Quantitative Sequencing and its Application to Studies of the Human Small-Intestine Microbiota

Thesis by

Jacob T. Barlow

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2022
(Defended August 26, 2021)

Jacob T. Barlow
ORCID: 0000-0002-1842-4835

ACKNOWLEDGEMENTS

This thesis could tell a separate story about the incredible people who supported me during each step of the journey. I'd like to express my sincere gratitude for their support, guidance, friendship, and care.

To my advisor, Dr. Rustem Ismagilov, thank you for believing in me from the start. I know your support was instrumental in my acceptance at Caltech and you've always made me feel respected as an individual. You gave me the room to grow and test my own ideas but also provided instrumental support and guidance when needed. I will always respect you for your ability to have constructive discussions even when we completely disagreed. This created an environment where I enjoyed questioning everything and I will take this with me wherever I end up next.

To the rest of my committee members, thank you for your time, critical thoughts, and support. Dr. Sarkis Mazmanian, our many discussions in the first few years at Caltech helped shaped many of my projects. Even when my questions were out of complete curiosity with no relevance to my current projects you took the time to address them. Dr. Matt Thomson, your ability to take complicated data transformation techniques and simplify them have been instrumental in my projects. Your open-door policy provided the perfect environment for me to feel comfortable asking even the simplest of questions. Dr. Long Cai, thank you for being the outside voice on my committee and making me take a step back and truly think about the impact of my work in the context of the broader field.

To other members of the Caltech community that have been instrumental during my PhD: Dr. Justin Bois, Dr. Konstantin Zuev, and Dr. Chace Tydell, your lectures may have been routine to you, but they provided the foundation and knowledge critical to my work. Your enthusiasm and care for students showed through each of your lectures. Dr. Karen Lencioni and the rest of the OLAR staff, thank you for your incredible support and help on all animal experiments. Dr. Lauriane Queenee, thank you for helping keep our lab safe and for your timely responses to our many inquiries.

To the external collaborators I had the pleasure of working with, thank you for helping apply the tools we developed in the lab to ask real questions about human physiology. Dr. Bana Jabri, Dr. Dustin Shaw, Dr. Zach Earley, Dr. Mark Pimentel, Dr. Gabriela Leite, Dr. Elaine Hsiao, Dr. Christine Olson, Dr. Sarkis Mazmanian, and Dr. Wei-Li Wu, I enjoyed each and every one of our collaboration projects. Your work is truly inspiring, and I feel privileged to have played a small role.

To Dr. Aiden Aceves, I want to thank you for being the best of friends since day 1 in data analysis. You and Marlene have been incredible friends and I know that our paths will cross in the future. We both said we would graduate in 4 years and here we are. We did it!

To Dr. Natasha Shelby, you truly are the glue that holds the lab together. You are the most genuinely kind individual I have ever met and your kindness spills over into the lab. You touch each and every project in the lab and never receive the full credit you deserve. I cannot thank you enough for your kind words, friendship, editing skills, organization skills, and motivational support.

To each member of the Ismagilov lab, thank you for being an incredible group of individuals. Each of you has touched this thesis and my life in different ways. Dr. Nathan Schoepp, Dr. Tahmineh Khazaei, Dr Said Bogatyrev, Dr. Asher Preska-Steinberg, and Dr. Justin Rolando, thank you for helping train me in all aspects of the lab and deal with my endless questions. To Emily Savela, you have been the best colleague and friend since day one. Getting placed at the desk next to you was one of the more influential aspects of my PhD. I'm still upset we never got to collaborate on a project together, but I hope you know that you have had an impact on each project in this thesis. To Michael Porter, I enjoyed commiserating with you over each failed microbiome project. We may not have much to show for it, but I know we were close to something. To Anna Romano, I can't thank you enough for the friendship and support since you joined the lab. We accomplished some of the most ridiculous time crunch experiments and without your support I know this work would not be completed. To Matt Cooper, I'm glad I was able to transition you to the dark side of the lab. Thank you for help pushing the limits of the methodologies in this thesis and being a great friend. To Natalie Wu-Woods, I cannot thank you enough for the incredible work you've put in since you've joined the lab. Without your hard work and dedication there is no way the work in this thesis would have been completed. To all other members of the Ismagilov lab, you have each played a role in this work and my PhD journey. I wish you all the best and hope that our paths continue to cross in the future.

To my family, your unending love and support for my passions have been clear my entire life. You have given me the confidence to try without the fear of failure. I want to express my deepest gratitude for all that you have provided for me. I have enjoyed my time in California but part of me has always yearned to be back home close to your love and support.

I've saved the most influential person in my life for last. To my loving wife, Tori Barlow, you are the reason all the efforts and struggles of these last 4 years are worth it. The thoughts of our future together helped push me through every failed experiment, late night, and frustrating reviewer. You both helped keep my ego in check and remind me of what is truly important in life. You pushed us to explore California to its fullest while we were here and helped me detach from the lab when I needed it most. Some of the clearest thoughts of my PhD have come after one of our many trips where we forgot about work and just enjoyed each other's company. You bring unfathomable joy to my life and there is no one else I would have rather had by my side during this journey. Thank you for everything!

ABSTRACT

Our understanding of the interplay between microbial species and the hosts they live on and in is continually expanding. New insights have focused not only microorganisms that drive specific disease states but also those that help maintain human health. As research drives towards mechanistic understanding of host-microbe relationships new quantitative tools are needed to help interrogate these complex interactions. Chapter I of this thesis discusses formulation of a method for rapid detection of antibiotic resistance in *Neisseria gonorrhoeae*. Our approach identified RNA signatures from transcriptional profiling of Neisseria gonorrhoeae after 10-minute antibiotic exposure. Utilization of these RNA markers allowed for rapid identification of antibiotic susceptibility or resistance to the antibiotic ciprofloxacin. Chapter II shifts focus to the development of a quantitative sequencing technique for the measurement of absolute taxon abundances in complex microbial communities. Combining the precision of digital PCR with the high-throughput nature of 16S rRNA gene amplicon sequencing allowed for simultaneous quantitative profiling of all bacterial taxa in host-associated microbial communities. We extensively characterized our quantitative sequencing methodology in the presence of high host nucleic acid levels and low microbial loads to understand the limits of quantification and detection in complex sample types. Last, Chapter III applies the quantitative sequencing technology from Chapter II to investigate the microbial community of the human small intestine, specifically the duodenum. Data from the duodenum of 250 individuals revealed a wide range of total microbial loads and a distinct subset of microbes, termed disruptor taxa, that were associated with small intestinal bacterial overgrowth (SIBO) and GI symptom severity.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Khazaei T., **Barlow J.,** Schoepp N., and Ismagilov R. "RNA markers enable phenotypic test of antibiotic susceptibility in Neisseria gonorrhoeae after 10 minutes of ciprofloxacin exposure." *Scientific Reports.* (2018) 8:11606. doi:10.1038/s41598-018-29707-w.

   Author contributions:

   Tahmineh Khazaei: Co-designed the study with NGS. Developed the computational pipeline for processing and analyzing RNA sequencing data and for selection of RNA markers. Using this pipeline, identified porB and rpmB as the top markers for this study. Established the RNA extraction protocol for *Neisseria gonorrhoeae* samples (e.g. best kit/protocol to use). Performed RNA extraction of *Neisseria gonorrhoeae* samples from all the AST experiments in this study for RNA sequencing and performed the quality assessment of the extracted RNA. (This step was after the initial AST exposures performed by NGS or JTB). Wrote the manuscript and generated all the final figures for publication.

   Jacob T. Barlow: Worked side-by-side with TK on day-to-day experimental optimization of RNA AST pipeline on CDC strains, including choice of using 16S rRNA as a control marker. Generated glycerol stocks used for the 50 CDC strains using NGS's protocol. Set up all cultures and ran all antibiotic exposures for the 50 CC strains using NGS's protocols. Ran all qPCR and dPCR experiments for assessing changes in RNA markers. Designed and implemented statistical thresholding method for determination of differential genes. Used thresholding method to choose 4 additional markers tested. Designed and optimized primers for the additional markers. Generated visualization strategy for plots in figures 2, 3, 4, and 5 before handing off to TK to prepare final versions for paper.

   Nathan G. Schoepp: Designed "high throughput" sample handling and exposure workflow for isolates used by JTB and TK. Obtained initial set of *Neisseria gonorrhoeae* isolates from UCLA. Established *Neisseria gonorrhoeae* culturing and quantification methods which included 1) selecting and screening medias (including the one ultimately used in AST experiments) and 2) selecting and testing primers from literature for specificity, speed, and LOD. Performed initial AST exposures using *Neisseria gonorrhoeae* isolates, which TK then extracted and sequenced. Designed primers used in final manuscript for rpmB and porB markers. Assisted JTB in primer design for other markers by demonstrating primer design workflow and tools. Made minor contributions to manuscript including minor edits, and providing TK with graphics used in Fig. 1.

2. **Barlow J.,** Bogatyrev S., and Ismagilov R. "A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities." *Nature communications* 11.1 (2020): 1-13**.** doi:10.1038/s41467-020-16224-6

   Author contributions:

   Jacob T. Barlow: Validated limits of digital PCR assay with mock microbial communities in germ-free tissues. Designed, performed, and analyzed experiments to validate accuracy of quantitative sequencing with dPCR anchoring. Established the quantitative limits of an individual taxon's absolute abundance. Conducted the ketogenic animal study. Analyzed all data; created all figures; and wrote the paper.

   Said R. Bogatyrev: Co-developed the idea of quantitative sequencing with dPCR anchoring for absolute quantification of total microbial loads and taxa absolute abundances in lumenal and mucosal samples. Contributed the method for quantitative sequencing with dPCR anchoring in lumenal and mucosal samples. Contributed ideas and provided support for animal study design. Contributed ideas for data analysis and representation.

3. **Barlow J.*,** Leite G.*, Romano A., Sedighi R., Chang C., Celly S., Rezaie A., Mathur R., Pimentel M., and Ismagilov R. "Quantitative sequencing clarifies the role of disruptor taxa, oral microbiota, and strict anaerobes in the human small-intestine microbiome." *Microbiome;* doi:10.1186/s40168-021-01162-2.

   *Authors contributed equally to this work

   Author contributions:

   Jacob T. Barlow: Major contributor on idea of applying absolute taxon load via quant seq to small intestinal aspirate samples. Minor contributor to selection of samples during study design. Major contributor to library prep for quant-seq. Performed all digital PCR experiments, developed and performed analysis for all figures, generated all figures, wrote original draft of paper, led revisions for final paper, and managed all data and code archiving in public repositories.

   Gabriela Leite: Major contributor on idea of applying absolute taxon load via quant-seq to small-intestinal aspirate samples. Major contributor to selection of samples during study design. Major contributor to patient data curation. Developed method for DNA extraction from duodenal aspirates. Minor contributor to DNA extraction. Reviewed and edited original draft of paper.

   Anna E. Romano: Major contributor to library prep for quant-seq. Reviewed and edited original draft of paper.

Rashin Sedighi: Major contributor to patient recruitment and blood/saliva collection.

Christine Chang: Major contributor to patient recruitment and blood/saliva collection.

Shreya Celly: Performed all duodenal bacterial cultures and sample processing prior to DNA extraction. Major contributor to DNA extraction.

Ali Rezaie: Supervision of patient data curation. Major contributor to patient upper endoscopy procedure and sample collection from duodenum. Reviewed and edited original draft of paper.

Ruchi Mathur: Supervision of patient recruitment. Reviewed and edited original draft of paper.

Mark Pimentel: Supervision of study. Major contributor to patient upper endoscopy procedure and sample collection from duodenum. Reviewed and edited original draft of paper.

TABLE OF CONTENTS

C h a p t e r I

RNA MARKERS ENABLE PHENOTYPIC TEST OF ANTIBIOTIC SUSCEPTIBILITY IN NEISSERIA GONORRHOEAE AFTER 10 MINUTES OF CIPROFLOXACIN EXPOSURE

This chapter was originally published in "Khazaei T., **Barlow J.,** Schoepp N., and Ismagilov R. "RNA markers enable phenotypic test of antibiotic susceptibility in Neisseria gonorrhoeae after 10 minutes of ciprofloxacin exposure." *Scientific Reports.* (2018) 8:11606. doi:10.1038/s41598-018-29707-w."

