Optimizing mouse metatranscriptome profiling by selective removal of redundant nucleic acid sequences

Morgan Roos[1*], Samuel Bunga[1*], Asako Tan[1], Erica Maissy[2], Dylan Skola[1], Alexander Richter[2], Daniel S. Whittaker[3], Paula Desplats[4], Amir Zarrinpar[2,5,6,7], Rick Conrad[9] & Scott Kuersten[1]

[1]Illumina, Inc., San Diego, CA, USA
[2]Division of Gastroenterology, University of California, San Diego, La Jolla, CA, USA
[3]Department of Neurosciences, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093.
[4]Departments of Neurosciences and Pathology, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093.
[5]Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, La Jolla, CA, USA
[6]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA
[7]Division of Gastroenterology, Jennifer Moreno Department of Veterans Affairs Medical Center, La Jolla, CA, USA
[8]Retired
*These authors contributed equally to this work.

Abstract

Metatranscriptome (MetaT) sequencing is a critical tool for profiling the dynamic metabolic functions of microbiomes. In addition to taxonomic information, MetaT also provides real-time gene expression data of both host and microbial populations, thus permitting authentic quantification of the functional (enzymatic) output of the microbiome and its host. The main challenge to effective and accurate MetaT analysis is the removal of highly abundant rRNA transcripts from these complex mixtures of microbes, which can number in the thousands of individual species. Regardless of methodology for rRNA depletion, the design of rRNA removal probes based solely upon taxonomic content of the microbiome typically requires very large numbers of individual probes, making this approach complex to commercially manufacture, costly, and frequently technically infeasible. In previous work [1], we designed a set of depletion probes for human stool samples using a design strategy based solely on sequence abundance, completely agnostic of the microbiomal species present. Here, we show that the human-based probes are less effective when used with mouse cecal samples. However, adapting additional rRNA depletion probes specifically to cecal content provides both greater efficiency and consistency for MetaT analysis of mouse samples.

Importance

Sequencing total RNA from microbiome samples is seriously impaired by the overwhelming proportion of rRNA to mRNA content. As much as 99% of sequencing reads can be assigned to the rRNA content, thus removal of these abundant transcripts is critical to MetaT analysis. The use of Ribo Zero Plus rRNA depletion probes designed for human gut microbiomes proved to be less effective and more inconsistent across mouse cecal donor samples, a common experimental system for microbiome studies. In the present work, we have extended and refined a taxonomically-neutral probe design method for mouse cecal content. The additional probes were carefully chosen to limit the number needed for effective depletion to reduce both the cost and risk of introducing bias to MetaT analysis. Our results demonstrate this method as efficient and consistent for rRNA removal in mouse cecal samples, thus providing a significant increase in the number of mRNA-rich sequencing reads for MetaT analysis.

Introduction

The diversity and metabolic state of the estimated ~1000 species contributing to the intestinal microbiota is immensely important to the health and well-being of the host [2]. Diet, overall health, and even disease state of the host can, conversely, affect the activity and enzymatic expression patterns of the intestinal microbiome. Our understanding of the microbiome has evolved over the past decades such that it is now understood that the interplay between the microbial population and its host is a dynamic and vital interaction [3]. To further understand this interaction, it is crucial that we look at not just the composition of the microbes present, but also the metabolic contribution from this population to assess the health and well-being of the host.

In the last decades, the use of Next Generation Sequencing (NGS) to interrogate microbiomes, especially from the gut contents of host organisms, has become more cost-effective and technologically attainable [4, 5]. Genetic profiling of the microbial populations present in gut microbiomes is traditionally performed by methods such as 16S rDNA sequencing [6-8]. More recently, as NGS has become far less expensive, whole genome bacterial shotgun sequencing has become more commonplace and provides a greater depth of knowledge of the bacterial diversity and inferred function(s) in these complex communities [9-11]. However, metagenome (MetaG) sequencing is mostly limited to taxonomic identification of genus and/or species composition of the samples. The functional metabolic state of the microbes can, at best, be deduced based on assumptions of the relative contributions from the taxonomic groups present. Elucidating the profile of mRNAs being expressed by these populations (their transcriptomes) is more informative.

Using transcriptome profiling of microbiomes (i.e. metatranscriptomics) can provide a wealth of valuable information about not just the taxonomic composition, but also the metabolic activity of the microbial population [12]. The dynamics of gene expression changes between microbiome and host can be simultaneously monitored to establish the health or disease states [13]. Metatranscriptome (MetaT) sequencing offers a much more detailed view of the overall activity of the microbes by providing real time functional gene expression information, such as what gene families and enzymatic activities are taking place at the time of sample collection and extraction [14, 15]. Furthermore, it can be quite informative to track these dynamic changes over time, especially in response to dietary interventions, drug treatments, and disease progression of the host [2, 16]. Whereas tracking changes to the taxonomic composition of the gut by MetaG sequencing may be relatively stable upon intervention, MetaT offers the potential for a more detailed and dynamic profiling of the most transcriptionally active participants in the sample [17].

A major challenge for MetaT analysis is the presence of highly abundant rRNA transcripts. Bacterial small subunit (SSU or 16S) and large subunit (LSU or 23S) rRNA transcripts dominate total RNA samples extracted from gut microbiomes. Indeed, sequencing total RNA from stool samples without rRNA removal can typically result in >95% of the reads matching LSU and SSU rRNA transcripts. Methods to remove the vast diversity of microbiome rRNA can be contrasted with the relatively easy methods available for removing host eukaryotic rRNA from samples. Since polyA tails are added post-transcriptionally to coding mRNAs as part of eukaryotic RNA processing, the mRNAs themselves can be preferentially enriched using oligo-dT capture beads. Additionally, removal of the host rRNA is more feasible since it only requires using sequences from a single species. Several commercially available rRNA removal kits and methods can be utilized to remove rRNA from commonly studied species, such as human, mouse, and rat and provides the means to sequence both mRNA and non-coding transcripts in these sample types. Routinely used methods for removing rRNA from total RNA include enzymatic (i.e. RNase H) depletion, CRISPR-based approaches, or physical removal using hybridization with antisense biotinylated probes and streptavidin magnetic beads [18-21]. However, these methods designed for samples from a single eukaryotic host are not effective for complex microbiome sample types since the microbial rRNA is from a multitude of source organisms whose sequences are evolutionarily divergent [1]. Removal of these abundant transcripts through targeted probe design, whether using physical or enzymatic means, is a much more complex procedure to permit the deep sequencing of microbial mRNA.

In a recent study, a 'rational' probe design strategy was established where raw sequencing data was utilized to collect and filter abundant sequences from human gut microbiome samples [1]. The probe set that was designed demonstrated effective and efficient enzymatic depletion of rRNA from both human adult and infant stool samples with minimal bias introduced. Robust depletion of rRNA within multiple human microbiome sample types, including stool, tongue, and vagina resulted in >60% of sequencing reads available for MetaT analysis. This method is commercially available as the Ribo-Zero Plus Microbiome (RZPM) kit.

In this work, we sought to first test the ability of RZPM to deplete rRNA from a set of mouse microbiome samples and, if needed, make use of a similar rational probe design strategy to create supplemental probe pools optimized for mouse gut (specifically, cecal) contents. Since there is currently no commercially available solution for rRNA depletion of mouse microbiome sample types, an important goal is to provide a list of additional probes that can be obtained at minimal cost and combined with RZPM for more effective performance. Additionally, this effort is intended to provide improvements in both consistency and efficiency of depletion for mouse-specific MetaT analysis. We demonstrate that addition of the supplemental probes can provide ~75% of the mRNA-rich reads available for MetaT analysis. These extra probes add a minimal cost per sample to the existing RZPM kit yet provide an additional ~15% of sequencing reads for functional data analysis.

Materials/Methods:

Animal Work

All animal research was performed in accordance with University of California, San Diego Institutional Animal Care and Use Committee. Cecal contents were manually removed (not washed) from the cecum and saved for RNA extractions. For terminal ileum, the distal ~2cm of the ileum was collected including ileal content. Samples were immediately flash frozen and powderized, and RNA was extracted using the MagMAX mirVana Total RNA isolation kit (Thermo Fisher #A27828).

RNA-Seq library preparation

Extracted RNA (~100 -200 ng) was prepped for sequencing using the Illumina Stranded Total RNA Prep with Ribo-Zero Plus® Microbiome kit following the manufacturer's protocol. The 5050 and 2025 probe sets were purchased from IDT as an oPools product at 50 pmol per oligo. For each tube, the lyophilized pellet obtained from the manufacturer was resuspended in 50 μl of nuclease-free water (for a final concentration of 1 pmol/μl/oligo) and 1 μl was used per depletion reaction. Ribo-Zero Plus depletion was performed with either (1) no probes (volume supplemented with nuclease-free water), (2) Depletion Pool 1 (DP1) and Depletion Pool Microbiome (DPM) (1 μl each), (3) DP1, DPM, and the 5050 probe set (1 μl each), or (4) DP1, DPM, and 2025 probe set (1 μl each). RNA-Seq libraries were sequenced on Illumina's NovaSeq 6000 at 2x150 bp.

RNA-Seq Analysis

Post sequencing, FASTQ's were downsampled to 20 million total reads (10 million clusters) each using BaseSpace Sequencing Hub (BSSH) FASTQ Toolkit app and then processed with BSSH applications. For metatranscriptome analysis, the Microbiome Metatranscriptomics application was used. The BBDuk output was utilized for analysis of rRNA content and for rRNA-based taxonomic analysis. As part of filtering out the rRNA reads, the application bins the rRNA reads into a fastq file for each sample. The rRNA reads were then processed with the BSSH DRAGEN Metagenomics application for species analysis and proportion. For host transcriptome and general differential gene expression analysis, the BSSH RNA Express application was used. For normalization of Gene Family Abundance, HUMAnN quantifies genes and pathways in units of RPKs (reads per kilobase). To make the results comparable across samples, we converted these raw abundances to relative abundances using the human renorm table script with the -u relab parameter (as described in the HUMAnN3 documentation at https://github.com/biobakery/humann). This process follows a Total Sum Scaling (TSS) approach, which normalizes the total abundance in each sample to 1, which helps control for differences in sequencing depth [30].

Probe Design

Following sequencing and RNA-Seq analysis, 7 cecal samples with more than 30% of reads aligning to rRNA were identified. The rRNA reads were collected and processed with variations of a probe design pipeline described in previous work [1]. To select the most efficient and c ost-effective probe counts, the samples were processed with the probe design pipeline multiple times, varying the number of the top most abundant regions from 20 to 50 and altering the spacing between probes from 25 nt to 50 nt. Out of the matrix of results, two options were selected based on the number of probes yielded and level of stringency. These probe sets were then assessed for their performance in library preparation and sequencing.

