H., V.R., E.L., and J.J. conducted experiments. A.T., S.M., J.M. analyzed data. S.K., B.L.F., and D.M.G. wrote the manuscript. All authors contributed to review and revision of the manuscript.

Funding

This work was funded by Illumina, DNA Genotek, and Diversigen. No grant funding was obtained for this study.

Availability of data and materials

Sequencing data for this project is available through the National Center for Biotechnology Information (NCBI) Sequence Read Archive BioProject PRJNA812896.

Declarations

Ethics approval and consent to participate

All samples were collected using experimental protocols approved under DNA Genotek's IRB protocol (RD-PR-00087). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects.

Consent for publication

Not applicable.

Competing interests

A.T., J.K., D.K., F.H., V.R., G.P.S., and S.K. are employees of Illumina, Inc. A.M., J.M., and B.L.F. are employees of DNA Genotek. S.M., E.L., J.J., and D.M.G. are current or former employees of Diversigen, Inc. A.M. and B.L.F. are inventors on a provisional patent submitted for the DNA/RNA stabilization chemistry in OMR-205 (United States Patent Application No. 63/208,212). A.T., J.K., D.K. and S.K. are inventors on a patent application pending WO 2021/127191.

Received: 13 April 2023 Accepted: 3 October 2023

Published online: 20 October 2023

References

- Choi I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13:260–70.
- Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nat Med*. 2018;24:392 NIH Public Access.
- Durack J, Lynch SV. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med*. 2019;216:20–40 The Rockefeller University Press.
- Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. *Genome Med*. 2016;8:51 BioMed Central.
- Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci USA*. 2014;111(22):E2329–38.
- Abu-Ali GS, Mehta RS, Lloyd-Price J, Mallick H, Branck T, Ivey KL, et al. Metatranscriptome of human fecal microbial communities in a cohort of adult men. *Nat Microbiol*. 2018;3:356 NIH Public Access.
- Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenkov T, Niaz F, et al. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA*. 2010;107:7503–8.
- Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. 2019;569:655 Nature Publishing Group.
- Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science*. 2013;341:295–8.
- Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol*. 2012;13:R23.
- Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2017;8(1). <https://doi.org/10.1002/wrna.1364>.
- Zhao S, Zhang Y, Gamini R, Zhang B, Von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep*. 2018;8:4781.
- August JT, Ortiz PJ, Hurwitz J. Ribonucleic Acid-dependent Ribonucleotide Incorporation I. Purification and properties of the enzyme. *J Biol Chem*. 1962;237:3786–93.
- Modak A, Srinivasan PR. Purification and Properties of a Ribonucleic Acid Primer-independent Polyriboadenylate Polymerase from *Escherichia coli*. *J Biol Chem*. 1973;248:69–6910.
- Mohanty BK, Kushner SR. Analysis of the function of *Escherichia coli* poly(A) polymerase I in RNA metabolism. *Mol Microbiol*. 1999;34:1094–108.
- O'Hara EB, Chekanova JA, Ingle CA, Kushner ZR, Peters E, Kushner SR. Polyadenylation helps regulate mRNA decay in *Escherichia coli*. *Proc Natl Acad Sci USA*. 1995;92:1807–11.
- Culviner PH, Guegler CK, Laub MT. A Simple, Cost-Effective, and Robust Method for rRNA Depletion in RNA-Sequencing Studies. *MBio*. 2020;11(2):e00010–20.
- Huang Y, Sheth RU, Kaufman A, Wang HH. Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. *Nucleic Acids Res*. 2020;48:E20.
- Prezza G, Heckel T, Dietrich S, Homberger C, Westermann AJ, Vogel J. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA*. 2020;26:1069–78.
- Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol BioMed Central Ltd*. 2016;17:1–13.
- Reck M, Tomasch J, Deng Z, Jarek M, Husemann P, Wagner-Döbler I. Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation. *BMC Genomics*. 2015;16:494.
- Wahl A, Huptas C, Neuhaus K. Comparison of rRNA depletion methods for efficient bacterial mRNA sequencing. *Sci Rep*. 2022;12:5765.
- Ojala T, Häkkinen A-E, Kankuri E, Kankainen M. Current concepts, advances, and challenges in deciphering the human microbiota with metatranscriptomics. *Trends Genet*. 2023;39:686–702.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* Nature Publishing Group. 2012;9:357–9.
- Bushnell B. BBMap short read aligner and other bioinformatic tools. Berkeley, CA: Joint Genome Institute; 2019.
- Quast C, Priesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41:D590–6.

27. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics Oxford Academic*. 2012;28:3211–7.
28. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, et al. MetaTrans: an open-source pipeline for metatranscriptomics. *Sci Rep*. 2016;6:1–12.
29. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*. 2012;486:207–14.
30. Stewart CJ, Ajami NJ, O'Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*. 2018;562:583–8.
31. Turrioni F, Peano C, Pass DA, Foroni E, Severgnini M, Claesson MJ, et al. Diversity of bifidobacteria within the infant gut microbiota. *PLoS One*. 2012;7:e36957.
32. Chen X, Lu Y, Chen T, Li R. The Female Vaginal Microbiome in Health and Bacterial Vaginosis. *Front Cell Infect Microbiol*. 2021;11:631972.
33. Ma B, Forney LJ, Ravel J. The vaginal microbiome: rethinking health and diseases. *Annu Rev Microbiol*. 2012;66:371.
34. Vangay P, Johnson AJ, Ward TL, Al-Ghalith GA, Shields-Cutler RR, Hillmann BM, et al. US Immigration Westernizes the Human Gut Microbiome. *Cell*. 2018;175:962–972.e10.
35. Booijink CCGM, Boekhorst J, Zoetendal EG, Smidt H, Kleerebezem M, De Vos WM. Metatranscriptome analysis of the human fecal microbiota reveals subject-specific expression profiles, with genes encoding proteins involved in carbohydrate metabolism being dominantly expressed. *Appl Environ Microbiol Appl Environ Microbiol*. 2010;76:5533–40.
36. Wiciński M, Sawicka E, Gebałski J, Kubiak K, Malinowski B. Human Milk Oligosaccharides: Health Benefits, Potential Applications in Infant Formulas, and Pharmacology. *Nutrients*. 2020;12:266.
37. Hugenholtz F, Ritari J, Nylund L, Davids M, Satokari R, De Vos WM. Feasibility of Metatranscriptome Analysis from Infant Gut Microbiota: Adaptation to Solid Foods Results in Increased Activity of Firmicutes at Six Months. *Int J Microbiol*. 2017;2017:9547063.
38. Davila AM, Blachier F, Gotteland M, Andriamihaja M, Benetti PH, Sanz Y, et al. Intestinal luminal nitrogen metabolism: Role of the gut microbiota and consequences for the host. *Pharmacol Res*. 2013;68:95–107.
39. Neis EPJG, Dejong CHC, Rensen SS. The role of microbial amino acid metabolism in host metabolism. *Nutrients*. 2015;7:2930–46.
40. Egan M, Dempsey E, Ryan CA, Ross RP, Stanton C. The Sporobiota of the Human Gut. *Gut Microbes Gut Microbes*. 2021;13:1–17.
41. Boudry G, Charton E, Le Huerou-Luron I, Ferret-Bernard S, Le Gall S, Even S, et al. The Relationship Between Breast Milk Components and the Infant Gut Microbiota. *Front Nutr*. 2021;8:629740.
42. Skillington O, Mills S, Gupta A, Mayer EA, Gill CIR, Del Rio D, et al. The contrasting human gut microbiota in early and late life and implications for host health and disease. *Nutr Heal Aging IOS Press*. 2021;6:157–78.
43. Azad MB, Konya T, Maughan H, Guttman DS, Field CJ, Chari RS, et al. Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. *CMAJ CMAJ*. 2013;185:385–94.
44. Wang Z, Neupane A, Vo R, White J, Wang X, Marzano SYL. Comparing Gut Microbiome in Mothers' Own Breast Milk- and Formula-Fed Moderate-Late Preterm Infants. *Front Microbiol*. 2020;11:891.
45. Stincone A, Prigione A, Cramer T, Wamelink MMC, Campbell K, Cheung E, et al. The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biol Rev Camb Philos Soc*. 2015;90:927–63.
46. Harwood CS, Burchhardt G, Herrmann H, Fuchs G. Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. *FEMS Microbiol Rev Oxford Academic*. 1998;22:439–58.
47. Yadav M, Lomash A, Kapoor S, Pandey R, Chauhan NS. Mapping of the benzoate metabolism by human gut microbiome indicates food-derived metagenome evolution. *Sci Rep*. 2021;11:1–11.
48. Deng Z, Luo XM, Liu J, Wang H. Quorum Sensing, Biofilm, and Intestinal Mucosal Barrier: Involvement the Role of Probiotic. *Front Cell Infect Microbiol*. 2020;10:538077.
49. Buret AG, Motta JP, Allain T, Ferraz J, Wallace JL. Pathobiont release from dysbiotic gut microbiota biofilms in intestinal inflammatory diseases: a role for iron? *J Biomed Sci*. 2019;26:1.
50. Chi YH, Koo SS, Oh HT, Lee ES, Park JH, Phan KAT, et al. The physiological functions of universal stress proteins and their molecular mechanism to protect plants from environmental stresses. *Front Plant Sci*. 2019;10:750.
51. Vargas-Blanco DA, Shell SS. Regulation of mRNA Stability During Bacterial Stress Responses. *Front Microbiol*. 2020;11:2111.
52. Petersen LM, Bautista EJ, Nguyen H, Hanson BM, Chen L, Lek SH, et al. Community characteristics of the gut microbiomes of competitive cyclists. *Microbiome*. 2017;5:98.
53. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. 2010;38:1767 Oxford University Press.
54. R Core Team. R: A language and environment for statistical computing. 3.5.0. Vienna, Austria: R Foundation for Statistical Computing; 2018.
55. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag; 2016.
56. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing Reference Genome Assemblies. *PLoS Biol*. 2011;9:e1001091.
57. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res Oxford Academic*. 2019;47:W636–41.
58. Meinicke P. UProC: tools for ultra-fast protein domain classification. *Bioinformatics Bioinformatics*. 2015;31:1382–8.
59. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28:27 Oxford University Press.
60. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res Oxford Academic*. 2021;49:D192–200.
61. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics Oxford University Press*. 2014;30:2114–20.
62. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun*. 2016;7:1–9.
63. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics*. 2018;19:175.
64. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14:R36.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