Abstract

Antimicrobial-resistant *Neisseria gonorrhoeae* is an urgent public-health threat with continued worldwide incidents of infection and rising resistance to antimicrobials. Traditional culture-based methods for antibiotic susceptibility testing are unacceptably slow (1–2 days), resulting in the use of broad-spectrum antibiotics and the further development and spread of resistance. Critically needed is a rapid antibiotic susceptibility test (AST) that can guide treatment at the point-of-care. Rapid phenotypic approaches using quantification of DNA have been demonstrated for fast-growing organisms (e.g. *E. coli*) but are challenging for slower-growing pathogens such as *N. gonorrhoeae*. Here, we investigate the potential of RNA signatures to provide phenotypic responses to antibiotics in *N. gonorrhoeae* that are faster and greater in magnitude compared with DNA. Using RNA sequencing, we identified antibiotic-responsive transcripts. Significant shifts (>4-fold change) in transcript levels occurred within 5 min of antibiotic exposure. We designed assays for responsive transcripts with the highest abundances and fold changes, and validated gene expression using digital PCR. Using the top two markers (*porB* and *rpmB*) we correctly determined the antibiotic susceptibility and resistance of 49 clinical isolates after 10 min exposure to ciprofloxacin. RNA signatures are therefore promising as an approach on which to build rapid AST devices for *N. gonorrhoeae* at the point-of-care, which is critical for disease management, surveillance, and antibiotic stewardship efforts.

Introduction

Neisseria gonorrhoeae is the second most common sexually transmitted bacterial infection in the United States, with about 460,000 cases reported in 2016, an 18.5% rise since 2015Prevention[1]. Worldwide, it is estimated that about 78 million new *N. gonorrhoeae* infections occur annually[2]. *N. gonorrhoeae* infections can lead to heart and nervous system infections, infertility, ectopic pregnancies, newborn blindness, and increased risk for other sexually transmitted infections, including HIV[3]. The CDC has identified *N. gonorrhoeae* as one of the three most urgent drug-resistant bacterial threats[3]. *N. gonorrhoeae* has developed resistance to all of the most commonly used antibiotics (including penicillins, sulfonamides, tetracyclines, and fluoroquinolones) leaving only one last effective class of antibiotics, cephalosporins. However, there have even been worldwide reported cases of decreased susceptibility to the cephalosporin ceftriaxone[4-8], and therefore an imminent threat of widespread untreatable *N. gonorrhoeae*. An important factor leading to the widespread development of antibiotic resistance is the liberal use and misuse of antibiotics. Critically needed is a rapid antibiotic susceptibility test (AST) that can guide treatment at the point-of-care – both to provide correct treatment and to facilitate antibiotic stewardship.

The gold standard for determining *N. gonorrhoeae* susceptibility to antibiotics is the culture-based agar dilution test, which is unacceptably slow (1–2 days). More rapid genotypic approaches, involving detection of gene mutations, are available for a subset of antibiotics in *N. gonorrhoeae*[9,10], but such approaches are inherently limiting, as they require knowledge of the mechanisms of resistance. Moreover, *N. gonorrhoeae* is naturally competent for transformation, and can take up gonococcal DNA from the environment and recombine it with its own genome, resulting in frequent gene mutations[11,12]. Given the high rate at which new resistance emerges, relying solely on genotypic methods is not an acceptable long-term solution. Phenotypic methods involving growth measurements have enabled faster ASTs that are independent of resistance mechanisms[13-16]. However, such growth-based methods are challenging for *N. gonorrhoeae*, which is slow-growing and fastidious[17]. Another phenotypic

approach for antibiotic susceptibility testing is quantification of nucleic acids[18,19]. We have previously demonstrated a rapid (30 min) phenotypic AST using quantification of DNA replication by digital PCR (dPCR) to assess the antibiotic susceptibility of *Escherichia coli* in clinical urine samples[20]. However, AST methods that quantify changes in DNA replication require a longer antibiotic-exposure step for slow-growing pathogens such as *N. gonorrhoeae,* which has a doubling time of about 60 min[21], compared with the 20 min doubling time of *E. coli*[22].

A complementary approach to DNA quantification is measuring the pathogen's RNA response to antibiotic exposure. Transcriptional responses are among the earliest cellular changes upon exposure to antibiotics[23], far before phenotypic changes in growth can be observed. Quantifying changes in RNA signatures is therefore a particularly appealing approach for slow-growing organisms. RNA has previously been used to differentiate antibiotic susceptibility and resistance in organisms where the transcriptional response is well characterized[24,25]. More recently, RNA sequencing (RNA-Seq) has been used to measure the transcriptome response of *Klebsiella pneumoniae* and *Acinetobacter baumanii* to antibiotic exposure[25]. Although the *N. gonorrhoeae* transcriptome has been previously sequenced[26,27], to our knowledge, no one has characterized the transcriptome response of *N. gonorrhoeae* to antibiotic exposure. Unlike most bacteria, *N. gonorrhoeae* lacks the classic transcriptional SOS response to DNA damage whereby DNA repair is induced and the cell cycle is arrested[28,29]. The SOS response promotes survival to certain antibiotic classes, such as the fluoroquinolones, which act by directly inhibiting DNA synthesis[30]. The recA or recA-like proteins are essential for the induction of the SOS response[28]. However, neither *recA* transcripts nor recA protein levels increase in *N. gonorrhoeae* upon exposure to DNA damaging agents[31,32].

In this work, we explore the transcriptome response of *N. gonorrhoeae* upon exposure to ciprofloxacin. Ciprofloxacin is a fluoroquinolone and functions by inhibiting the enzymes topoisomerase II (DNA gyrase) and topoisomerase IV, thereby inhibiting cell division[33]. Ciprofloxacin was chosen in this study to gain insight into transcriptional changes that occur

upon DNA damage in an organism lacking the classic SOS response. Here, we address the following questions: (1) How does the transcriptome of *N. gonorrhoeae* respond to ciprofloxacin exposure? (2) What is the shortest antibiotic exposure time at which we can still observe significant changes (>4-fold) in RNA expression? (3) Which transcripts provide the largest and most abundant fold-changes per cell, which is an important consideration for clinical samples that have low numbers of pathogens? (4) Will candidate markers respond consistently across a large pool of isolates with wide genetic variability?

Results

We used RNA-seq to study the transcriptome response of susceptible and resistant isolates of *N. gonorrhoeae* after 5, 10, and 15 min of ciprofloxacin exposure (Fig. 1.1). Each clinical isolate was initially split into two tubes, where one tube was exposed to the antibiotic (+) and the other served as the control with no antibiotic exposure (-). Samples were collected for RNA-seq prior to antibiotic exposure and every 5 min for 15 min. We calculated the fold change in gene expression between the control and treated samples – defined as the control:treated ratio (C:T ratio); genes that demonstrated significant fold-change differences between the susceptible and resistant isolates were identified as differentially expressed. To account for biological variability, three pairs of susceptible and resistant isolates were used in this study. Candidate markers were selected from the pool of differentially expressed genes and were validated using droplet dPCR (see Methods).

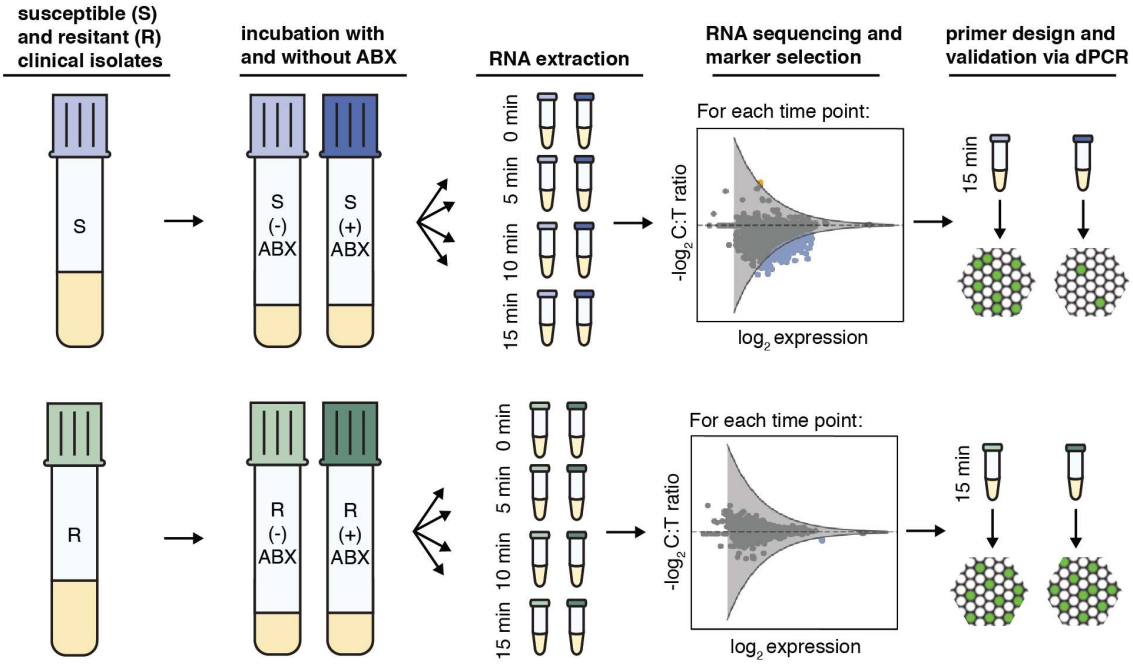


Figure 1.1. The workflow for selection and validation of RNA markers for phenotypic measurements of antibiotic susceptibility and resistance. Susceptible and resistant isolates of *Neisseria gonorrhoeae* are exposed to antibiotics (ABX) for 5, 10, and 15 min. Samples are collected for RNA sequencing at time zero and every 5 min thereafter. Genes demonstrating fold changes in expression (control:treated ratio (C:T ratio)) greater than the threshold of significance (gray line) are identified as differentially expressed (blue: downregulated and orange: upregulated). Candidate markers are selected from the pool of differentially expressed genes and validated by digital PCR.

Temporal shifts in global gene expression upon antibiotic exposure

We observed global shifts in RNA expression in susceptible isolates in as early as 5 min after antibiotic exposure (Fig. 1.2a). The distribution of fold changes in gene expression levels (C:T ratios) indicated global shifts toward negative $\log_2$ fold-change values (downregulation). The magnitude of fold change at which most genes were distributed was approximately 2-fold. The tail of the distribution illustrates that a few genes responded to antibiotic exposure with changes as large as 6-fold within 5 min. Increasing the antibiotic exposure time further shifted the distribution to larger negative $\log_2$ fold-change values. The transcriptional response in resistant isolates was tightly distributed around a fold-change

value of 1 at every time point, indicating that the transcriptome of the resistant isolates did not respond significantly to antibiotic exposure (Fig. 1.2a).

To identify genes that were differentially expressed between control and treated samples, we defined a threshold of significance (Fig. 1.2b). The threshold of significance took into account technical variability and was calculated from the C:T ratios at t = 0 min of all biological replicates that were sequenced (three susceptible and three resistant isolates). For each of the six gene expression datasets (one for each isolate), we plotted the $-\log_2$(C:T ratio) against the $-\log_2$(expression) for all genes and fit a negative exponential curve to the outer edge of each plot. We then averaged the curves from all six datasets and added a 90% confidence interval to the average curve by assuming a Gaussian fit for the error distribution, which we define as our threshold of significance. Genes with a $-\log_2$(C:T ratio) value above or below the upper and lower thresholds were identified as differentially expressed. Downregulated genes (fold changes below the significance threshold) appeared as early as 5 min after antibiotic exposure (blue dots, Fig. 1.2b). Two upregulated genes (fold changes above the significance threshold) appeared after 10 min of exposure (orange dots, Fig. 1.2b).