Results

Efficacy of rRNA depletion of mouse cecal samples

In previous work, we demonstrated the utility of using the enzyme-based RiboZero Plus methodology to deplete rRNA from mixed-source RNA extracted from complex human microbiomes [1]. We developed a strategy to efficiently design removal probes for complex mixtures of microbes based solely on sequence data without referencing assumptions about specific bacterial genera or species being present. We refer to this as ‘rational’ or ‘taxonomically neutral’ design of probes. This probe set was subsequently commercialized in the Ribo Zero Plus Microbiome kit as DPM (Depletion Pool Microbiome). Combining DPM with the human, mouse, and rat probes from the Ribo Zero Kit (DP1) demonstrated effective depletion of both host and microbial rRNA sequences across several human microbiome sample types and the resulting RNAseq libraries did not show significant bias from the additional probes used for depletion [1] .

Since the initial design used microbiome sequences from human stool samples as the training set to provide new content, we investigated how it would perform with microbiome samples obtained from the common mouse, *Mus musculus*, a laboratory model system routinely used for studying the gut microbiome. Numerous studies using methods such as 16S rDNA and whole genome shotgun sequencing of murine gut content have been published [16, 22, 23]. However, mouse MetaT analysis has been somewhat more limited due to the lack of relevant molecular tools to aid in the generation of mRNA-rich RNA for RNA-Seq library preparation [15].

To determine the effectiveness of RZPM depletion from mouse microbiomes, matched samples from ~60 individual mice (56 cecal, 62 terminal ileum, and 62 liver samples) were obtained. The functional analysis of these samples is part of a larger study on aging and Alzheimer’s that will be discussed in detail elsewhere (manuscript in progress). For this work, we initially utilized the commercial RZPM probe sets (DP1 + DPM) to determine its effectiveness and extent of rRNA depletion in these three sample types. Total RNA extracted from each sample was subjected to rRNA depletion using the RZPM kit, converted into RNAseq libraries using the Illumina Stranded Total RNA Kit, and sequenced (150 cycle paired end) on a NovaSeq 6000 instrument (Figure 1A). The data was analyzed using the Microbiome Metatranscriptomics application in Basespace Sequencing Hub (BSSH). Briefly, the raw data was quality-filtered and then analyzed for rRNA content by BBDUK alignment using the Silva database, a comprehensive rRNA database, as the alignment reference [1, 24]. The pipeline reports the proportion of reads that match rRNA sequences or host (mm9) transcriptome (Figure 1A). The number of reads that match rRNA sequences differs depending on sample type. Both ileum and liver samples have very low rRNA reads from microbiome content and instead are dominated by host transcriptome (mm9). This suggests that these sample types are predominantly composed of mouse cells and rRNA depletion was effective at removing the confounding host rRNA reads. In contrast, the cecal samples display very little host contribution and instead contain a much higher proportion of microbial (”retained”, blue) content and a variable amount of contaminating rRNA reads. This suggests that the cecal samples contain very little mouse RNA and are mostly composed of bacterial content. The proportion of rRNA reads differs for each individual sample and ranges from ~5% to as high as nearly ~50%. This high variability is likely due to differences in total microbial content from each individual mouse and the technical limitations of the RZPM probes designed against human microbiomes to deplete the rRNA content effectively in relatively dissimilar mouse samples. If the samples with higher rRNA contamination are the result of a single or even few specific species, it would be a straightforward and relatively simple task to design new probes directly against the rRNA sequences from those species. However, if the contamination is from many species, probe design would potentially become too cumbersome and would likely benefit from a more taxonomic-free approach. To ascertain the validity of this assumption, the filtered rRNA reads from a subset of both undepleted and depleted cecal samples were collected and processed using the BSSH Dragen Metagenomics Pipeline to confer proportional taxonomic identification to the rRNA content (Figure 1C). The results suggest that, following depletion, there is not a single or even a few taxa that dominate the remaining left over rRNA reads in these samples, suggesting the contaminating reads are spread across multiple species. To optimize the performance of rRNA depletion of the cecal samples, we undertook a probe design strategy that is taxonomically neutral and instead relies on the most abundant rRNA sequences in the sample [1].

Probe design and testing

We took an approach to probe design similar to a strategy used to create a probe set for rRNA removal from human gut microbiome samples [1]. The strategy relies on sequence abundance and not on taxonomic information to establish probe sequences. A second consideration is that since Ribo-Zero Plus relies on the use of DNase to remove the probes following enzymatic depletion, too many extra probes could saturate the ability of DNase to digest them. This could potentially lead to excess oligonucleotides being incorporated into the library preparation and result in undesired sequences contaminating the analysis results. For the current set of studies, we therefore examined approaches to reduce the quantity of probes being used for each depletion (Figure 2). Effective spacing, filtering, and alignment during probe design minimizes both the risk of DNase saturation and cost per experiment, which is especially relevant for studies using model organisms like mouse where potentially large numbers of samples are processed.

As previously established [1], following alignment to the rRNA database, the short sequencing reads tend to cluster into longer stretches, or regions, of rRNAs. We therefore limited the number of additional mouse specific probes by altering two main parameters of the design process: the number of abundant rRNA regions per sample used for designs and the spacing of the probes across the regions of interest. The region numbers and spacing option strategy is summarized in Figure 2. We chose 7 of the 56 cecal samples demonstrating high rRNA levels (~32-55% of total reads) and subjected them to further analysis. The abundant rRNA regions were collected from each of these samples and ranked according to coverage depth. For each of the 7 individual samples, we categorized and collected the top 50, 30, 25, or 20 most abundant regions and then for each category the samples were combined. To limit redundant sequences, the regions within each category were subjected to pairwise alignment. If any two sets of regions demonstrate at least 80% sequence identity, only one is randomly selected for further processing. Within the remaining regions, 50 nucleotide antisense probes were designed; however, the gap between adjacent probes was altered from 25-50 nucleotides apart in 5 nucleotide increments. The expectation is that probes designed against a larger number of regions (i.e. top 50 vs top 20) and spaced closer together (i.e. Gap 25 vs Gap 50) will inherently result in a larger number of probes than selecting a smaller number of regions and probes spaced further apart. The results of the number of probes designed per category from this pipeline are shown in Table 1 and demonstrate the expected trends. For example, designs based upon the Top 50 regions and spaced 25 nucleotides apart results in 492 probes, but if spaced apart by 50 nucleotides results in 380 probes. In addition, comparing the use of more vs less abundant regions within each spacing category also demonstrates the expected trends.

While not practical to test every probe design option, we chose to continue with two probe designs that essentially fall at each end of the design spectrum. Furthermore, considering cost and convenience of probe synthesis, we wanted to limit the number of probes to <384 oligonucleotides. For these reasons, we chose to move forward with the probe designs for the Top 50 regions, spaced 50 nucleotides apart (the 5050 probe set; 380 oligos) and the designs for the Top 20 spaced 25 nucleotides apart (the 2025 probe set; 317 oligos). These probe sets were synthesized, combined with both DP1 and DPM, and tested for rRNA depletion efficiency in a subset of the cecal samples used in the above experiments.

Improved efficiency of rRNA depletion using mouse-specific supplementary probes

Fifteen samples were chosen from the initial 56 tested previously, including 6 taken from the learning set used for probe design, and were subjected to three different depletion options; depleted solely with DP1 and DPM, or depleted with DP1 and DPM plus the 5050 or 2025 probe sets (Figure 3A). In 15/15 samples, the addition of the extra probes resulted in a reduction of rRNA reads, regardless of whether the samples were used for probe design or not. Furthermore, for the majority of samples (9/15), the 2025 probe set demonstrated better depletion than the 5050 set. This demonstrates that the addition of either of the new probe sets improves rRNA depletion performance across multiple samples/donors. Perhaps a more practical measure of the performance of the added probes is the percentage of retained reads that are ultimately used for MetaT analysis. The proportion of retained reads from the samples with or without the supplemental probes are summarized in Figure 3B. Depletion with DP1 and DPM results in a sample mean of ~60% retained, but also indicate a relatively broad range from ~40-90%. Addition of the 2025 probes significantly increased the average percent retained to ~75% ($P < 0.01$, see Supplementary Table 2), but also improved the consistency of depletion between samples as well. The range of the percent retained narrowed to ~60-90%. The 5050 probe set also improved the amount of retained reads recovered, but was not statistically significant compared depletion with DP1 and DPM (Supplementary Table 2), indicating that the 2025 probes demonstrate better performance in these samples. These results show that the 2025 probe design strategy was able to both improve the efficiency (average % mRNA reads) and consistency (spread between samples) of rRNA depletion from the cecal samples, resulting in an average recovery of ~15% more reads for functional data analysis.

To understand how the addition of the extra probe sets could impact the taxonomic distribution of rRNA reads, we collected the remaining rRNA content from representative samples, including two of the samples used for probe design and two that were not, and performed taxonomic analysis (Figure 3C). The purpose was to determine if any of the leftover rRNA reads are overtly dominated by a particular species in the samples following depletion; which might suggest the probe design strategy somehow missed certain abundant sequences. As was observed previously (Figure 1C), in the samples used for probe design, CC53 and CC54, no particular species substantially dominates the rRNA content following depletion with DP1 and DPM (Figure 3C). Conversely, the remaining rRNA content of the two samples not used for probe design (CC41 and CC63), is less evenly distributed and contains a larger proportion of *Lachnospiraceae*. Addition of either the 5050 or 2025 probe sets do not appear to bias the representation of any particular genus or species in the overall taxonomic profile of the samples, generally compressing the proportion of reads assigned to all of them. However, one noticeable exception is *Lachnospiraceae* in samples CC41 and CC63. Despite these samples not being part of the training set, addition of the extra probes did provide a change in the proportion of reads assigned to this species. In particular, CC41 shows a major reduction in *Lachnospiraceae* rRNA reads following the addition of the 5050 probe set and a more modest depletion upon addition of the 2025 probes. This suggests that even though these two samples were not used for the learning set, the generated probes are still effective at removing rRNA content not specifically included in the design. These results demonstrate that this probe design strategy can provide depletion outside the taxonomic range of the samples used in the training set, reinforcing the value of this strictly abundance-based and species-agnostic approach to the design method.

Analysis of off target effects from additional probes

An important consideration when supplementing additional probes for rRNA depletion is whether there are any off-target or otherwise non-specific effects of the probes on the remaining content that could impact the ability to perform either host transcriptome or MetaT analysis. Any of the additional probes designed against the most abundant sequences could potentially hybridize to host or microbial mRNAs and result in depleting those regions of the transcript and potentially causing erroneous gene expression profiling results. To investigate this possibility, we first analyzed the mouse transcriptome by gene expression comparisons of host transcripts with or without the additional probes (Figure 4A). The cecal samples contain very little host transcriptome (Figure 1B), so we utilized a subset of the liver samples to perform differential gene expression analysis of the host transcriptome, comparing depletion with only DP1 and DPM vs. the inclusion of the additional probes (Figures 4A, 4B). In both cases, addition of the supplemental probes for depletion does not notably impact host gene expression in the mouse live samples. with a Pearson R^2 correlation of 0.96 when adding either the 5050 or 2025 probe-sets compared to just DP1 and DPM (Figure 4A). We also compared gene family relative abundance within the metatranscriptome (Figures 4B and 4C) with or without addition of the probe sets. To examine any possible detrimental effects the additional probes may have on the MetaT results, we performed a rank correlation analysis of the cecal samples from Figure 3, specifically focusing on the correlation between the relative abundance of gene pathways. Two samples were chosen for this analysis: CC53 (used for probe design; Figures 4C, 4D, 4E) and CC41 (used for probe design; Figures 4F, 4G, 4H). Spearman's rank correlation (ρ) demonstrates pairwise comparisons with quite good concordance between probe depletion strategies, where ρ values range from ~0.92-0.95. This indicates that inclusion of the additional probes has no obvious detrimental effects on MetaT analysis. Together, the results suggest that rRNA depletion using either the 5050 or 2025 probe set causes minimal bias for transcriptome profiling of these complex sample types.