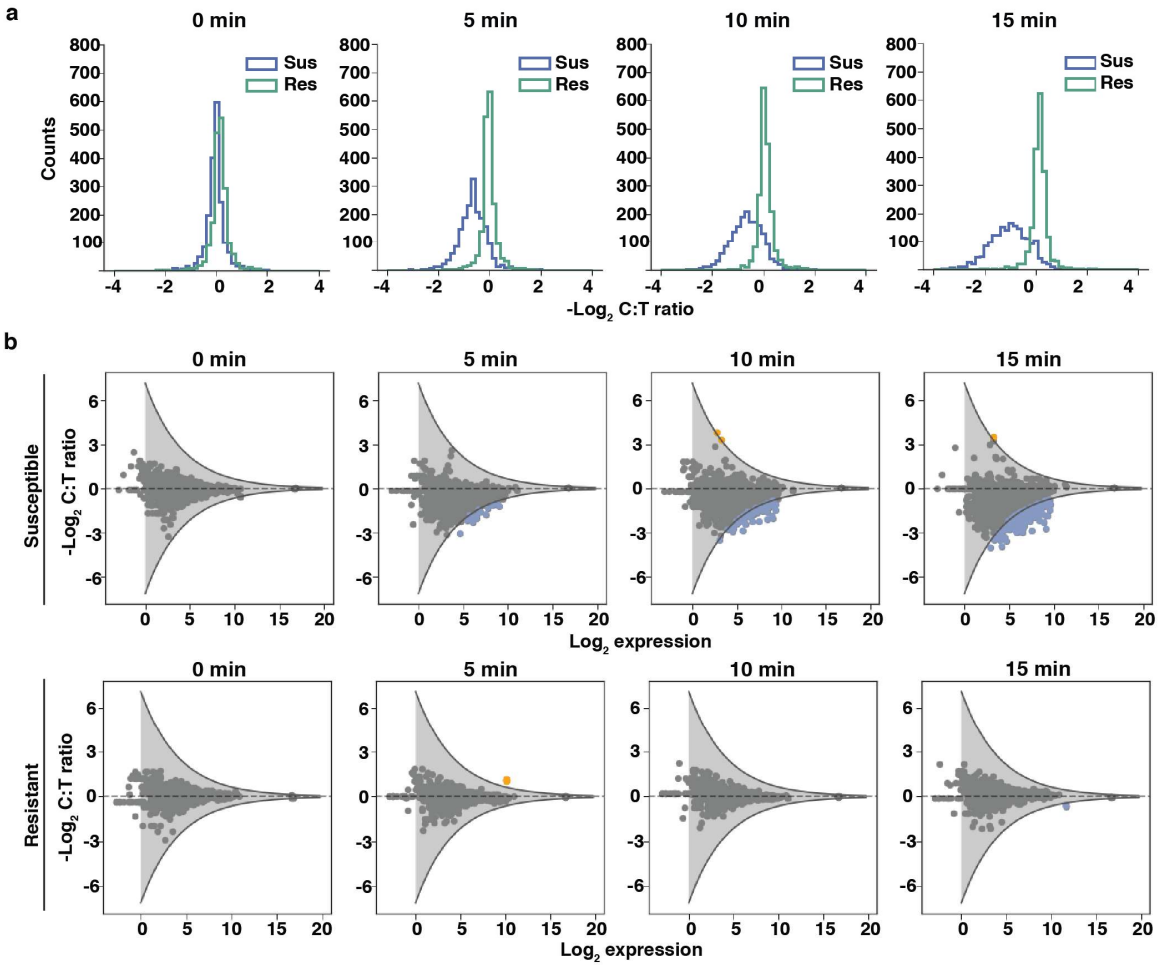


Figure 1.2. Temporal shifts in global gene expression upon ciprofloxacin exposure in *Neisseria gonorrhoeae*. (**a**) The distribution of -log$_2$(C:T ratios) for a susceptible isolate (Sus) and resistant isolate (Res) at 0, 5, 10, and 15 min. (**b**) The fold change in gene expression between control and treated samples (C:T ratio) versus expression in the control sample at 0, 5, 10, and 15 min for one susceptible isolate and one resistant isolate. Genes with C:T ratios above or below the significance threshold are identified as differentially expressed (blue: downregulated; orange: upregulated). Thresholds for statistical significance of fold change (gray lines) are determined by fitting a negative exponential curve (with 90% confidence interval) to the outer edge of the -log$_2$ C:T ratios measured at time zero (see Methods).

Selection of candidate markers that are consistent in response and abundant

RNA expression in response to antibiotics can be heterogeneous among different isolates of the same species[34]; thus, it is important to select candidate markers from differentially expressed genes that respond consistently across isolates of *N. gonorrhoeae.* To identify these candidate markers, we exposed three different pairs of susceptible isolates (minimum inhibitory concentrations (MICs) <= 0.015mg/mL) and resistant isolates (MICs 2.0 mg/mL, 4.0mg/mL, and 16.0mg/mL) to ciprofloxacin for 15 min and extracted RNA for sequencing (see workflow in Fig. 1.1). We found 181, 41, and 410 differentially expressed genes in susceptible isolates 1, 2, and 3, respectively (Fig. 1.3a). Among the differentially expressed genes, 38 genes responded consistently across the three pairs of susceptible and resistant isolates (i.e. responses overlapped in all three susceptible isolates, whereas all three resistant isolates were non-responsive) (Supplementary Table S2.1 online). These genes spanned a variety of biochemical functions in the cell. We selected six candidate transcript markers for further analysis based on the following criteria: (1) high fold change; (2) high expression levels (>75 transcripts per million, TPM); and (3) representative of different biochemical pathways. The selected candidate markers were: *porB* (membrane protein), *rpmB* (ribosomal protein), *tig* (molecular chaperone), *yebC* (transcriptional regulator), *pilB* (pilus assembly ATPase), and *cysK* (cysteine synthase). The candidate marker with the highest abundance and largest fold change upon antibiotic exposure was *porB*, which is a membrane channel forming protein and the site of antibiotic influx into the cell[35].

A high level of gene expression was one of our criteria for selection of candidate markers from the sequencing data. High expression of candidate markers is not only important for sensitivity and limits of detection, but is particularly important for clinical samples with low numbers of pathogen cells. One of the advantages of RNA compared with DNA as a nucleic acid marker is its natural abundance in the cell. Because the gene expression values obtained from sequencing are relative values, our next step was to quantify the absolute

copies per cell for the candidate markers. In our quantification approach, we plated clinical isolate samples after 15 min of ciprofloxacin exposure to obtain cell numbers in colony forming units (CFU/mL). We designed primers for the candidate markers (see Methods and Supplementary Table S1.2) and measured their absolute concentration using dPCR. The concentrations were converted to per cell values using the cell counts from plating (Fig. 1.3b). Additionally, we used the RNA sequencing data to obtain transcriptome-wide estimates of transcript copies per cell. In the sequencing approach, we added external RNA control consortium (ERCC) spike-ins to the lysis buffer step of the extraction protocol in order to capture any loss of RNA throughout the extraction steps. By linear regression, we captured the relationship between ERCC copies added to the samples and ERCC quantified by sequencing. Using the linear regression, we converted gene expression values from RNA sequencing (in TPM) to approximate copy numbers per cell (see Methods). The transcript copies per cell estimated for the candidate markers using the sequencing approach were within the same order of magnitude as the absolute copies per cell measured by digital PCR (Fig. 1.3b).

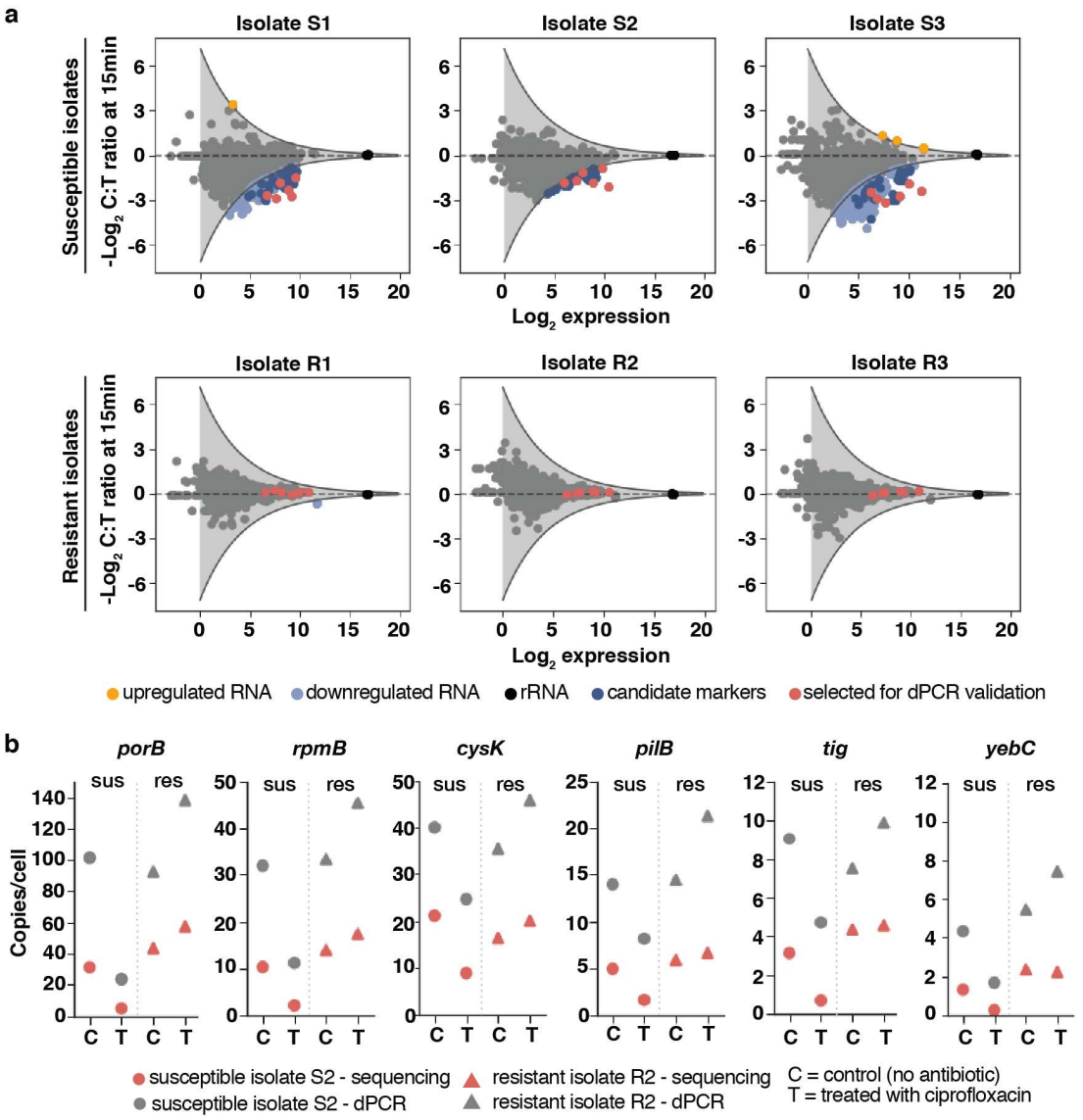


Figure 1.3. Selection of candidate RNA markers for phenotypic antibiotic susceptibility testing in *Neisseria gonorrhoeae* and measurements of candidate marker abundances per cell (**a**) Genes that are differentially expressed (light blue) across three pairs of resistant and susceptible clinical isolates are identified as candidate markers (dark blue). Six candidate markers that span different biological functions were selected for validation (red). (**b**) Copies/cell values for the candidate markers are determined from RNA sequencing (red) and dPCR (gray) (see Methods). Data is shown for one pair of susceptible (S2) and resistant (R2) isolates at 15 min of ciprofloxacin exposure.

Validation of candidate markers by dPCR

We next asked how the relative changes observed through RNA-seq compare with direct gene expression measurements by dPCR. We designed dPCR assays for candidate markers, which involved measuring the absolute expression of the candidate marker in both control and treated samples, and calculating the C:T ratio. In this assay, the 16S rRNA was also measured and used to normalize the C:T ratio of the candidate markers. In the three susceptible isolates that were sequenced we found that rRNA consistently showed the smallest fold change (<1.06) in response to ciprofloxacin compared with all other genes in *N. gonorrhoeae*. Therefore, to account for experimental variations in the antibiotic exposure and RNA extraction steps between control and treated samples, we used the 16S rRNA as an intracellular control for normalizing the C:T ratios (see Methods). We found that the C:T ratios measured by the dPCR assay agreed with the C:T ratios obtained through sequencing (Fig. 1.4), confirming that both approaches accurately capture the transcriptional response to antibiotic exposure.

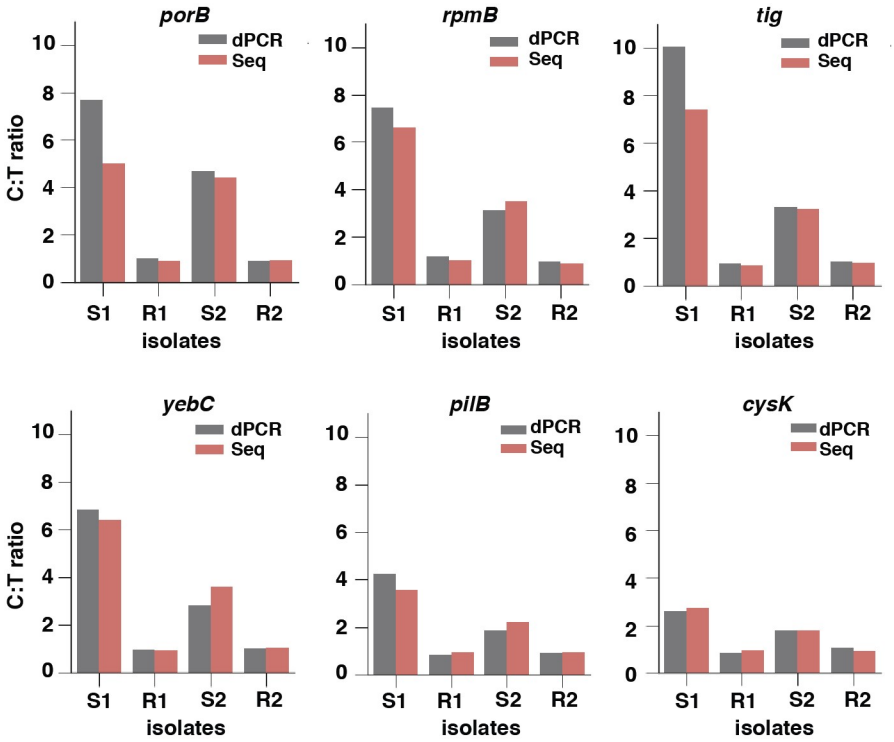


Figure 1.4. Validation of the RNA sequencing approach using digital PCR (dPCR) with six candidate markers. Control:treated ratios (C:T ratios) determined by RNA sequencing (red) were validated against C:T ratios measured by dPCR (gray). The dPCR C:T ratios were normalized using ribosomal RNA (rRNA) by dividing the C:T ratio of the candidate marker by the C:T ratio of 16S rRNA. This normalization step is not required for sequencing data because sequencing depth normalizes the values (see Methods). Markers were validated using two susceptible (S1 and S2) and two resistant (R1 and R2) isolates at 15 min of ciprofloxacin exposure.