Discussion

In this work, we describe a novel strategy to adapt and optimize a probe pool initially created for the enzymatic depletion of abundant rRNA sequences in human gut microbiome samples to enable effective and consistent use in mouse cecal samples. Efficient depletion of rRNA sequences in these sample types greatly increases the informational content for use in MetaT analysis. The main goal in these types of studies is to determine the gene expression profile and functional/metabolic characteristics of the microorganisms residing in the host. Knowledge about host-microbe interplay relies on understanding both microbiome contribution as well as the host response and how it can vary with specific gene expression contributions from the microbial community. At our current level of understanding, it is not clear to what extent host response is dependent on specific microorganisms being present. Thus, being able to compare the microbiome population composition to the metatranscriptomic profile during controlled stimulation experiments is critical.

The approach to probe design used in this work is agnostic towards the taxonomic content of the samples and instead relies on collecting abundant microbial rRNA sequences from several mouse cecal donors, setting coverage thresholds, and varying probe spacing options to efficiently create additional probes optimized for mouse cecal microbial rRNA content. We initially tested 56 individual cecal donors using the RZPM probe sets (DP1 & DPM) designed against human stool and determined that for many donors, the rRNA depletion resulted in <30% of the sequencing reads matching microbial rRNA. However, several donors demonstrated higher rRNA content; in some cases, up to ~50% of reads. We chose 7 donors with rRNA content >30% for use as a training set for additional probe design. Abundant rRNA sequences were collected from these individual samples and merged to create abundant rRNA 'regions' and then ranked by median coverage depth from highest to lowest. To investigate probe design efficiency, we combined two options. First, we filtered sequences based upon specific coverage thresholds and second, designed anti-sense probes with various spacing across the rRNA regions selected by coverage. This matrix of options produced probe design pools with various numbers of individual oligos that trend as generally expected; designs against higher numbers of abundant regions and spaced closer together generate more probes than using fewer regions and spacing the probes further apart. We then chose two design pools that represent compromises between these two extremes (the green and blue cells in Table 1) and tested their performance on both the same cecal samples used for training as well as samples that were not. Both design options demonstrate improved rRNA depletion performance, but the more 'packed' version (using the top 20 most abundant regions and probes spaced 25 nt apart; named the 2025 probe set) results in a statistically significant increase in the percentage on non-rRNA reads generated for MetaT analysis. In addition, the ranges of rRNA content from different donors is generally more compact and consistent when compared to the use of the human gut-centric probes by themselves.

The microbiome is present virtually throughout the human body [25]. However, most areas of the human biome in healthy individuals contain only trace amounts of microbes, with the bulk present within the digestive and respiratory tracts, which provide an enriched conditions for their growth due to contact with the external environment [26]. This knowledge has resulted in a great deal of investigation into gut microbiomes and the role they play in the health of their hosts, as well as the effect of the host's diet and health on microbial composition and gene expression. Several studies have suggested an interplay between the microbiome metabolic activity and the response of the host to various interventions. The diet of the host is perhaps an obvious example where consumption of certain nutrients clearly results in changes to the gut flora and activity [3, 27]. Therefore, it is important to consider the ability to monitor both host and microbe response to aspects such as dietary stress, disease states, or even medicinal treatment. With this in mind, we performed rRNA depletion and RNA-Seq analysis of three distinct mouse microbiome types: cecal, terminal ileum, and liver. The proportional amount of microbial vs host transcriptome varies considerably between these sample types. The liver and terminal ileum samples contain predominantly host RNA, while the cecal content is mostly microbial. The existing RZPM kit is designed to remove the common rRNA transcripts from human, mouse, and rat samples via the use of the DP1 probe pool. Furthermore, previous work established the use of DPM for microbial rRNA removal from human stool, vaginal, and oral microbiomes. This provides an integrated tool to allow the simultaneous gene expression profiling of both host and human microbiome. A major goal of this current work is to provide tools that will maximize the amount of useful MetaT information available through the use of an optimized set of supplemental probe pools for mouse cecal content.

In addition to providing a manageable set of depletion probes, we demonstrate that the use of these supplemental probes does not introduce unwanted bias to either the host- or the meta-transcriptome, in terms of microbial species or metabolic pathway gene expression. Our results suggest that the combination of the 2025 probe set with the RZPM probes (DP1 and DPM) provides an effective tool to enable MetaT analysis of mouse cecal microbiomes.

The ability to remove abundant unwanted sequences from total RNA is critical for any whole transcriptome RNA-Seq analysis [12]. Sequencing total RNA without rRNA removal generally results in the vast majority of sequencing reads aligning to rRNA. Microbiome samples are especially susceptible; typically >95% of all reads are identified as microbial rRNA (Figure 1). In some studies, this rRNA content has been used to track the taxonomic profile of microbiomes and rank species abundance based upon the proportion of reads assigned to each species. Much like metagenome analysis, the proportion of each species can then be used as a proxy to infer metabolic or functional activity of the population [28, 29]. However, microbiomes are extremely complex systems, containing hundreds to thousands of different species and the proportional existence of a particular taxon or species related to its metabolic contribution to particular enzymatic pathways is inferential at best. Abundant species may be relatively quiescent or vice versa, and small species populations may be extremely significant to the metabolic state of the population. Furthermore, microbiomes are typically highly dynamic environments/ecosystems that will quickly adapt to changes in diet, disease state, and the various medicines used to treat them. It is for these reasons that having the means to effectively profile microbial gene expression changes via MetaT analysis is critical to our understanding of microbiomes as well as their influence and interactions with their hosts.

Disclosures

AZ is a co-founder and acting CMO of Endure Biotherapeutics. He holds equity in the company.

AZ is supported by NIH R01 EB030134, R01 AI163483, U01 CA265719, and VA Merit BLR&D Award I01 BX005707. Some authors receive institutional support from NIH P30 DK120515, P30 DK063491, P30 CA014195, and UL1 TR001442.

References

1. Tan, A., et al., *Rational probe design for efficient rRNA depletion and improved metatranscriptomic analysis of human microbiomes.* BMC Microbiol, 2023. **23**(1): p. 299.
2. Gilbert, J.A., et al., *Current understanding of the human microbiome.* Nat Med, 2018. **24**(4): p. 392-400.
3. McDonald, D., et al., *American Gut: an Open Platform for Citizen Science Microbiome Research.* mSystems, 2018. **3**(3).
4. Wensel, C.R., et al., *Next-generation sequencing: insights to advance clinical investigations of the microbiome.* J Clin Invest, 2022. **132**(7).
5. Maghini, D.G., et al., *Quantifying bias introduced by sample collection in relative and absolute microbiome measurements.* Nat Biotechnol, 2024. **42**(2): p. 328-338.
6. Bashiardes, S., G. Zilberman-Schapira, and E. Elinav, *Use of Metatranscriptomics in Microbiome Research.* Bioinform Biol Insights, 2016. **10**: p. 19-25.
7. Schlaberg, R., *Microbiome Diagnostics.* Clin Chem, 2020. **66**(1): p. 68-76.

8. Ley, R.E., et al., *Microbial ecology: human gut microbes associated with obesity.* Nature, 2006. **444**(7122): p. 1022-3.
9. Laudadio, I., et al., *Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome.* Omics, 2018. **22**(4): p. 248-254.
10. Lloyd-Price, J., et al., *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature, 2017. **550**(7674): p. 61-66.
11. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
12. Zhang, Y., et al., *Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling.* Annu Rev Biomed Data Sci, 2021. **4**: p. 279-311.
13. Ojala, T., E. Kankuri, and M. Kankainen, *Understanding human health through metatranscriptomics.* Trends Mol Med, 2023. **29**(5): p. 376-389.
14. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nat Methods, 2018. **15**(11): p. 962-968.
15. Chung, Y.W., et al., *Functional dynamics of bacterial species in the mouse gut microbiome revealed by metagenomic and metatranscriptomic analyses.* PLoS One, 2020. **15**(1): p. e0227886.
16. Fogelson, K.A., et al., *The Gut Microbial Bile Acid Modulation and Its Relevance to Digestive Health and Diseases.* Gastroenterology, 2023. **164**(7): p. 1069-1085.
17. Franzosa, E.A., et al., *Relating the metatranscriptome and metagenome of the human gut.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. E2329-38.
18. Culviner, P.H., C.K. Guegler, and M.T. Laub, *A Simple, Cost-Effective, and Robust Method for rRNA Depletion in RNA-Sequencing Studies.* mBio, 2020. **11**(2).
19. Gu, W., 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, 2016. **17**: p. 41.
20. Wahl, A., C. Huptas, and K. Neuhaus, *Comparison of rRNA depletion methods for efficient bacterial mRNA sequencing.* Sci Rep, 2022. **12**(1): p. 5765.
21. Huang, Y., et al., *Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics.* Nucleic Acids Res, 2020. **48**(4): p. e20.
22. Nguyen, T.L., et al., *How informative is the mouse for human gut microbiota research?* Dis Model Mech, 2015. **8**(1): p. 1-16.
23. Beresford-Jones, B.S., et al., *The Mouse Gastrointestinal Bacteria Catalogue enables translation between the mouse and human gut microbiotas via functional mapping.* Cell Host Microbe, 2022. **30**(1): p. 124-138.e8.
24. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
25. Turnbaugh, P.J., et al., *The human microbiome project.* Nature, 2007. **449**(7164): p. 804-10.
26. Rastogi, S., et al., *Possible role of gut microbes and host's immune response in gut-lung homeostasis.* Front Immunol, 2022. **13**: p. 954339.
27. Pérez-Gerdel, T., et al., *Impact of Intermittent Fasting on the Gut Microbiota: A Systematic Review.* Adv Biol (Weinh), 2023. **7**(8): p. e2200337.
28. David, L.A., et al., *Host lifestyle affects human microbiota on daily timescales.* Genome Biol, 2014. **15**(7): p. R89.
29. Johnson, J.S., et al., *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.* Nat Commun, 2019. **10**(1): p. 5029.