Validation of RNA markers across CDC isolates

Finally, we asked whether candidate markers respond consistently across a large pool of isolates with genetic variability. We chose the two candidate markers with the highest abundances and fold changes (*porB* and *rpmB*) to determine the susceptibility of 49 clinical isolates, with a wide range of MIC values (Supplementary Table S1.3 online), from the *N. gonorrhoeae* panel of the Centers for Disease Control (CDC) Antimicrobial Resistance Isolate Bank. The MIC values were representative of the population-wide distribution values reported by the European Committee on Antimicrobial Susceptibility Testing[36]. We exposed each clinical isolate to ciprofloxacin for 10 min and measured the fold change in expression of the two candidate markers between the control and treated sample using dPCR (Fig. 1.5). Both markers correctly classified all 49 CDC isolates, based on Clinical and Laboratory Standards Institute (CLSI) breakpoint values, as 9 susceptible and 40 resistant strains.

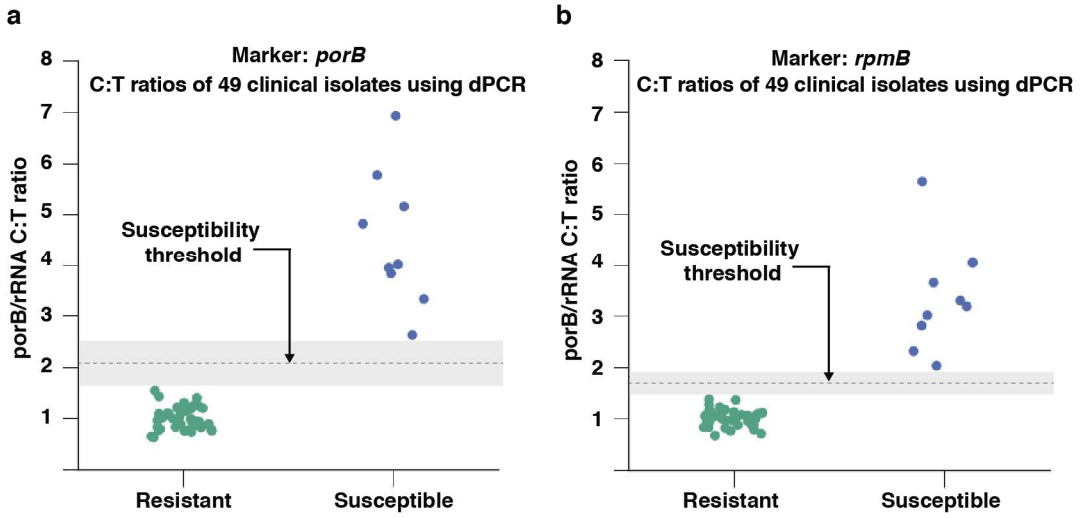


Figure 1.5. Antibiotic susceptibility testing of 49 clinical isolates using (a) *porB* and (b) *rpmB* as RNA AST markers. Antibiotic susceptibility of 49 clinical isolates (9 susceptible and 40 resistant) from the *Neisseria gonorrhoeae* panel of the CDC bacteria bank was determined using the "normalized" C:T ratios (C:T ratio of marker/C:T ratio of 16S rRNA). Clinical isolates were exposed to ciprofloxacin for 10 min and the concentration of RNA markers was measured by digital PCR.

Discussion

In this work, we demonstrate that antibiotic-responsive transcripts can be used as suitable markers for a rapid phenotypic AST in *N. gonorrhoeae.*

When characterizing the global transcriptional response of *N. gonorrhoeae* to antibiotic exposure, we observed a significant change in response in as early as 5 min. The nature of the response was a global downregulation in transcript levels. Among the candidate markers, all exhibited downregulation in response to ciprofloxacin. We specifically looked at *gyrA* and *parC*, which are known genotypic markers of resistance to ciprofloxacin, and differential expression was not observed. We also looked at the *recA* transcript because *recA* is one of the prominent genes in the SOS response, and as expected, because *N. gonorrhoeae* does not have a true SOS system[28,29], we did not find *recA* levels to increase. Whereas *recA* is a specific cellular response to overcome DNA damage, the global downregulation that we observed suggests a general shift away from growth and cell proliferation. Among the 38 candidate markers, 15 were ribosomal proteins (including one of the top markers, *rpmB*), which play a prominent role in assembly and function of the ribosomes and are essential for cell growth. Mutations in ribosomal proteins have been reported to confer resistance to different classes of antibiotics[37]. Another top marker identified in this study was *porB*, which is a membrane channel forming protein (porin) responsible for uptake of small nutrients and the site of antibiotic influx into the cell. The expression of porins is highly regulated in response to environmental stimuli[38]. Reducing permeability to decrease intracellular antibiotic concentration is a known mechanism for bacteria to confer antibiotic resistance[37]. The downregulation of *porB* observed in this study can be attributed to a halt in growth

processes caused by ciprofloxacin damage and possibly an attempt to reduce influx of antibiotic.

A key aim of this study was to identify RNA markers that would yield a measurable response after only a short antibiotic exposure (<15 min) to ensure this approach can fit within the required timescale for a rapid AST. It is possible that longer exposure times could provide additional insight into the biological response of *N. gonorrhoeae* to ciprofloxacin, but this was not the focus of our study. Furthermore, the short exposure times potentially introduce a bias in selection of transcripts present at low abundances. For transcripts present at high abundance to display the same fold change as low abundance transcripts, a substantially higher number of mRNA molecules must be transcribed, which would require longer timescales. As an example, a 4-fold change from 1 to 4 transcripts requires 3 additional mRNA to be produced, whereas a 4-fold change from 20 to 80 requires 60 mRNA to be transcribed. This bias also holds true in downregulation, where mRNA continues to be transcribed in the control samples, whereas transcript levels drop in treated samples due to degradation of RNA, and/or a reduction in the rate of transcription.

We identified candidate markers with consistent differential expression across three sets of susceptible and resistant pairs. Among the candidate markers, one of our criteria for selection was transcript abundance, which is of particular importance in clinical samples with low cell numbers. Furthermore, marker abundance affects measurement sensitivity and limits of detection, as has been previously demonstrated in AST methods based on quantification of DNA replication[20]. To measure the abundance of the candidate markers, we used both dPCR measurements and ERCC spike-ins for RNA sequencing to obtain approximate RNA copies/cell. Both methods yielded results within the same order of magnitude. To our knowledge, this is the first quantitative measurement of RNA abundance per cell in *N. gonorrhoeae*.

We separately validated the performance of the two most abundant candidate markers, *porB* and *rpmB*, with 49 clinical isolates. Both markers were consistent in their ability to correctly determine susceptibility or resistance of all 49 clinical isolates. *porB* demonstrated C:T ratios

between 2.5 to 7 and *rpmB* demonstrated C:T ratios between 2 and 6 after 10 min of antibiotic exposure in the nine susceptible clinical isolates. The large fold changes highlight the significance of using RNA response as an AST marker compared with quantification of DNA replication. Our previous work using dPCR quantification of DNA replication demonstrated C:T ratios between 1.2 and 2.4 for 15 min of antibiotic exposure in *E. coli*[20], which has a doubling time approximately 3 times shorter than *N. gonorrhoeae*.

We performed an alignment search of *porB* against other prokaryotes and found it to be specific to the *Neisseria* genus. AST markers should be specific to the pathogen of interest because additional bacterial species are likely to be present in clinical samples. Additional experiments with mixtures of bacteria would be required to further confirm the specificity of the markers identified in this study. We additionally measured the 16s rRNA to normalize C:T ratios, which inherently enables pathogen identification as well. A combination of identification and susceptibility testing in a single integrated platform is important for correct and rapid diagnosis.

This paper demonstrates that RNA markers can be used to determine antibiotic susceptibility of *N. gonorrhoeae* after short antibiotic exposure times, a requirement for a rapid phenotypic AST. *N. gonorrhoeae* is a fastidious slow-growing organism, presenting challenges to growth-based AST methods. Additional work will be needed to yield a clinic-ready, rapid RNA-based AST for *N. gonorrhoeae.* Additional background matrices of clinical samples, both urine and swab samples, that could possibly affect speed and sensitivity of an AST, must be further evaluated. Digital isothermal chemistries, such as digital loop-mediated isothermal amplification (dLAMP) should be considered to speed up quantification times relevant to point-of-care settings[20]. Follow-up studies should also examine the transcriptional response of *N. gonorrhoeae* to other classes of antibiotics and identify responsive RNA markers for class-specific antibiotics. Overall, as a first step, the work described here demonstrates the promise for a phenotypic RNA-based approach for a rapid AST of *N. gonorrhoeae* at the point-of-care, which is critically needed for disease management, surveillance, and antibiotic stewardship.

Methods

Antibiotic exposure for RNA sequencing

Antibiotic susceptible and resistant clinical isolates were obtained from the University of California, Los Angeles, Clinical Microbiology Laboratory. Isolates were plated from glycerol stocks onto Chocolate Agar plates and grown in static incubation overnight (37 °C, 5% CO_2). Cells were re-suspended in Hardy Fastidious Broth (HFB) and incubated for 45 min (37 °C, 5% CO_2) with shaking (800 rpm) to an OD_{600} between 1 and 5. Cultures were diluted (5X) into HFB. Each isolate culture was split into "treated" and "control" tubes. Ciprofloxacin was added to the "treated" tubes (final concentration of 0.5 µg/mL) and water was added to the "control" tubes; cultures were incubated (static; 37 °C, 5% CO_2) for 15 min. During incubation, samples were collected for RNA sequencing at 5, 10, and 15 min (300 µL aliquot of sample was mixed into 600 µL of Qiagen RNA Protect Reagent (Qiagen, Hilden, Germany) for immediate RNA stabilization). In addition, a sample was collected for RNA sequencing immediately before ciprofloxacin was added. To quantify CFU, the sample at t = 15 min was serially diluted (10x), plated on a Chocolate Agar plate, and incubated overnight (37 °C, 5% CO_2).

Antibiotic exposure for clinical isolates

Antibiotic susceptible and resistant clinical isolates were obtained from the *N. gonorrhoeae* panel of the CDC Antimicrobial Resistance Isolate Bank. Isolates were plated from glycerol stocks onto Chocolate Agar plates and grown in static incubation overnight (37 °C, 5% CO_2). Cells were re-suspended in pre-warmed HFB + 5 mM sodium bicarbonate and incubated for 30 min (37 °C, 5% CO_2) with shaking (800 rpm) to an OD_{600} between 1 and 5. Cultures were diluted (100X) into HFB + 5 mM sodium bicarbonate. Each isolate culture was split into treated (0.5 µg/mL final concentration of ciprofloxacin) and control (water instead of antibiotic) samples. Samples were incubated at 37 °C for 10 min on a static hot plate. A 90 µL aliquot of each sample was placed into 180 µL of Qiagen RNA Protect Reagent for immediate RNA stabilization. A 5 µL aliquot of each sample was plated onto a Chocolate

Agar plate and incubated overnight (37 °C, 5% CO_2) as a control for the exposure experiments. If the expected growth phenotypes (i.e. resistant = growth; susceptible = no growth) were not observed for any single sample in the plating control, the exposure experiment was repeated for the set of samples. From the 50 total isolates available from the *N. gonorrhoeae* panel of the CDC Antimicrobial Resistance Isolate Bank, 49 were used in this study. One isolate was excluded from this study because we suspected that it had been contaminated; we did not detect *porB* primer amplification using qPCR.