30. Beghini, Francesco, et al. "Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3." elife 10 (2021): e65088.

Figure legends:

Figure 1: Depletion of rRNA from mouse cecal, ileum and liver samples for RNAseq library preparation and sequencing. A) Summary of library preparation and sequencing. Total RNA samples were rRNA depleted and converted into stranded Total RNA libraries and sequenced. The resulting raw fastq files were quality filtered with Trimmomatic to remove any reads that were less than 50 bases and below Q20. The remaining reads were then aligned with BBDUK to rRNA sequences using the Silva database as the reference. The detected rRNA reads are collected as a new fastq file that was used for taxonomic analysis to determine proportion of species contributing to the rRNA contamination. Reads filtered for rRNA are further processed for MetaT analysis using Humann3 to assign gene families and metabolic pathways. B) Analysis of host and rRNA content of 56 cecal, 62 ileum and 62 liver samples (average of 2 replicates shown, see supplentary file 1 for ordered list of sample names). Liver and ileum samples are dominated by host transcriptome (red shading, mm9 genome reference) while cecal samples are primarily metatranscriptome reads (Blue = % retained; the proportion of reads that are not host or rRNA used for MetaT analysis). LSU and SSU refer to rRNA large subunit and small subunit, respectively, for each species type listed. Host rRNA refers to reads that align to mouse rRNA. Rfam refers to other RNA family types, like tRNA. The asterisks at the top indicate the samples chosen for probe design in figure 2 (CC6, CC7, CC9, CC51, CC53, CC54, CC58). C) Taxonomic analysis of the collected rRNA reads from either undepleted (left side) or samples depleted (right side) with DP1 & DPM indicating that no particular species dominate the remaining rRNA content of the samples following depletion. Samples CC54, CC53, CC41, CC63 shown (see supplemental table XX). The asterisks at the top indicate two of the samples chosen for probe design in figure 2.

Figure 2: Probe design strategy based on most abundant reads. A) From the 56 cecal samples that were initially processed, 7 samples with %rRNA reads >30% were chosen as the learning set for additional probe design. For each sample, the rRNA reads were collected and aligned to the Silva database to determine the coverage depth across regions of rRNA transcripts. Any regions that were covered >500x were collected and ranked from highest to lowest. Two filtering options were used to minimize the number of probes needed. First, the regions that make up the top 50, 30, 25 or 20 most abundant sequences were binned and the equivalent regions from each sample were combined. Second, for each combined set of regions, antisense probes were designed with various spacing options ranging from 25 to 50 nt apart. The probes can then be ordered, combined with DP1 and DPM and the Total RNA samples re-tested for rRNA depletion and overall RNAseq performance.

Table 1: The number of probes designs relative to the two main criteria used; the number of abundant regions used and the spacing of the probes across the target regions. Two pools were chosen for synthesis and further analysis. The 5050 pool is composed of 380 probes designed against the top 50 most abundant regions and spaced 50 nt apart. The 2025 pool of 317 probes is designed to target the top 20 most abundant regions and spaced 25 nt apart.

Figure 3: Inclusion of supplementary probes improves rRNA depletion and increases the proportion of reads available for MetaT analysis. A) Percentage of rRNA reads in samples depleted using standard probes (DP1 + DPM) or with supplementary probes 5050 or 2025 included. Samples on the left were used for supplementary probe designs and samples on the right were not. In both cases, the majority of samples (14/15) demonstrate a decrease in the rRNA reads upon inclusion of the additional probes. Inclusion of the 2025 probe set results in the least amount of rRNA reads. B) Inclusion of the 2025 probe set both increases the percentage of reads available for MetaT analysis by ~15% and provides greater

consistency between individual donors. C) Taxonomic representation of rRNA reads remaining following depletion conditions comparing the standard probes alone (left, DP1 + DPM) or when the 5050 (middle) or 2025 (right) supplementary probes are included. Four representative samples are shown; two of the samples (CC54 & CC53) were used for probe design and two samples (CC41 & CC63) were not. The results indicate that the newly designed probes do not result in targeting the rRNA from only a few specific taxa but are more dispersed across several taxa.

Figure 4: Inclusion of supplementary probes does not result in significant bias to MetaT analysis. A) and B) Gene expression comparison of liver samples from 3 mice treated as replicates (LVR20, LVR25, LVR49) comparing depletion conditions using the standard probes (DP1_DPM) vs inclusion of the 5050 (A) or 2025 (B) probes sets. In both cases, the Pearson R2 values of 0.96 suggests no significant bias is introduced to host transcriptome upon inclusion of the supplementary probes. Scatter plots of normalized gene family abundances, where each point corresponds to a gene family, comparing depletion conditions in either a sample used for probe design (CC54; C-E) or a sample that was not (CC41; F-H). The X and Y axes show the fraction of total abundance contributed by that gene family in each sample. The Spearman's rank coefficient, or Rho (ρ), indicates how similarly the gene families are distributed between the two samples, higher values suggest a stronger correlation. In all comparisons ρ demonstrates strong agreement (>0.91) suggesting minimal bias is introduced to the metatranscriptomes by inclusion of the supplementary probes.

Supplementary Table 1: List of mouse Cecum, Ileum and Liver RNA samples used for RNAseq analysis.

Supplementary Table 2: Results of two-tailed t-test of significance. One way analysis of % retained by depletion method (see Figure 3B). The comparison of DP1+DPM vs DP1+DPM+2025 is significant with a P < 0.01.

Supplementary Figure 1: Taxonomic analysis of the collected rRNA reads from the top 100 most abundant species (sorted by sample CC53 – No Depletion) either from undepleted (left side) or samples depleted with DP1 & DPM (right side) indicating that no particular species dominate the remaining rRNA content of the samples following depletion. See Figure 1C for the Top 15 species. Samples CC54, CC53, CC41, CC63. The asterisks at the top indicate two of the samples chosen for probe design in Figure 2.

Supplementary Figure 2: Taxonomic analysis of the collected rRNA reads from the top 100 most abundant species (sorted by sample CC53 – DP1+DPM) either from samples depleted with DP1+DPM (left side), DP1+DPM+5050 (middle), or DP1+DPM+2025 (right side) indicating that no particular species dominate the remaining rRNA content of the samples following depletion. See Figure 3C for the Top 15 species. The asterisks at the top indicate two of the samples chosen for probe design in Figure 2.

Figure 1

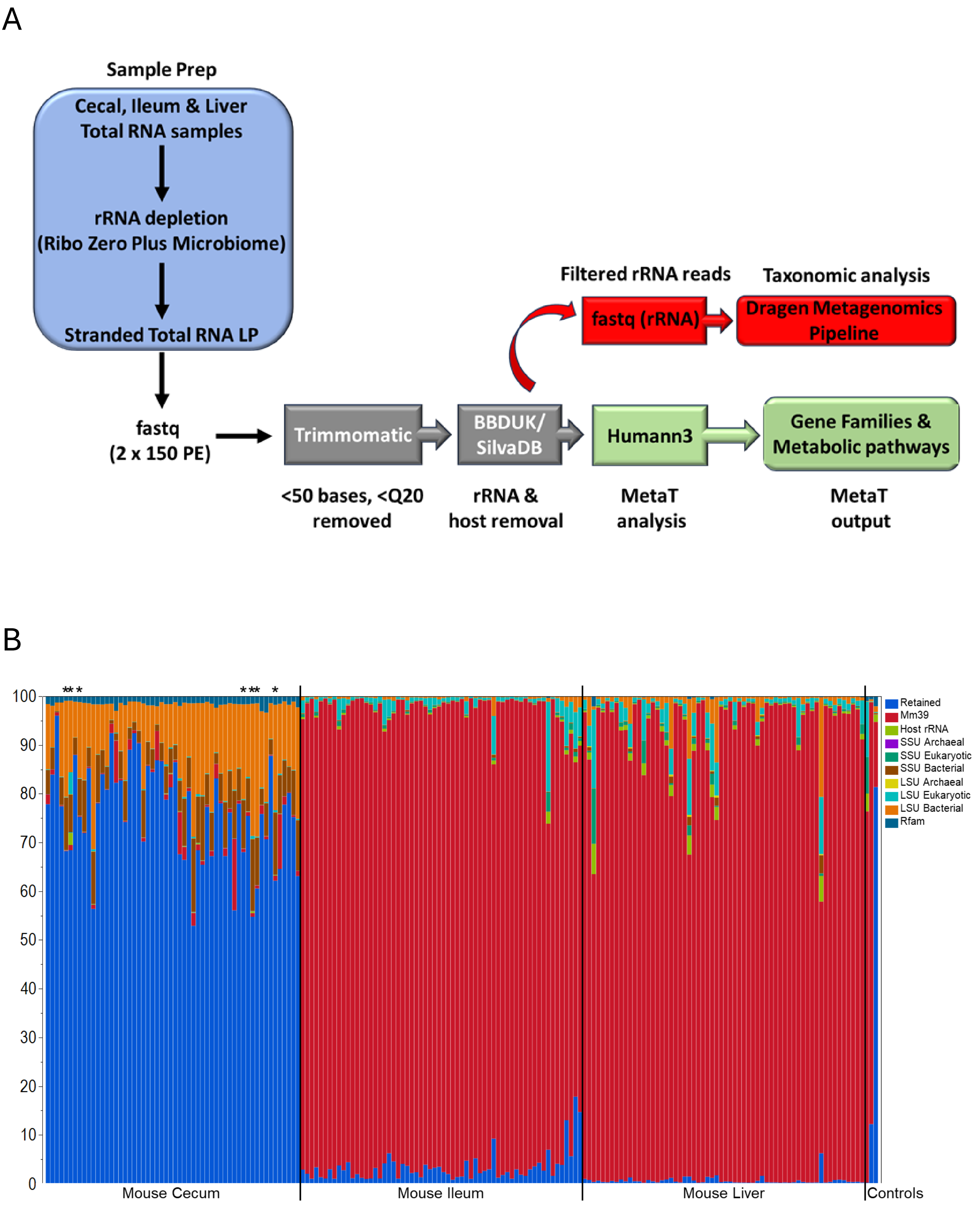

A
Sample Prep
Cecal, Ileum & Liver
Total RNA samples
rRNA depletion
(Ribo Zero Plus Microbiome)
Stranded Total RNA LP
fastq
(2 x 150 PE)
Trimmomatic
<50 bases, <Q20
removed
BBDUK/
SilvaDB
rRNA &
host removal
Filtered rRNA reads
fastq (rRNA)
Taxonomic analysis
Dragen Metagenomics
Pipeline
Humann3
MetaT
analysis
Gene Families &
Metabolic pathways
MetaT
output
B
** *
** *
*
100
90
80
70
60
50
40
30
20
10
0
Mouse Cecum
Mouse Ileum
Mouse Liver
Controls
Retained
Mm39
Host rRNA
SSU Archaeal
SSU Eukaryotic
SSU Bacterial
LSU Archaeal
LSU Eukaryotic
LSU Bacterial
Rfam

Figure 1 - expanded

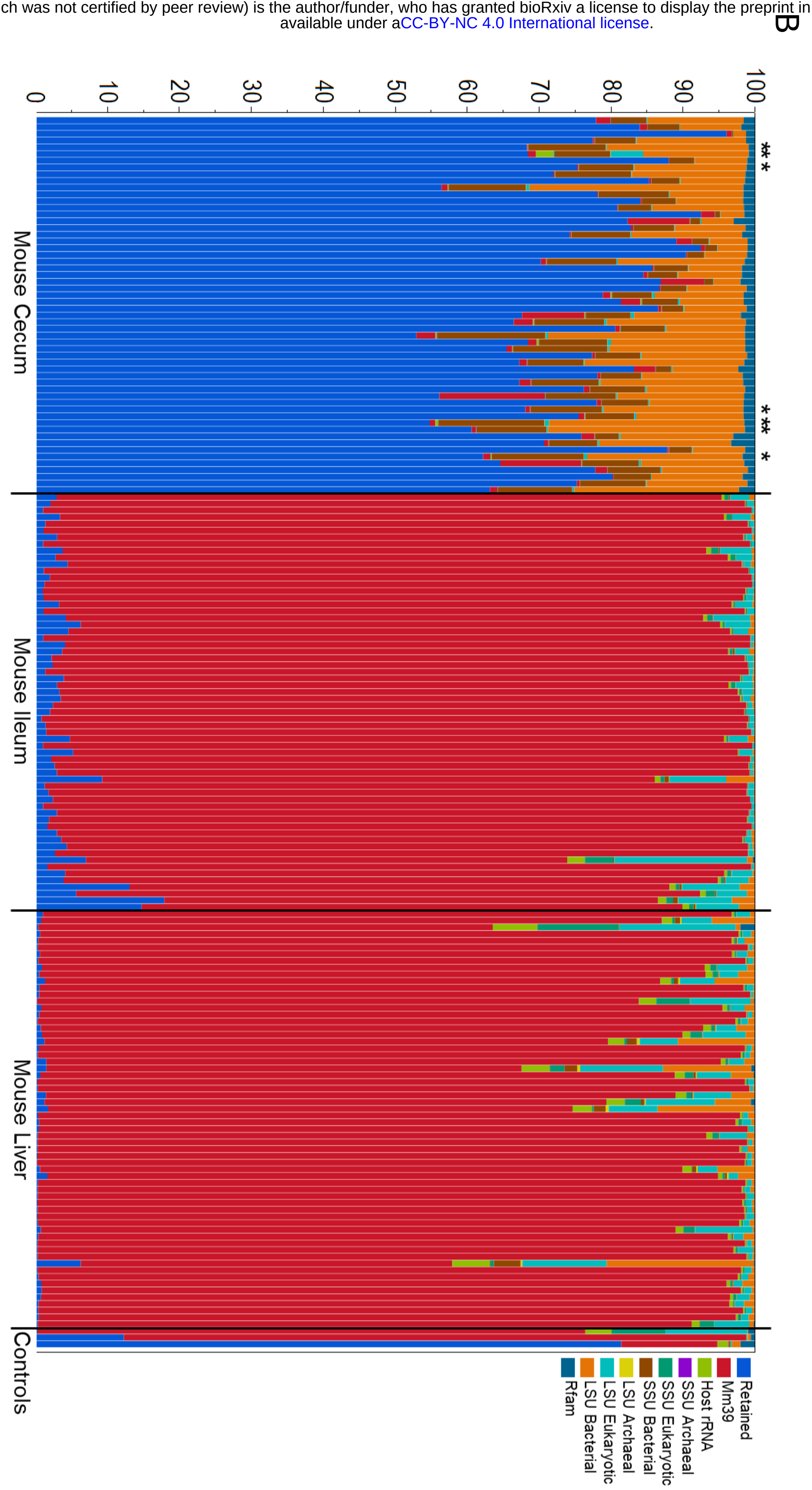


Figure 1

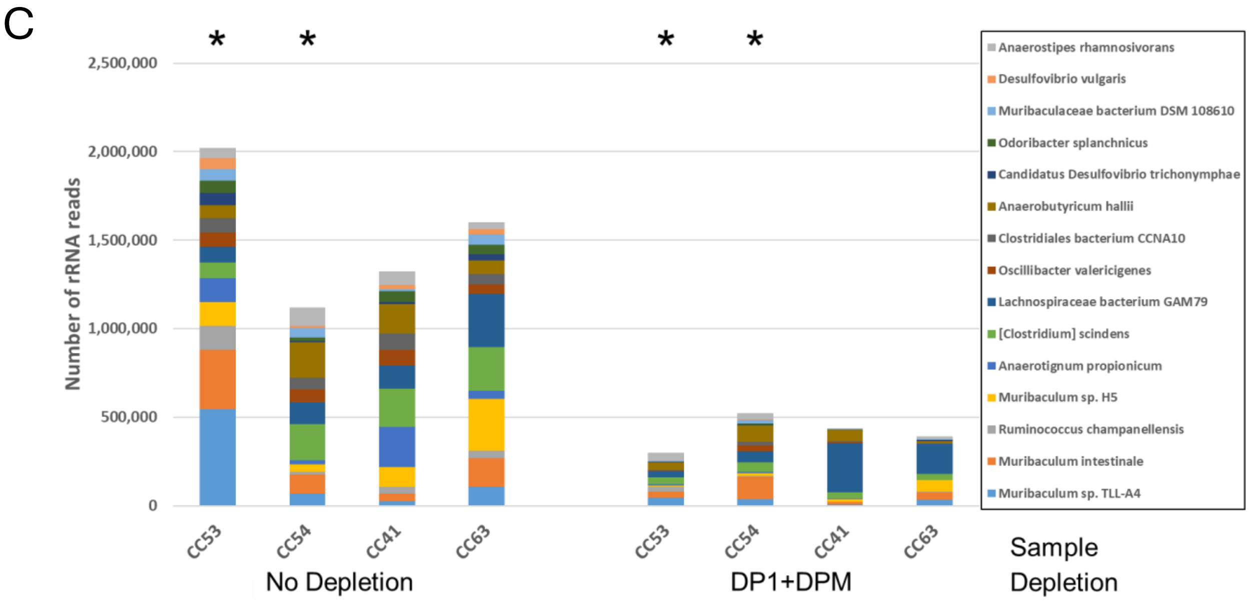


Figure 1: Depletion of rRNA from mouse cecal, ileum and liver samples for RNAseq library preparation and sequencing. A) Summary of library preparation and sequencing. Total RNA samples were rRNA depleted and converted into stranded Total RNA libraries and sequenced. The resulting raw fastq files were quality filtered with Trimmomatic to remove any reads that were less than 50 bases and below Q20. The remaining reads were then aligned with BBDUK to rRNA sequences using the Silva database as the reference. The detected rRNA reads are collected as a new fastq file that was used for taxonomic analysis to determine proportion of species contributing to the rRNA contamination. Reads filtered for rRNA are further processed for MetaT analysis using Humann3 to assign gene families and metabolic pathways. B) Analysis of host and rRNA content of 56 cecal, 62 ileum and 62 liver samples (average of 2 replicates shown, see supplentary file 1 for ordered list of sample names). Liver and ileum samples are dominated by host transcriptome (red shading, mm9 genome reference) while cecal samples are primarily metatranscriptome reads (Blue = % retained; the proportion of reads that are not host or rRNA used for MetaT analysis). LSU and SSU refer to rRNA large subunit and small subunit, respectively, for each species type listed. Host rRNA refers to reads that align to mouse rRNA. Rfam refers to other RNA family types, like tRNA. The asterisks at the top indicate the samples chosen for probe design in figure 2 (CC6, CC7, CC9, CC51, CC53, CC54, CC58). C) Taxonomic analysis of the collected rRNA reads from either undepleted (left side) or samples depleted (right side) with DP1 & DPM indicating that no particular species dominate the remaining rRNA content of the samples following depletion. Samples CC54, CC53, CC41, CC63 shown (see Supplemental Table 1). The asterisks at the top indicate two of the samples chosen for probe design in Figure 2. Only the top 15 most abundant species (sorted by sample CC53-No Depletion) are shown. A more comprehensive analysis of the top 100 species is shown in Supplementary Figure 1.

Figure 2

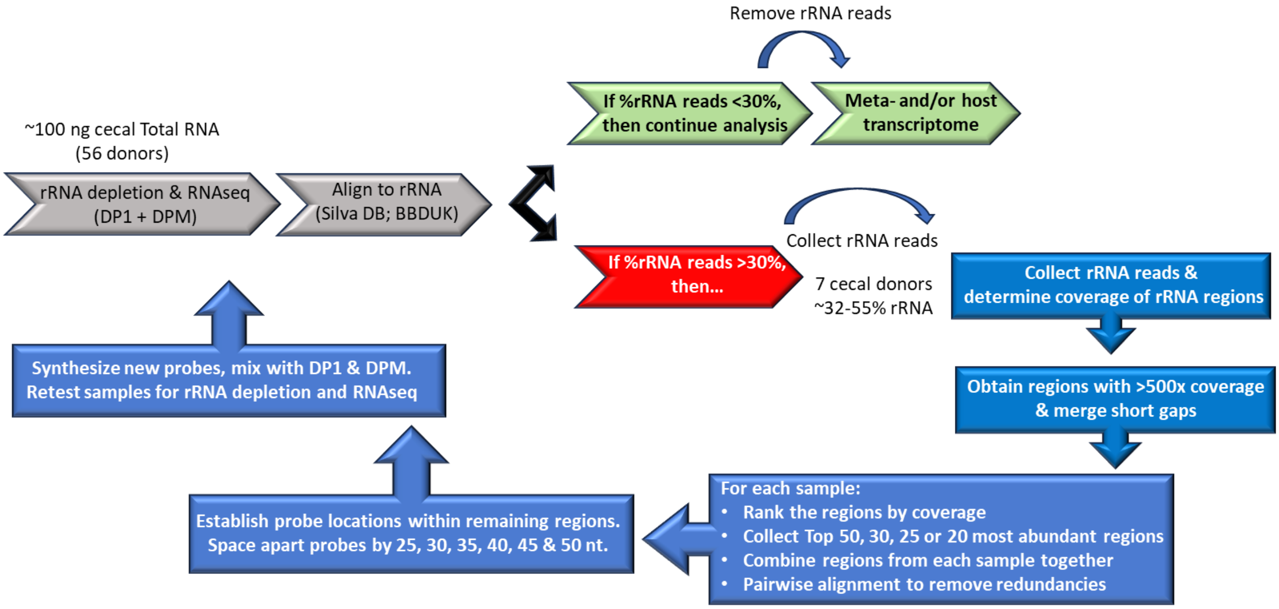
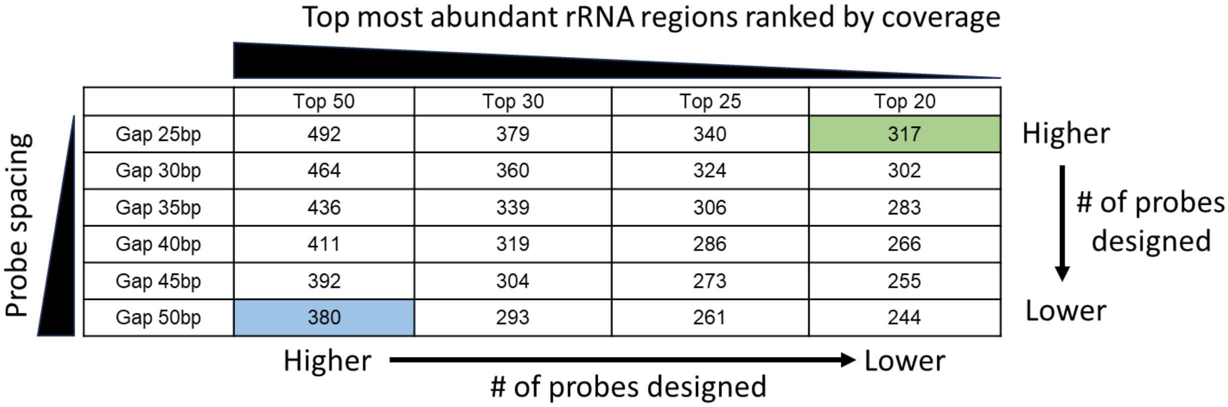