RNA sequencing and analysis

RNA was extracted using the Enzymatic Lysis of Bacteria protocol of the Qiagen RNeasy Mini Kit and processed according to the manufacturer's protocol. DNA digestion was performed during extraction using the Qiagen RNase-Free DNase Set. The quality of extracted RNA was measured using an Agilent 2200 TapeStation (Agilent, Santa Clara, CA, USA). Extracted RNA samples were prepared for sequencing using the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswitch, MA, USA) and the NEBNExt Multiplex Oligos for Illumina. Libraries were sequenced at 50 single base pair reads and a sequencing depth of 10 million reads on an Illumina HiSeq 2500 System (Illumina, San Diego, CA, USA) at the Millard and Muriel Jacobs Genetics and Genomics Laboratory, California Institute of Technology. Raw reads from the sequenced libraries were subjected to quality control to filter out low-quality reads and trim the adaptor sequences using Trimmomatic (version 0.35). The reads were aligned to the FA 1090 strain of *N. gonorrhoeae* (NCBI Reference Sequence: NC_002946.2) using Bowtie2 (version 2.2.5) and quantified using the Subread package (version 1.5.0-p1). A pseudocount of 1 was added to the gene quantification; gene expression was defined in transcripts per million (TPM).

Marker selection

For each gene, we defined the C:T ratio as the gene expression (TPM) in the control sample divided by the gene expression (in TPM) in the treated sample. We plotted the $-\log_2$(C:T) against the $-\log_2$(expression in TPM) for all genes. To identify genes that were differentially expressed between control and treated samples, we defined a threshold of significance. The threshold of significance was calculated from the C:T ratios at t = 0 min for the biological replicates that were sequenced (three susceptible and three resistant isolates). For each of the six gene expression datasets (one for each isolate), we fit a negative exponential curve to the outer edge of each plot and then averaged the curves from all six datasets. Finally, we added a 90% confidence interval to the average curve by assuming a Gaussian fit for the error distribution, which is our threshold of significance. Genes with a $-\log_2$(C:T) value above or below the upper and lower thresholds were identified as differentially expressed. Genes that were differentially expressed consistently (either always above or always below the thresholds) among the three susceptible isolates and were not differentially expressed among the three resistant isolates were defined as candidate markers.

Copies/cell measurements from sequencing

To measure copies per cell using sequencing data, we added 2uL of (1/1000 dilution) ERCC RNA Spike-In Mix (Thermo Fisher Scientific, Waltham, MA, USA) to the lysis buffer in the RNeasy Mini Kit to each individual sample. We calculated the number of copies of each ERCC transcript in the sample, by accounting for dilution and multiplying by Avogadro's number (manufacturer's concentrations were reported in attomoles/μL). We plotted the relationship between $\log_2$(ERCC copies added) against $\log_2$(gene expression in TPM) and performed a linear regression in the region of linearity. We used the linear regression to convert TPM values to total RNA copies in each sample. Finally, using the CFU measured for each sample from plating (described in the "Antibiotic exposure for RNA sequencing" section), the total RNA copies were converted to copies per cell.

Validation with droplet digital PCR (dPCR)

Primers were designed for candidate markers using Primer-BLAST[39] and primer alignments were verified using SnapGene. Expression of candidate markers was quantified using the Bio-Rad QX200 droplet dPCR system (Bio-Rad Laboratories, Hercules, CA, USA). The concentration of the components in the dPCR mix used in this study were as follows: 1× EvaGreen Droplet Generation Mix (Bio-Rad), 150U/mL WarmStart RTx Reverse Transcriptase, 800U/mL RiboGaurd RNase Inhibitor, 500 nM forward primer, and 500 nM reverse primer. The RNA extraction comprised 5% of the final volume in the dPCR mix. The remaining volume was nuclease-free water. For each isolate, candidate marker expression was quantified in the control and treated samples and the fold-change difference (C:T ratio) was calculated. To account for potential differences between the control and treated samples that could arise from experimental variability and extraction efficiency, we used ribosomal RNA (rRNA) as an internal control because from our sequencing data, we found that rRNA was not affected by antibiotic exposure in the time frame of this study. To normalize by rRNA, we quantified the 16S rRNA in the control and treated samples by dPCR and calculated an rRNA C:T ratio. We then divided the C:T ratio of each marker by the rRNA C:T ratio. All dPCR C:T ratios reported in this paper are the normalized C:T ratios.

References

1 Centers for Disease Control and Prevention. Sexually transmitted disease surveillance, 2016 (2017).

2 World Health Organization. Global action plan to control the spread and impact of antimicrobial resistance in *Neisseria gonorrhoeae* (2012).

3 Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2013 (2013).

4 Unemo, M. *et al.* High-level cefixime-and ceftriaxone-resistant *Neisseria gonorrhoeae* in France: novel penA mosaic allele in a successful international clone causes treatment failure. *Antimicrob. Agents Chemother.* **56**, 1273-1280 (2012).

5 Ohnishi, M. *et al.* ceftriaxone-resistant *Neisseria gonorrhoeae*, Japan. *Emerg. Infect. Dis.* **17**, 148 (2011).

6 Lahra, M. M., Ryder, N. & Whiley, D. M. A new multidrug-resistant strain of *Neisseria gonorrhoeae* in Australia. *N. Engl. J. Med.* **371**, 1850-1851 (2014).

7 Cámara, J. *et al.* Molecular characterization of two high-level ceftriaxone-resistant *Neisseria gonorrhoeae* isolates detected in Catalonia, Spain. *J. Antimicrob. Chemother.* **67**, 1858-1860 (2012).

8 Deguchi, T. *et al.* New clinical strain of *Neisseria gonorrhoeae* with decreased susceptibility to ceftriaxone, Japan. *Emerg. Infect. Dis.* **22**, 142 (2016).

9 Hemarajata, P., Yang, S., Soge, O., Humphries, R. & Klausner, J. Performance and verification of a real-time PCR assay targeting the gyrA gene for prediction of ciprofloxacin resistance in *Neisseria gonorrhoeae*. *J. Clin. Microbiol.* **54**, 805-808 (2016).

10 Pond, M. J. *et al.* Accurate detection of *Neisseria gonorrhoeae* ciprofloxacin susceptibility directly from genital and extragenital clinical samples: towards genotype-guided antimicrobial therapy. *J. Antimicrob. Chemother.* **71**, 897-902 (2016).

11 Aas, F. E., Løvold, C. & Koomey, M. An inhibitor of DNA binding and uptake events dictates the proficiency of genetic transformation in *Neisseria gonorrhoeae*: mechanism of action and links to type IV pilus expression. *Mol. Microbiol.* **46**, 1441-1450 (2002).

12 Hamilton, H. L. & Dillard, J. P. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. *Mol. Microbiol.* **59**, 376-385 (2006).

13 Baltekin, Ö., Boucharin, A., Tano, E., Andersson, D. I. & Elf, J. Antibiotic susceptibility testing in less than 30 min using direct single-cell imaging. *Proc. Natl. Acad. Sci. U.S.A.*, 201708558 (2017).

14 Liu, T.-T. *et al.* A high speed detection platform based on surface-enhanced Raman scattering for monitoring antibiotic-induced chemical changes in bacteria cell wall. *PloS One* **4**, e5470 (2009).

15 Broeren, M., Maas, Y., Retera, E. & Arents, N. Antimicrobial susceptibility testing in 90 min by bacterial cell count monitoring. *Clin. Microbiol. Infect.* **19**, 286-291 (2013).

16 Fredborg, M. *et al.* Real-time optical antimicrobial susceptibility testing. *J. Clin. Microbiol.* **51**, 2047-2053 (2013).

17 Spence, J. M., Wright, L. & Clark, V. L. Laboratory maintenance of *Neisseria gonorrhoeae*. *Curr. Protoc. Microbiol.*, 4A. 1.1-4A. 1.26 (2008).

18 Mezger, A. *et al.* A general method for rapid determination of antibiotic susceptibility and species in bacterial infections. *J. Clin. Microbiol.* **53**, 425-432 (2015).

19 Rolain, J., Mallet, M., Fournier, P. & Raoult, D. Real-time PCR for universal antibiotic susceptibility testing. *J. Antimicrob. Chemother.* **54**, 538-541 (2004).

20 Schoepp, N. G. *et al.* Rapid pathogen-specific phenotypic antibiotic susceptibility testing using digital LAMP quantification in clinical samples. *Sci. Transl. Med.* **9**, eaal3693 (2017).

21 Tobiason, D. M. & Seifert, H. S. The obligate human pathogen, *Neisseria gonorrhoeae*, is polyploid. *PLoS Biol.* **4**, e185 (2006).

22 Cooper, S. & Helmstetter, C. E. Chromosome replication and the division cycle of *Escherichia coli* Br. *J. Mol. Biol.* **31**, 519-540 (1968).

23 Sangurdekar, D. P., Srienc, F. & Khodursky, A. B. A classification based framework for quantitative description of large-scale microarray data. *Genome Biol.* **7**, R32 (2006).

24 Barczak, A. K. *et al.* RNA signatures allow rapid identification of pathogens and antibiotic susceptibilities. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6217-6222 (2012).

25 Bhattacharyya, R. *et al.* Rapid Phenotypic Antibiotic Susceptibility Testing Through RNA Detection. *Open Forum Infect. Dis.* **4**, S33-S33, 10.1093/ofid/ofx162.082 (2017).

26 McClure, R. *et al.* The gonococcal transcriptome during infection of the lower genital tract in women. *PloS One* **10**, e0133982 (2015).

27 Remmele, C. W. *et al.* Transcriptional landscape and essential genes of Neisseria gonorrhoeae. *Nucleic Acids Res.* **42**, 10579-10595 (2014).

28 Stohl, E. A., Gruenig, M. C., Cox, M. M. & Seifert, H. S. Purification and characterization of the RecA protein from *Neisseria gonorrhoeae*. *PloS One* **6**, e17101 (2011).

29 Schook, P. O., Stohl, E. A., Criss, A. K. & Seifert, H. S. The DNA-binding activity of the Neisseria gonorrhoeae LexA orthologue NG1427 is modulated by oxidation. *Mol. Microbiol.* **79**, 846-860 (2011).

30 Qin, T.-T. *et al.* SOS response and its regulation on the fluoroquinolone resistance. *Ann. Transl. Med.* **3** (2015).

31 Black, C. G., Fyfe, J. A. & Davies, J. K. Absence of an SOS-like system in *Neisseria gonorrhoeae*. *Gene* **208**, 61-66 (1998).

32 Stohl, E. A. & Seifert, H. S. *Neisseria gonorrhoeae* DNA recombination and repair enzymes protect against oxidative damage caused by hydrogen peroxide. *J. Bacteriol.* **188**, 7645-7651 (2006).

33 LeBel, M. Ciprofloxacin: chemistry, mechanism of action, resistance, antimicrobial spectrum, pharmacokinetics, clinical trials, and adverse reactions. *Pharmacotherapy* **8**, 3-30 (1988).

34 Gao, Q. *et al.* Gene expression diversity among *Mycobacterium tuberculosis* clinical isolates. *Microbiology* **151**, 5-14 (2005).

35 Quillin, S. J. & Seifert, H. S. *Neisseria gonorrhoeae* host adaptation and pathogenesis. *Nat. Rev. Microbiol.* **16**, 226-240 (2018).

36 The European Committee on Antimicrobial Susceptibility Testing. Ciprofloxacin/*Neisseria gonorrhoeae* International MIC Distribution - Reference Database 2018-04-03. https://mic.eucast.org/Eucast2/regShow.jsp?Id=35702 (2018).

37 Gomez, J. E. *et al.* Ribosomal mutations promote the evolution of antibiotic resistance in a multidrug environment. *Elife* **6** (2017).

38 Fernández, L. & Hancock, R. E. Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. *Clin. Microbiol. Rev.* **25**, 661-681 (2012).

39 Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).

Supplementary Information

Supplementary Table S1.1. List of candidate markers and their expression in transcripts per million (TPM) and copies per cell for susceptible isolate S2 and resistant isolate R2 after 15 min of ciprofloxacin exposure. The genome used for alignment was *N. gonorrhoeae* FA1090 (NCBI Reference Sequence: NC_002946.2).