~100 ng cecal Total RNA
(56 donors)
rRNA depletion & RNAseq
(DP1 + DPM)
Align to rRNA
(Silva DB; BBDUK)
Remove rRNA reads
If %rRNA reads <30%,
then continue analysis
Meta- and/or host
transcriptome
Collect rRNA reads
If %rRNA reads >30%,
then...
7 cecal donors
~32-55% rRNA
Collect rRNA reads &
determine coverage of rRNA regions
Obtain regions with >500x coverage
& merge short gaps
For each sample:
• Rank the regions by coverage
• Collect Top 50, 30, 25 or 20 most abundant regions
• Combine regions from each sample together
• Pairwise alignment to remove redundancies
Establish probe locations within remaining regions.
Space apart probes by 25, 30, 35, 40, 45 & 50 nt.
Synthesize new probes, mix with DP1 & DPM.
Retest samples for rRNA depletion and RNAseq

Figure 2: Probe design strategy based on most abundant reads. A) From the 56 cecal samples that were initially processed, 7 samples with %rRNA reads >30% were chosen as the learning set for additional probe design. For each sample, the rRNA reads were collected and aligned to the Silva database to determine the coverage depth across regions of rRNA transcripts. Any regions that were covered >500x were collected and ranked from highest to lowest. Two filtering options were used to minimize the number of probes needed. First, the regions that make up the top 50, 30, 25 or 20 most abundant sequences were binned and the equivalent regions from each sample were combined. Second, for each combined set of regions, antisense probes were designed with various spacing options ranging from 25 to 50 nt apart. The probes can then be ordered, combined with DP1 and DPM and the Total RNA samples re-tested for rRNA depletion and overall RNAseq performance.

Table 1



	Top 50	Top 30	Top 25	Top 20
Gap 25bp	492	379	340	317
Gap 30bp	464	360	324	302
Gap 35bp	436	339	306	283
Gap 40bp	411	319	286	266
Gap 45bp	392	304	273	255
Gap 50bp	380	293	261	244

Top most abundant rRNA regions ranked by coverage
Probe spacing
Higher
# of probes designed
Lower
Higher
Lower
# of probes designed

Table 1: The number of probes designs relative to the two main criteria used; the number of abundant regions used and the spacing of the probes across the target regions. Two pools were chosen for synthesis and further analysis. The 5050 pool is composed of 380 probes designed against the top 50 most abundant regions and spaced 50 nt apart. The 2025 pool of 317 probes is designed to target the top 20 most abundant regions and spaced 25 nt apart.

Figure 3

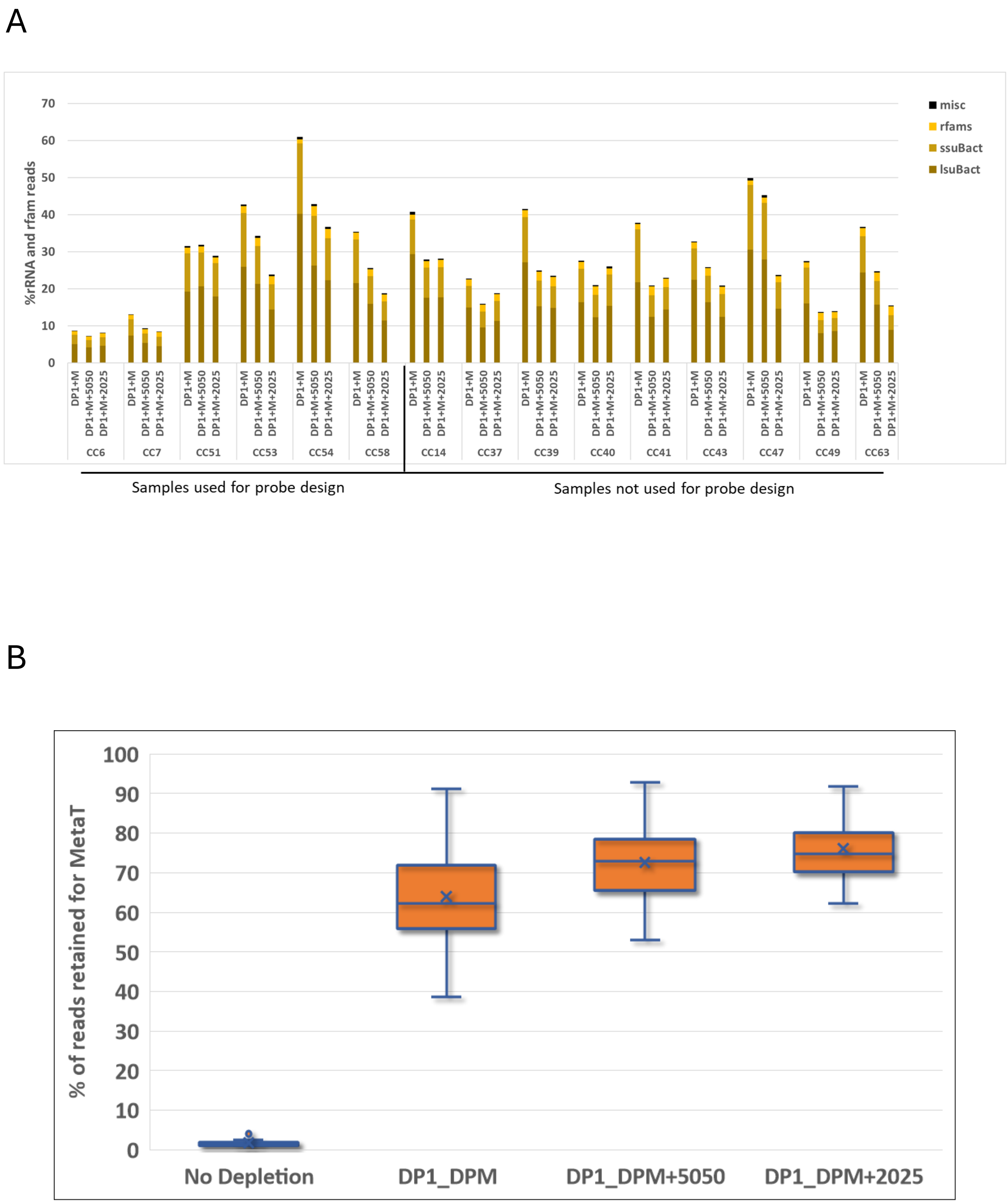

A
%rRNA and rfam reads
70
60
50
40
30
20
10
0
■ misc
■ rfams
■ ssuBact
■ lsuBact
DP1+M
DP1+M+5050
DP1+M+2025
CC6
CC7
CC51
CC53
CC54
CC58
CC14
CC37
CC39
CC40
CC41
CC43
CC47
CC49
CC63
Samples used for probe design
Samples not used for probe design
B
% of reads retained for MetaT
100
90
80
70
60
50
40
30
20
10
0
No Depletion
DP1_DPM
DP1_DPM+5050
DP1_DPM+2025

Figure 3

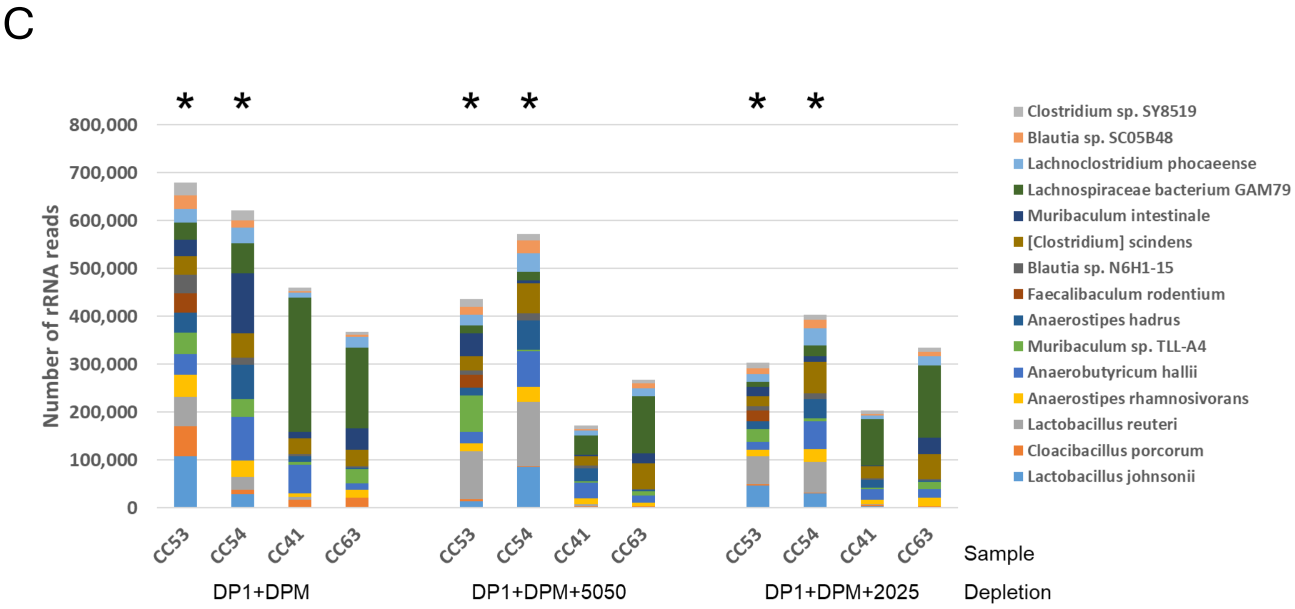


Figure 3: Inclusion of supplementary probes improves rRNA depletion and increases the proportion of reads available for MetaT analysis. A) Percentage of rRNA reads in samples depleted using standard probes (DP1 + DPM) or with supplementary probes 5050 or 2025 included. Samples on the left were used for supplementary probe designs and samples on the right were not. In both cases, the majority of samples (14/15) demonstrate a decrease in the rRNA reads upon inclusion of the additional probes. Inclusion of the 2025 probe set results in the least amount of rRNA reads. B) Inclusion of the 2025 probe set both increases the percentage of reads available for MetaT analysis by ~15% and provides greater consistency between individual donors. C) Taxonomic representation of rRNA reads remaining following depletion conditions comparing the standard probes alone (left, DP1 + DPM) or when the 5050 (middle) or 2025 (right) supplementary probes are included. Four representative samples are shown; two of the samples (CC54 & CC53) were used for probe design and two samples (CC41 & CC63) were not. The results indicate that the newly designed probes do not result in targeting the rRNA from only a few specific taxa but are more dispersed across several taxa. Only the top 15 most abundant species (sorted by sample CC53-DP1+DPM) are shown. A more comprehensive analysis of the top 100 species is shown in Supplementary Figure 2.