		Susceptible (S2) Control		Susceptible (S2) Treated		Resistant (R2) Control		Resistant (R2) Treated	
Locus Tag	**Gene Description**	**TPM**	**copies/cell**	**TPM**	**copies/cell**	**TPM**	**copies/cell**	**TPM**	**copies/cell**
NGO0340	cysteine synthase A (*cysK*)	894.1	21.1	505.2	8.9	551.8	16.3	600.0	20.0
NGO1837	50S ribosomal protein L4 *(rplD*)	474.9	10.8	262.2	4.4	403.6	11.9	425.4	13.8
NGO1843	elongation factor G (fusA)	433.4	9.8	224.9	3.8	432.9	12.8	503.5	16.6
NGO2024	50S ribosomal protein L13 (*rplM*)	415.0	9.4	213.5	3.6	455.3	13.5	503.5	16.6
NGO1845	30S ribosomal protein S12 (*rpsL*)	563.1	13.0	286.8	4.9	615.4	18.2	697.6	23.5
NGO1677	50S ribosomal protein L27 (rpmA)	410.7	9.3	192.2	3.2	500.6	14.8	497.6	16.4
NGO1844	30S ribosomal protein S7	520.0	11.9	241.3	4.0	520.1	15.4	651.6	21.9
NGO0171	50S ribosomal protein L19 (*rplS*)	379.2	8.5	175.0	2.9	328.5	9.7	353.2	11.3
NGO1834	30S ribosomal protein S19 (*rpsS*)	330.0	7.4	152.1	2.5	260.9	7.7	292.7	9.2
NGO0172	tRNA (guanine-N(1)-)-methyltransferase (*trmD*)	237.3	5.2	108.8	1.7	208.8	6.2	224.6	6.9
NGO1835	50S ribosomal protein L2 (*rplB*)	392.5	8.9	179.1	2.9	297.6	8.8	359.8	11.5
NGO1673	type IV pilus assembly protein (*pilB*)	225.9	4.9	101.5	1.6	199.3	5.9	214.9	6.6
NGO1833	50S ribosomal protein L22 (*rplV*)	343.8	7.7	147.9	2.4	292.1	8.6	304.3	9.6
NGO2173	50S ribosomal protein L32 (*rpmF*)	407.5	9.2	173.6	2.9	394.7	11.7	404.1	13.1
NGO0604	30S ribosomal protein S1 (*rpsA*)	437.9	9.9	185.3	3.1	456.3	13.5	493.9	16.2
NGO0016	preprotein translocase subunit (*secG*)	180.1	3.9	73.7	1.1	169.1	5.0	184.5	5.6
NGO2174	hypothetical protein	372.8	8.4	150.2	2.4	368.3	10.9	361.6	11.6
NGO2164	GMP synthase (guaA)	118.3	2.5	45.0	0.7	98.6	2.9	109.4	3.2
NGO1676	50S ribosomal protein L21 (*rplU*)	554.6	12.8	200.4	3.3	555.2	16.4	587.7	19.6

NGO1679	50S ribosomal protein L33 (*rpmG*)	283.8	6.3	101.4	1.6	298.5	8.8	284.3	8.9
NGO1658	hypothetical protein	98.4	2.1	33.8	0.5	118.3	3.5	116.1	3.4
NGO1440	macrolide transport protein MacA	143.3	3.1	48.6	0.7	132.3	3.9	139.7	4.2
NGO0174	30S ribosomal protein S16 (*rpsP*)	315.2	7.0	101.2	1.6	295.8	8.7	340.5	10.9
NGO0173	ribosome maturation factor RimM (*rimM*)	359.8	8.1	113.5	1.8	316.8	9.4	318.8	10.1
NGO0592	trigger factor (*tig*)	146.5	3.1	45.5	0.7	147.5	4.3	152.1	4.6
NGO1680	50S ribosomal protein L28 (*rpmB*)	452.8	10.3	130.3	2.1	470.2	13.9	525.4	17.3
NGO0620	aspartate alpha-decarboxylase	64.8	1.3	18.6	0.3	54.2	1.6	59.3	1.7
NGO1659	intracellular septation protein A	62.2	1.3	17.8	0.3	63.6	1.9	70.7	2.0
NGO1291	transcriptional regulator (yebC)	64.1	1.3	18.0	0.3	79.9	2.3	77.9	2.2
NGO0648	membrane protein	56.4	1.1	15.3	0.2	47.6	1.4	45.2	1.2
NGO0593	ATP-dependent Clp protease proteolytic subunit (*clpP*)	60.2	1.2	16.0	0.2	73.6	2.2	75.9	2.2
NGO1804	(3R)-hydroxymyristoyl-ACP dehydratase (*fabZ*)	91.0	1.9	24.0	0.3	74.6	2.2	73.5	2.1
NGO0618	membrane protein	81.4	1.7	20.1	0.3	66.8	2.0	70.2	2.0
NGO0619	2-dehydro-3-deoxyphosphooctonate aldolase	61.1	1.2	15.1	0.2	51.1	1.5	62.6	1.8
NGO1812	major outer membrane protein (*porB*)	1293.2	31.2	293.4	5.0	1459.1	43.3	1587.1	57.1
NGO1890	glutamate permease; sodium/glutamate symport carrier protein	35.0	0.7	7.5	0.1	40.3	1.2	48.9	1.3
NGO2098	diaminopimelate decarboxylase	26.0	0.5	4.9	0.1	18.6	0.5	18.6	0.5
NGO2100	frataxin-like protein (*cyaY*)	20.4	0.4	3.6	0.0	14.0	0.4	18.1	0.5

Supplementary Table S1.2. Primer sequences used for validation of candidate markers by digital PCR.

Candidate Marker	Gene Name	Forward Primer Sequence	Reverse Primer Sequence
porB	major outer membrane porin	GCTACGATTCTCCCGAATTTGCC	CCGCCKACCAAACGGTGAAC
rpmB	50S ribosomal protein L28	TTGCCCAACTTGCAATCACG	AGCACGCAAATCAGCCAATAC
tig	trigger factor	AAAGCCTTGGGTATTGCGG	TGACCAAAGCAACCGGAAC
yebC	YebC/PmpR family Transcriptional Regulator	GCTTTGGAAAAAGCAGCCG	GGTTTTGTTGTCGGTCAGGC
pilB	Type IV-A pilus assembly ATPase	GACTTTTGCCGCTGCTTTG	GCGCATTATTCGTGTGCAG
cysK	Cysteine synthase A	GAGGCTTCCCCCGTATTGAG	TTCAAAAGCCGCTTCGTTCG
16S rRNA	16S ribosomal RNA	ACTGCGTTCTGAACTGGGTG	GGCGGTCAATTTCACGCG

Supplementary Table S1.3. Minimum inhibitory concentration **(**MIC) values for the 49 *Neisseria gonorrhoeae* clinical isolates acquired from the CDC and FDA Antibiotic Resistance Isolate Bank[1].

MIC	Number of strains	Susceptible or Resistant
0.015	8	Susceptible
0.03	1	Susceptible
4	1	Resistant
8	6	Resistant
16	33	Resistant

[1] CDC and FDA Antibiotic Resistance Isolate Bank. Atlanta (GA): CDC. (2018)

Chapter II

A QUANTITATIVE SEQUENCING FRAMEWORK FOR ABSOLUTE ABUNDANCE MEASUREMENTS OF MUCOSAL AND LUMENAL MICROBIAL COMMUNITIES

This chapter was originally published in: "**Barlow J.,** Bogatyrev S., and Ismagilov R. "A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities." *Nature communications* 11.1 (2020): 1-13**.** doi:10.1038/s41467-020-16224-6."

Abstract

A fundamental goal in microbiome studies is determining which microbes affect host physiology. Standard methods for determining changes in microbial taxa measure relative, rather than absolute abundances. Moreover, studies often analyze only stool, despite microbial diversity differing substantially among gastrointestinal (GI) locations. Here, we develop a quantitative framework to measure absolute abundances of individual bacterial taxa by combining the precision of digital PCR with the high-throughput nature of 16S rRNA gene amplicon sequencing. In a murine ketogenic-diet study, we compare microbial loads in lumenal and mucosal samples along the GI tract. Quantitative measurements of absolute (but not relative) abundances reveal decreases in total microbial loads on the ketogenic diet and enable us to determine the differential effects of diet on each taxon in stool and small-intestine mucosa samples. This rigorous quantitative microbial analysis framework, appropriate for diverse GI locations enables mapping microbial biogeography of the mammalian GI tract and more accurate analyses of changes in microbial taxa in microbiome studies.

Introduction

One main goal of microbiome studies is to determine which taxa, if any, drive phenotypic changes among study groups.[1-3] The first step in this process is often to survey which microbial taxa differ in abundance between study groups (differentially abundant taxa). This survey is commonly performed by amplifying the 16S rRNA gene amplicon with "universal" primer sets before high throughput sequencing.[4] The output of these studies provides the relative, not absolute, abundance of each taxon in each sample. Researchers often then use standard statistical tests or microbiome specific packages to determine which taxa are differentially abundant.[5, 6]

Relative-abundance analyses are effective for determining the major microbial taxa in an environment (e.g., the human Microbiome Project). However, several researchers have pointed out the inherent limitations of comparing relative abundances between samples.[7-10]

In analyses of relative data, every increase in one taxon's abundance causes an equivalent decrease across the remaining taxa. Thus, the measurement of a taxon's relative abundance is dependent on the abundance of all other taxa, which can lead to high false positive rates in differential taxon analyses[8, 11-13] and negative-correlation biases in correlation-based analyses.[14, 15] Several methods (e.g., ALDEx2[16], Ancom[17], Gneiss[18], Differential Ranking[10]) acknowledge these biases and aim to address them by using the ratios among taxa, which are conserved regardless of whether the data are relative or absolute. These methods are particularly valuable because they enable improved re-analysis of existing datasets reporting relative abundances.[10, 16-18]

Despite such methodological advancements, analyses of relative abundance cannot fully capture how individual microbial taxa differ among samples or experimental conditions. Using the simple example of a community containing two taxa (Fig. 2.1), we see that an increase in the ratio between Taxon A and Taxon B could indicate one of five scenarios: (i) Taxon A increased (Fig. 2.1a), (ii) Taxon B decreased (Fig. 2.1b), (iii) A combination of 1 and 2, (iv) Taxon A and Taxon B increased but Taxon A increased by a greater magnitude, or (v) Taxon A and Taxon B decreased but Taxon B decreased by a greater magnitude (Fig. 2.1c). Knowing which of these five scenarios occurs when analyzing experimental data could drastically alter the interpretation of which taxa are positively or negatively associated with phenotypes. Thus, an inherent limitation of methods that use relative abundance is that they cannot determine whether an individual taxon is more abundant or less abundant (the direction of the change) or by how much (the magnitude of the change) between two experimental conditions or samples.

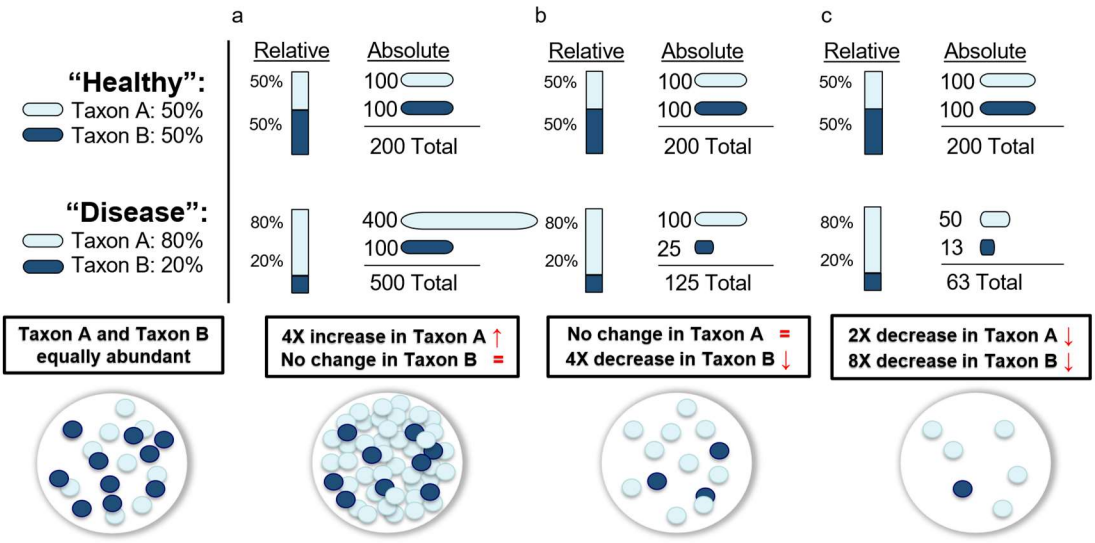


Figure 2.1: The value of absolute (compared with relative) quantification illustrated by three hypothetical scenarios. In this hypothetical, two taxa (Taxon A and Taxon B) are found in equal abundance (50:50) in a "healthy" state but in an 80:20 ratio in the "disease" state. Three possible scenarios arise: (a) Taxon A increases in abundance while Taxon B remains the same; (b) Taxon A remains unchanged while Taxon B decreases in abundance, and (c) Taxon A and Taxon B both decrease, but Taxon B decreases by a greater magnitude.

To overcome these limitations, several important methods have been developed for quantifying the absolute abundance of microbial taxa by using known "anchor" points to

convert relative data to absolutes. Spiked standards are commonly used in method calibration and have recently been applied to quantifying taxa in microbiome research.[19-23] These methods require a purified DNA sequence of known concentration from an organism not present in the sample and an estimate of the initial sample concentration to determine the amount of exogenous DNA to spike-in. Another group of anchoring methods, such as those that use flow cytometry[24], total DNA[25], or qPCR[26-28], measure the total concentration of cells, DNA, or amplicons to transform the relative abundances to absolute numbers. These methods have already demonstrated the value of quantitative microbiome analysis, yet microbiome researchers have not yet uniformly adapted these methods. One may speculate that this lack of adoption is because of real or potential limitations of these methods. For example, flow-cytometry based methods require dissociating the sample into single bacterial cells, which could require complex sample preparation and have not been validated with complex samples such as from gut mucosa. Total-DNA-based methods are limited to samples only containing microbial DNA (no host DNA), and spike-in or qPCR-based methods can be limited by amplification biases.[29, 30] To increase utilization of quantitative microbiome analyses, the following capabilities and validation need to be demonstrated: (i) performance across samples with microbial loads ranging from high, as in stool, to low, as in the small intestine; (ii) performance across biogeographically diverse sample types, from microbe-rich stool and colonic contents to host-rich mucosal samples; (iii) explicit evaluation of limits of quantification of the method, and how these limits depend on the starting microbial load, relative abundance of a specific target taxon in the sample, and sequencing depth.

To address this challenge, in this paper we establish a rigorous, absolute quantification framework based on digital PCR (dPCR) anchoring. We chose dPCR as our anchoring method because PCR is already part of sequencing protocols and has been extensively validated as a quantitative method in nucleic-acid measurements. To achieve precise measurements of absolute abundance from diverse sample types, we assessed the efficiency and evenness of the DNA extraction protocol. To minimize and quantify bias resulting from potentially uneven amplification of microbial 16S rRNA gene DNA, or non-specific amplification of host DNA, we utilized dPCR in a microfluidic format.[31-33]–. dPCR is an ultrasensitive method for counting single molecules of DNA or RNA.[34-36] By dividing a PCR reaction into thousands of nanoliter droplets and counting the number of "positive" wells (those with amplified template), dPCR yields absolute quantification without a standard curve. To understand the quantitative limits of our methodology, we measured the accuracy of each taxon's absolute abundance as a factor of both input DNA amount and individual taxon relative abundance.[37-39] We then evaluated this absolute quantification workflow by performing a murine ketogenic-diet study that illustrates how the selection of relative- vs. absolute-quantification analyses can result in different interpretations of the same experimental results. Many studies have shown that ketogenic diets can induce substantial compositional changes in gut microbiota,[40-42] so, we predicted it would serve as a good illustrative model for our workflow. Finally, we applied this workflow to an analysis of microbial loads along the entire gastrointestinal (GI) tract to highlight the importance of judicious selection of sample location when evaluating the impact of diet on host phenotype, and to highlight the applicability of this workflow to GI sites with diverse microbial loads.

Results

Efficient DNA extraction across microbial loads and sample types

To estimate the maximum quantity of sample we could extract before overloading the 20-µg column capacity, we measured total DNA and microbial DNA load across small intestine and large intestine lumenal and mucosal samples (Supplementary Figure 2.1). We then evaluated extraction efficiency across three tissue matrices (mucosa, cecum contents, and stool) to assess whether variation in levels of PCR inhibitors and non-microbial DNA interfered with microbial quantification. We spiked a defined 8-member microbial community into GI samples taken from germ-free (GF) mice. To assess quantitative limits, we performed a dilution series of microbial spike-in from 1.4 x 10^9 CFU/mL to 1.4 x 10^5 CFU/mL. dPCR quantification showed near equal and complete recovery of microbial DNA over 5 orders of magnitude (Fig. 2.2a). Overall, we measured ~ 2X accuracy in extraction across all tissue types (cecum contents, stool, SI mucosa) when total 16S rRNA gene input was greater than 8.3 x 10^4 copies (Supplementary Figure 2.2). Normalizing this sample input to the approximate maximum extraction mass (200 mg stool, 8 mg mucosa) yielded a lower limit of quantification (LLOQ) of 4.2 x 10^5 16S rRNA gene copies per gram for stool/cecum contents and 1 x 10^7 16S rRNA gene copies per gram for mucosa. Mucosal samples had a higher LLOQ because the high host DNA in this tissue type saturates the column, limiting total mass input.

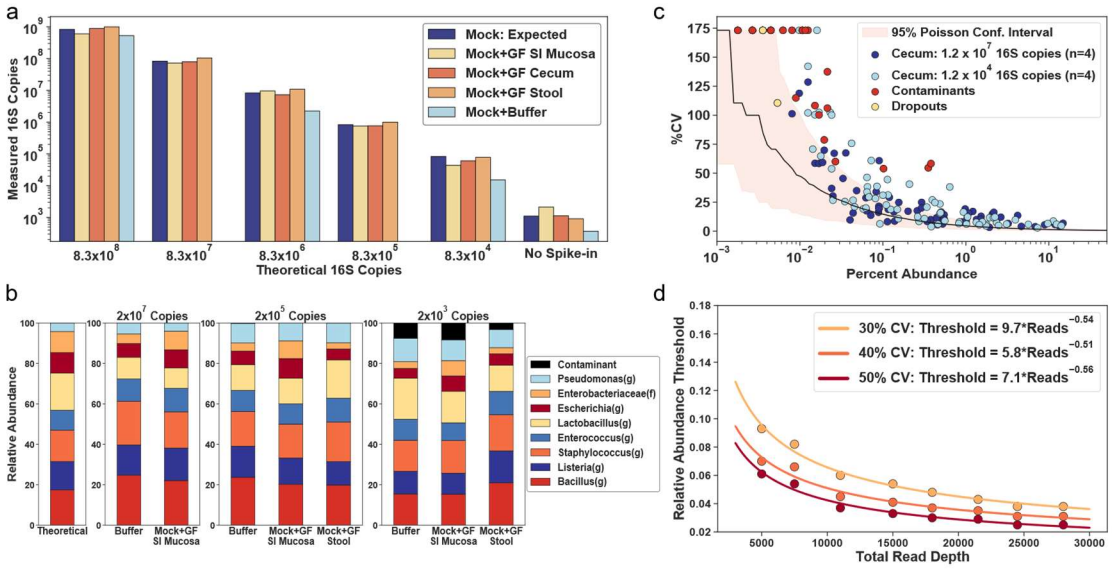


Figure 2.2: Lower limits of quantification for total microbial DNA extraction and 16S rRNA gene amplicon sequencing. (a) A comparison of theoretical and measured copies of the 16S rRNA gene with digital PCR using an eight-member microbial community spiked at a range of dilutions into germ-free (GF) mouse tissue from small-intestine (SI) mucosa, cecum, and stool. Each bar plot shows a single technical replicate for each matrix. (b) Relative abundance of the eight taxa as predicted and measured after 16S rRNA gene amplicon sequencing. (c) Correlation between the mean (n=4) relative abundance of each taxon and the coefficient of variation (%CV) using a cecum sample from a mouse on a chow diet with an initial template input of either 1.2 x 10^7 or 1.2 x 10^4 16S rRNA gene copies. Each analysis comprised four technical (sequencing) replicates. Taxa found only in the low-input sample were labeled contaminants (red points); taxa found in the high-input sample but not low input sample were labeled dropouts (yellow points). Red shading indicates the Poisson sampling 95% confidence interval (10,000 bootstrapped replicates) at a sequencing read depth of 28,000. (d) Relationship between relative abundance threshold (see text for details) and sequencing read depths at 30%, 40%, and 50% CV thresholds.

Next, to ensure extraction performance was consistent for both Gram-negative and Gram-positive microbes, we performed 16S rRNA gene amplicon sequencing using previously described improved primers and protocol[31, 33] on a subset of the extracted samples (Fig. 2.2b). It is important to note that all amplification reactions for 16S rRNA gene library prep were monitored with real-time qPCR and we stopped the reactions when they reached the late exponential phase to limit overamplification and chimera formation.[30-33, 43, 44] Extraction appeared less even among microbial taxa at lower total microbial DNA inputs (Fig. 2.2b). This discrepancy from the theoretical profile did not correlate with the presence of chimeric sequences (Supplementary Figure 2.3) and was likely a function of the reduced accuracy incurred when diluting complex microbial samples. Additionally, sequencing samples with low total microbial loads (<1 x 10^4 16S rRNA gene copies) resulted in the presence of contaminants, as confirmed by sequencing of negative-control extractions (Supplementary Table 2.1).

Quantitative limits of 16S rRNA gene amplicon sequencing

To establish the precision of relative-abundance measurements, we sequenced four replicates of DNA extractions from cecum samples. Libraries from one DNA extraction were prepared with either an input of 1.2 x 10^7 16S rRNA gene copies or 1.2 x 10^4 16S rRNA gene copies to determine the impact of starting DNA amount on sequencing variability. We calculated the coefficient of variation (%CV) for each taxon's relative abundance from amplicon sequencing the replicate samples. Each taxon's mean relative abundance (n=4) was then plotted against its corresponding coefficient of variation of the relative abundance (Fig. 2.2c). We defined "dropouts" as taxa present only in the high-DNA-input sample whereas we defined "contaminants" as taxa present only in the low-DNA-input sample. The two dropout taxa in the low input sample corresponded to the lowest abundance taxa from the high input DNA sample (yellow points, Fig. 2.2c). Most of the contaminant taxa had a relative abundance < 0.03%, but three taxa (*Pseudomonas(g)*, *Acinetobacter(g)*, *Rhizobiales(f)*) had relative abundances of 0.38%, 0.35%, and 0.1%, respectively. These three taxa were also the three most common contaminants in our negative-control extractions (Supplementary Table 2.1). The presence of contaminants in the sample containing 1.4 x 10^4 16S rRNA gene copies was consistent with the input amount at which we observed contaminants in our mixed microbial community dilutions (Fig. 2.2b). We calculated a bootstrapped Poisson sampling confidence interval at our sequencing depth (28,000 reads) to assess how close our accuracy limits were to the theoretical limits (red shading, Fig. 2.2c). At the low DNA input level of 1.2 x 10^4 16S rRNA gene copies, we began to reach the fundamental Poisson loading limit in our library-preparation reaction (Supplementary Figure 2.4a). We expected divergence of the %CV at ~0.01% abundance because at a read depth of 28,000 a relative abundance of 0.01% is a measure of ~3 reads whereas at a total 16S rRNA gene copy input of 1.4 x 10^4 a relative abundance of 0.01% is ~1 copy. Poisson statistics also helped us define the theoretical lower limits of relative-abundance measurements as a factor of sequencing depth (Supplementary Figure 2.4b).

We next wished to quantify an approximate threshold that would tell us, for a given sequencing depth, at what percentage of relative abundance we lose accuracy in our measurements (we defined this threshold as "relative abundance threshold"). To determine this threshold, we fit a negative exponential to the replicate data and identified the percentage

abundance at which 30% CV was observed. This threshold is a function of the sequencing depth, so we subsampled the data at decreasing read counts and repeated the exponential fitting method to calculate the relationship between the relative abundance threshold and sequencing depth (Fig. 2.2d). Greater sequencing depths yielded lower quantitative limits with diminishing returns, as expected. We found that the threshold for percentage abundance decreases with increasing sequencing depth with a square root dependence analogous to the square-root dependence of Poisson noise. This trend follows for %CV thresholds of 40% and 50% as well (Fig. 2.2d). This analysis provides a framework with which to impose thresholds on relative-abundance data that are grounded on the calculated limits of quantitation.

Absolute quantification of taxa via digital PCR (dPCR) anchoring

We calculated absolute abundances of taxa from sequencing data using dPCR measurement of total microbial loads as an anchor. Briefly, relative abundance of each taxon was measured by sequencing and these numbers were multiplied by the total number of 16S rRNA gene copies (obtained using the same universal primers from amplicon sequencing, without the barcodes) from dPCR. Next, we evaluated the accuracy of this quantitative sequencing approach. Typically, evaluation of quantitative accuracy and precision would involve the use of a mock microbial community (like the one used in Fig. 2.2). However, because we computed the absolute instead of relative abundances, we were able to use the actual gut-microbiota samples and compare the results to the dPCR data obtained with relevant taxa-specific primers. The 16S rRNA gene copy amount was then normalized to the mass of each extracted sample after correcting for volume losses (Materials and methods; Equation 1). We chose four representative taxa to encompass common gut flora of varying classification levels: *Akkermansia muciniphila(s)*, *Lachnospiraceae(f)*, *Bacteroidales(o)*, and *Lactobacillaceae(f)*. Like eubacterial primers, taxa-specific primer sets can (in principle) give rise to nonspecific amplification due to overlap with host mitochondrial DNA. To avoid nonspecific amplification, we ran temperature gradients with GF mucosal DNA and taxa-specific microbial DNA to identify the optimal annealing temperature for each primer set (Supplementary Figure 2.5). Each taxa-specific primer targets a separate region of the 16S rRNA gene than the universal primer set, thus keeping the gene copy number equivalent across primers. We observed high correlation coefficients between the taxa load determined by quantitative sequencing with dPCR anchoring and the taxa load measured by dPCR with taxa-specific primers (all $r^2 >= 0.97$, Fig. 2.3a) for all four taxa over a range of ~ 6 orders of magnitude. The ratio of the total load measurements obtained by quantitative sequencing with dPCR anchoring and by dPCR with taxa-specific primers showed unity agreement between three of the four primer sets with 2-fold deviation from the mean (Fig. 2.3b and Supplementary Figure 2.9). Sequencing quantification was consistently 2.5-fold higher than dPCR quantification for the species *Akkermansia muciniphila* (Fig. 2.3b). We cannot confirm amplification bias as a factor because the error did not depend on the number of cycles used in library preparation. An alternative factor could be a discrepancy in coverage/specificity between the taxon-specific and universal primer sets. We next tested the limits of the sequencing accuracy as a factor of input DNA load. A 10X dilution series of a cecum sample was created to cover input DNA loads of $1x10^8$ copies down to $1x10^4$ copies. Minimal differences in beta diversity (Aitchison distance) between the undiluted and diluted samples were observed with a trend towards increasing difference with decreasing DNA load (Fig. 2.3c). This negative correlation between beta diversity and microbial load is not

unexpected due to the higher presence of contaminant species from our negative controls in the lower input samples (Fig 2b).