Figure 4

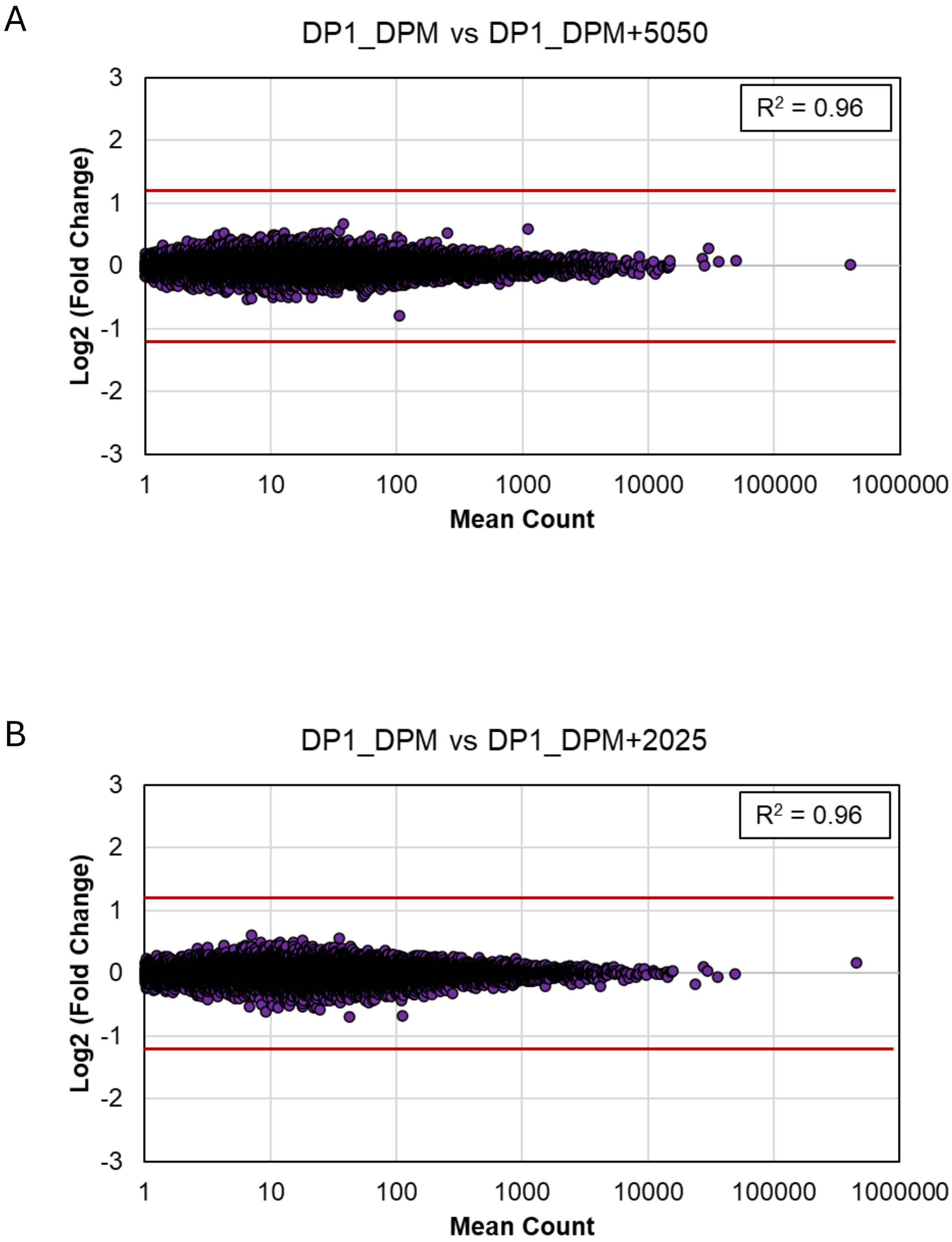

A
DP1_DPM vs DP1_DPM+5050
R^2 = 0.96
Log2 (Fold Change)
3
2
1
0
-1
-2
-3
1
10
100
1000
10000
100000
1000000
Mean Count
B
DP1_DPM vs DP1_DPM+2025
R^2 = 0.96
Log2 (Fold Change)
3
2
1
0
-1
-2
-3
1
10
100
1000
10000
100000
1000000
Mean Count

Figure 4

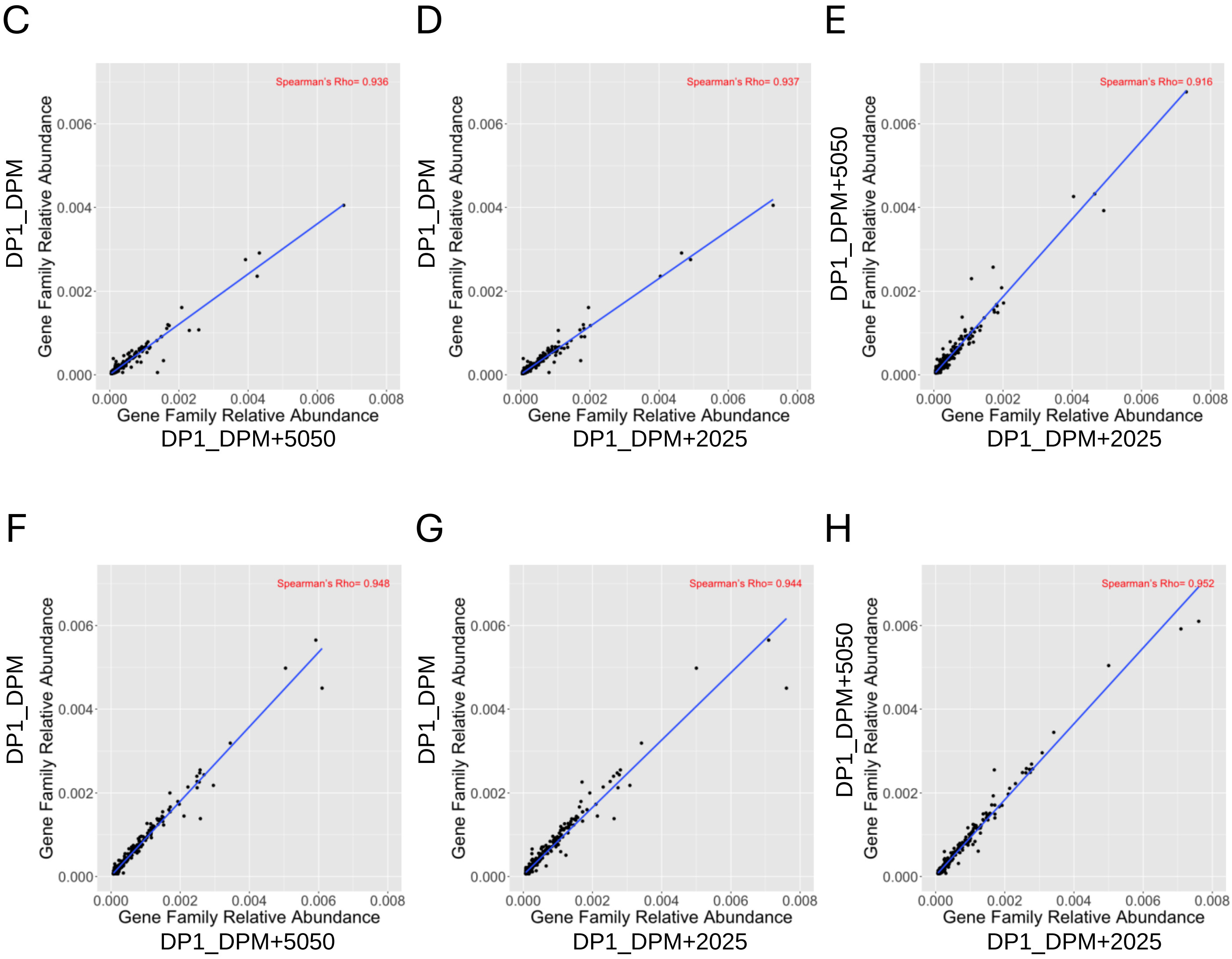


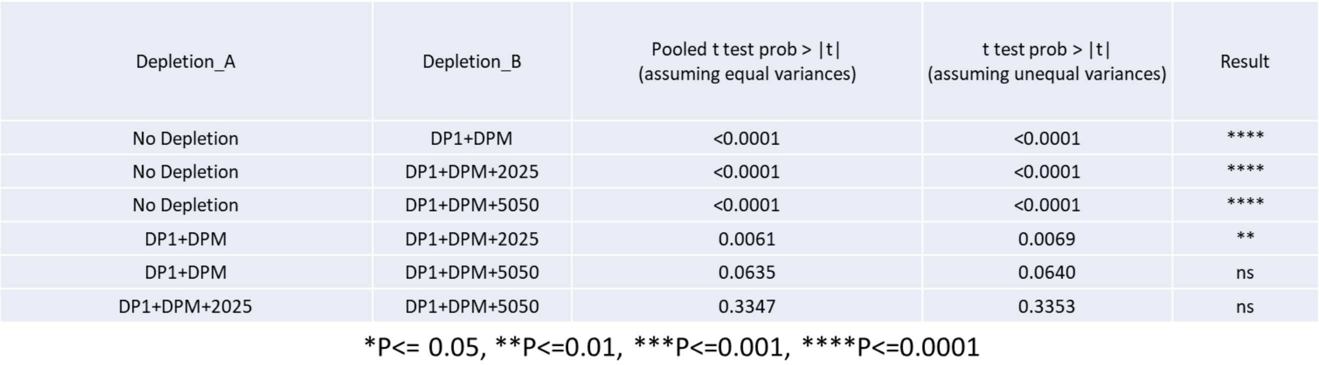
Figure 4: Inclusion of supplementary probes does not result in significant bias to MetaT analysis. A) and B) Gene expression comparison of liver samples from 3 mice treated as replicates (LVR20, LVR25, LVR49) comparing depletion conditions using the standard probes (DP1_DPM) vs inclusion of the 5050 (A) or 2025 (B) probes sets. In both cases, the Pearson R^2 values of 0.96 suggests no significant bias is introduced to host transcriptome upon inclusion of the supplementary probes. Scatter plots of normalized gene family abundances, where each point corresponds to a gene family, comparing depletion conditions in either a sample used for probe design (CC54; C-E) or a sample that was not (CC41; F-H). The X and Y axes show the fraction of total abundance contributed by that gene family in each sample. The Spearman's rank coefficient, or Rho (ρ), indicates how similarly the gene families are distributed between the two samples, higher values suggest a stronger correlation. In all comparisons ρ demonstrates strong agreement (>0.91) suggesting minimal bias is introduced to the metatranscriptomes by inclusion of the supplementary probes.

Supplementary Table 1

Mouse Name – Cecum	Mouse Name – Ileum	Mouse Name – Liver	Mouse Lab ID
N/A	IL1	LVR1	101
CC2	IL2	LVR2	102
CC3	IL3	LVR3	104
CC4	IL4	LVR4	105
CC5	IL5	LVR5	106
CC6	IL6	LVR6	107
CC7	IL7	LVR7	108
CC8	IL8	LVR8	109
CC9	IL9	LVR9	114
N/A	IL10	LVR10	115
CC11	IL11	LVR11	116
CC12	IL12	LVR12	117
N/A	IL13	N/A	118
CC14	IL14	LVR14	119
CC15	IL15	LVR15	120
N/A	IL16	LVR16	121
CC17	IL17	LVR17	122
CC18	IL18	LVR18	124
CC19	IL19	LVR19	125
CC20	IL20	LVR20	126
N/A	IL21	LVR21	127
CC22	IL22	LVR22	128
CC23	IL23	LVR23	129
CC24	IL24	LVR24	131
CC25	IL25	LVR25	132
N/A	IL26	LVR26	134
CC27	IL27	LVR27	135
CC28	IL28	LVR28	136
CC29	IL29	LVR29	142
CC30	IL30	LVR30	144
CC31	IL31	LVR31	145
CC32	IL32	LVR32	201
CC33	IL33	LVR33	202
CC34	IL34	LVR34	203
CC35	IL35	LVR35	204
CC36	IL36	LVR36	205
CC37	IL37	LVR37	206
CC38	IL38	LVR38	207
CC39	IL39	LVR39	208
CC40	IL40	LVR40	210
CC41	IL41	LVR41	211
CC42	IL42	LVR42	212
CC43	IL43	LVR43	216
N/A	IL44	LVR44	217
CC45	IL45	LVR45	218
CC46	IL46	LVR46	219
CC47	IL47	LVR47	220
CC48	IL48	LVR48	222
CC49	IL49	LVR49	223
CC50	IL50	LVR50	230
CC51	IL51	LVR51	231
CC52	IL52	LVR52	232
CC53	IL53	LVR53	233
CC54	IL54	LVR54	234
CC55	IL55	LVR55	250
CC56	IL56	LVR56	251
CC57	N/A	LVR57	317
CC58	IL58	LVR58	318
CC59	IL59	LVR59	320
CC60	IL60	LVR60	321
CC61	IL61	LVR61	322
CC62	IL62	LVR62	323
CC63	IL63	LVR63	335

Supplementary Table 1: List of mouse Cecum, Ileum and Liver RNA samples used for RNAseq analysis.

Supplementary Table 2

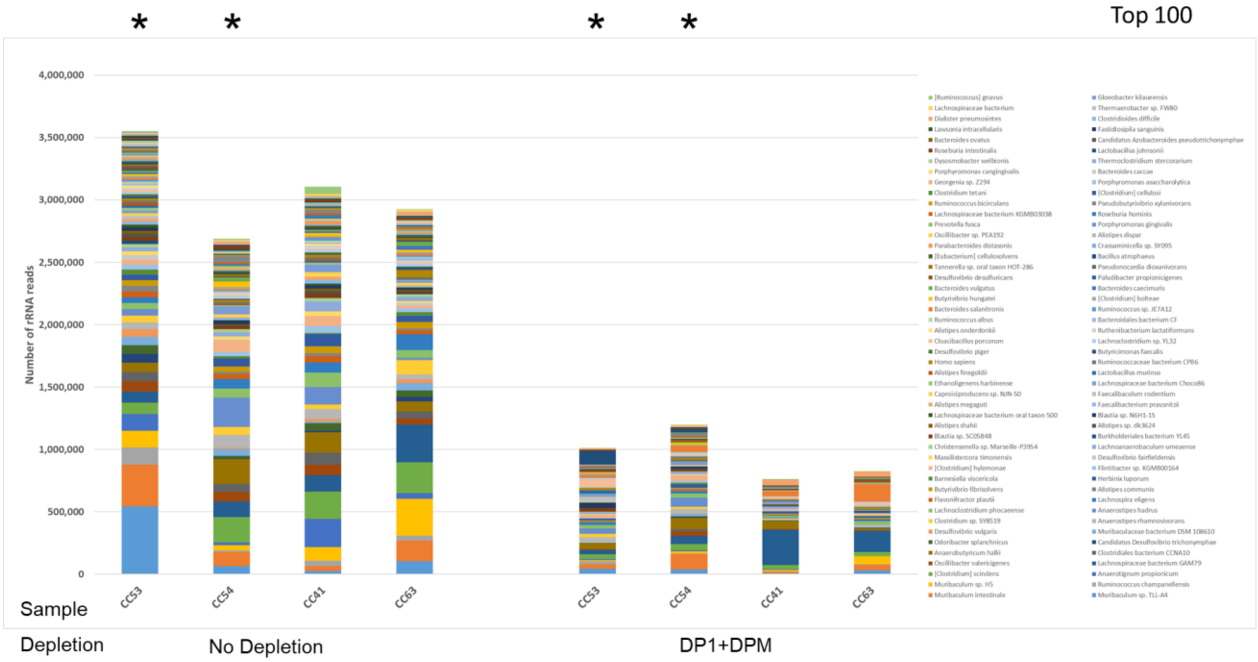


Depletion_A	Depletion_B	Pooled t test prob > |t| (assuming equal variances)	t test prob > |t| (assuming unequal variances)	Result
No Depletion	DP1+DPM	<0.0001	<0.0001	****
No Depletion	DP1+DPM+2025	<0.0001	<0.0001	****
No Depletion	DP1+DPM+5050	<0.0001	<0.0001	****
DP1+DPM	DP1+DPM+2025	0.0061	0.0069	**
DP1+DPM	DP1+DPM+5050	0.0635	0.0640	ns
DP1+DPM+2025	DP1+DPM+5050	0.3347	0.3353	ns

*P<= 0.05, **P<=0.01, ***P<=0.001, ****P<=0.0001

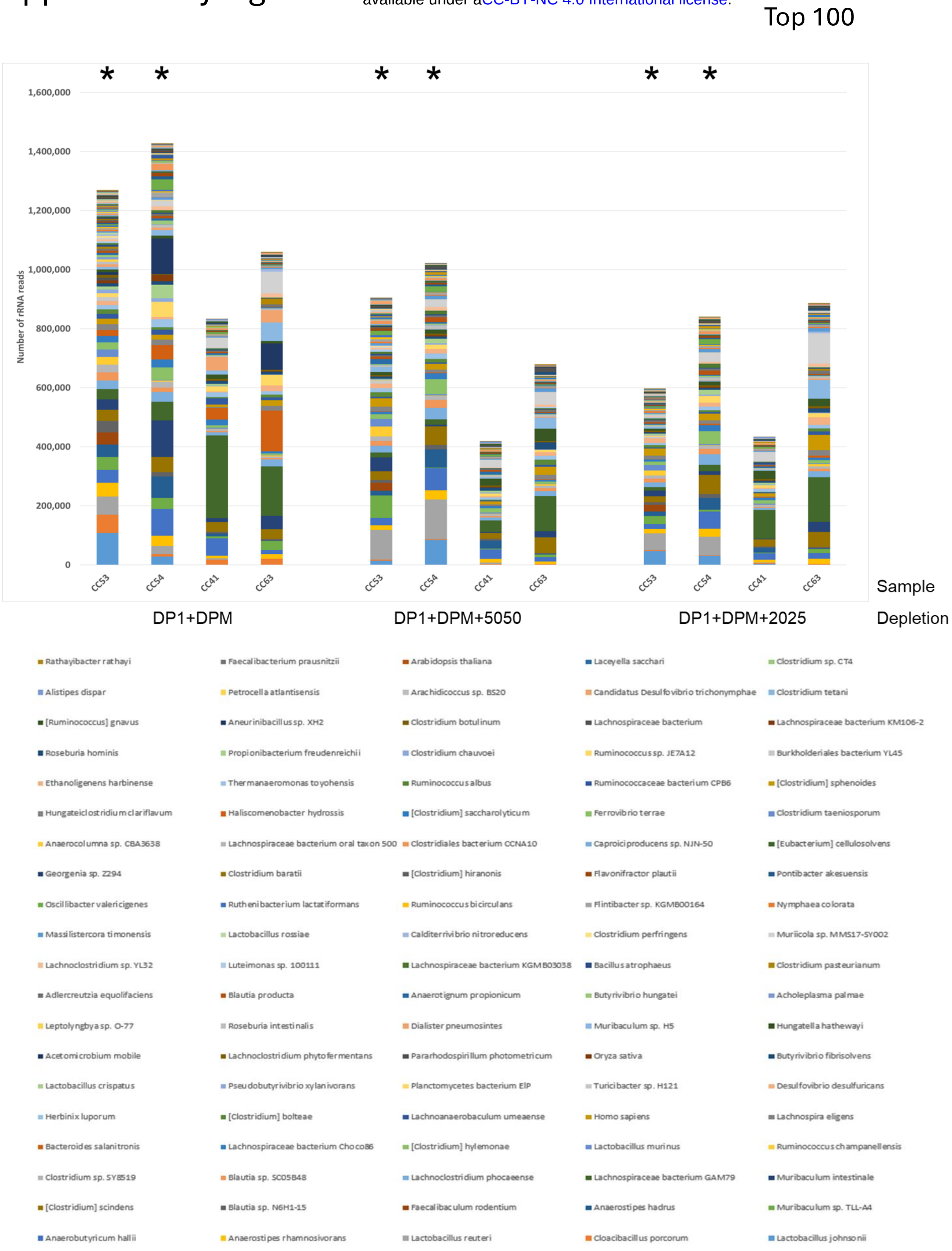
Supplementary Table 2: Results of two-tailed t-test of significance. One way analysis of % retained by depletion method (see Figure 3B). The comparison of DP1+DPM vs DP1+DPM+2025 is significant with a P < 0.01.

Supplementary Figure 1



Supplementary Figure 1: Taxonomic analysis of the collected rRNA reads from the top 100 most abundant species (sorted by sample CC53 – No Depletion) either from undepleted (left side) or samples depleted with DP1 & DPM (right side) indicating that no particular species dominate the remaining rRNA content of the samples following depletion. See Figure 1C for the Top 15 species. Samples CC54, CC53, CC41, CC63. The asterisks at the top indicate two of the samples chosen for probe design in Figure 2.

Supplementary Figure 2



Supplementary Figure 2: Taxonomic analysis of the collected rRNA reads from the top 100 most abundant species (sorted by sample CC53 – DP1+DPM) either from samples depleted with DP1+DPM (left side), DP1+DPM+5050 (middle), or DP1+DPM+2025 (right side) indicating that no particular species dominate the remaining rRNA content of the samples following depletion. See Figure 3C for the Top 15 species. The asterisks at the top indicate two of the samples chosen for probe design in Figure 2.