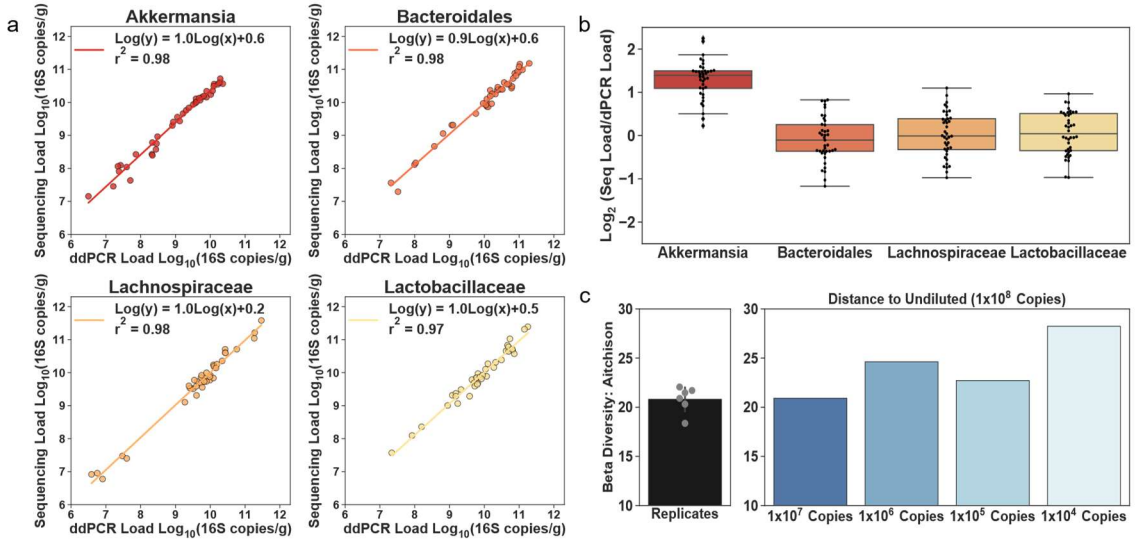


Figure 2.3: Digital PCR (dPCR) anchoring of 16S rRNA gene amplicon sequencing provides microbial absolute abundance measurements. Taxon-specific dPCR demonstrates low biases in abundance measurements calculated by 16S rRNA gene sequencing with dPCR anchoring. (a) Correlation between the Log_{10} abundance of four bacterial taxa as determined by taxa-specific dPCR and 16S rRNA gene sequencing with dPCR anchoring (relative abundance of a specific taxon measured by sequencing * total 16S rRNA gene copies measured by dPCR). (b) The Log_2 ratio of the absolute abundance of four bacterial taxa as determined either by taxa-specific dPCR or by 16S rRNA gene sequencing with dPCR anchoring (N = 32 samples). Data points are overlaid on the box and whisker plot. The body of the box plot goes from the first to third quartiles of the distribution and the center line is at the median. The whiskers extend from the quartiles to the minimum and maximum data points within the 1.5× interquartile range, with outliers beyond. All dPCR measurements are single replicates. (c) Analysis of beta diversity in cecum samples at a series of 10X dilutions (n = 1 for each dilution). Mean Aitchison distance for six pairwise comparisons of n = 4 sequencing replicates of the undiluted (10^8 copies) sample is shown for reference (error bar is standard deviation). Individual data points are overlaid on the replicates bar plot.

Absolute vs relative abundance analysis in a ketogenic-diet study

To test the impact of using a quantitative framework for 16S rRNA gene amplicon sequencing, we performed a ketogenic-diet study. Our goals were twofold. First, we wished to test whether absolute instead of relative microbial abundances can more accurately quantify changes in microbial taxa between study groups. Second, we wished to investigate how using a quantitative sequencing framework can guide the interpretation of changes in taxa across study conditions. We emphasize that our objective was not to make claims about the effect of a ketogenic diet on the microbiome, but rather to use this model as an illustration of the added benefits of using this quantitative sequencing framework.

After one week on a standard chow diet, 4-week old Swiss Webster mice were split into two groups (n=6 each): one was fed a ketogenic diet and the other a vitamin and mineral matched control diet (Supplementary Table 2.3). Stool was sampled immediately before the two diets were introduced (day 0), and again at days 4, 7 and 10. Additionally, on day 10, all mice

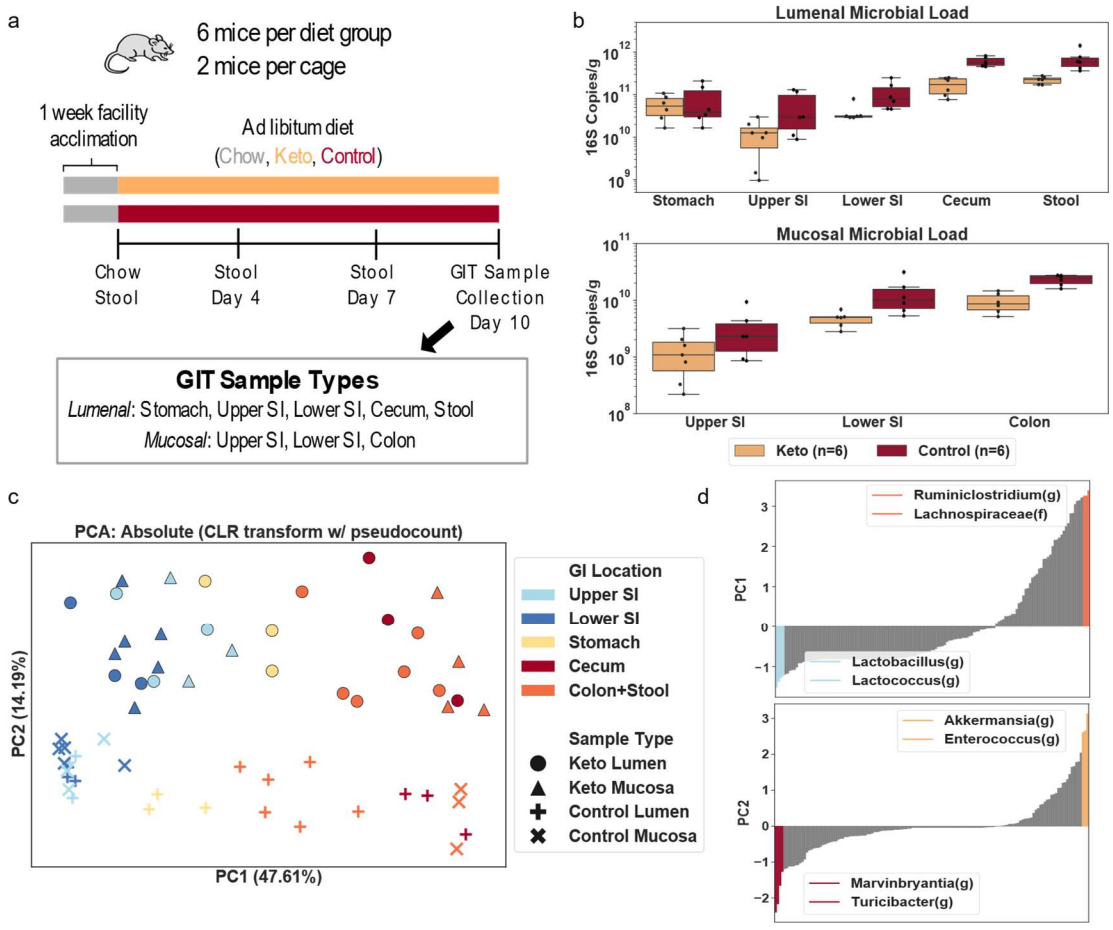
were euthanized and lumenal and mucosal samples were collected from throughout the GI tract (Fig. 2.4a). Microbial loads (quantified with dPCR) ranged from ~10^9 16S rRNA gene copies/g in small intestinal mucosa to ~10^{12} 16S rRNA gene copies/g in stool. On average, we observed lower microbial DNA loads in the mice on the ketogenic diet compared with mice on the control diet, except in the stomach, where loads were similar in mice on both diets (Fig. 2.4b).

All stool samples and roughly half of the samples for all other GI sites (evenly distributed across mice on the two diets) underwent 16S rRNA gene amplicon sequencing. Ordination methods (PCA, PCoA, NMDS, etc) are a common exploratory data analysis technique in the microbiome field. Common transformation techniques based on non-Euclidian distances (e.g., Bray-Curtis, UniFrac) can skew the accuracy of visualizations of relative data (Supplementary Figure 2.6a).[11] We used the centered log-ratio transformation (CLR, often used to compute the Aitchison distance) to handle compositional effects, and performed PCA on the transformed absolute abundance data for all samples from the final collection day (Fig. 2.4c). A clear separation along the first two principal components (PC) was observed. Separation along PC1 was related to the location within the GI tract whereas separation along PC2 was related to the diet. The PCA analysis suggested that stomach samples were distributed somewhere in-between small-intestine and large-intestine samples, possibly resulting from coprophagy in mice.[32, 33] Additionally, the mucosal and lumenal samples from the small intestine on the control diet seemed to be closer together than on the ketogenic diet (Fig. 2.4c).

We next investigated which taxa were contributing to separation in our principal component space. We calculated the scaled covariance between each taxon and the first two principal components by multiplying the eigenvectors by the square root of their corresponding eigenvalues. These values are also known as "feature loadings." Plotting these feature loadings from smallest to highest shows that *Lactobacillus(g)* and *Lactococcus(g)* had the greatest impact on separation along PC1 in the direction of the small intestinal samples whereas *Ruminiclostridium(g)* and *Lachnospiraceae(f)* separated in the direction of the large intestine (Fig. 2.4d). This matches with what we know about the major genera commonly present in the small and large intestine.[45] Along PC2 (the "diet axis"), the top two contributing taxa towards the control diet were *Turicibacter(g)* and *Marvinbryantia(g)*, while towards the ketogenic diet *Akkermansia(g)* and *Enterococcus(g)* had the greatest covariance. Although the CLR transformation preserves distances in principal component space regardless of whether the starting data are relative or absolute, it normalizes out the differences in total loads by looking at log ratios between each taxon's abundance and the geometric mean of the sample (Supplementary Figure 2.6b). In many cases, we want to know if the absolute load of a taxon is higher or lower under different conditions (e.g., in mice on ketogenic and control diets). When the total microbial load varies among samples, analyses of relative abundance cannot determine which taxa are differentially abundant (Fig. 1). To assess the impact of using absolute quantification in analyses, we analyzed microbiomes of stool samples from mice on ketogenic and control diets. PCA analysis on the CLR-transformed relative abundances of microbial taxa showed separation between the two diets (Fig. 2.5a). Feature loadings were analyzed as before, but this time total impact of each taxa on the PC space was plotted, which was defined as the sum of the feature loading vectors in

PC1 and PC2 (Fig. 2.5b). The same analysis was performed on the log-transformed absolute abundance data (Fig. 2.5a). Separation between diets is clear in both relative and absolute abundance analyses, but the contribution of each taxon to the separation differed in direction and magnitude. Comparing the magnitude of feature loadings for two taxa, *Akkermansia(g)* and *Acetatifactor(g)*, between the relative and absolute PCA plots showed obvious differences in the contribution of a given taxa to the separation in principal-component space. Analysis of relative-abundance data implies that *Akkermansia(g)* has the biggest contribution on separation between diets in PC space whereas the absolute abundance data implies that ~50% of the taxa in the sample have a greater contribution than *Akkermansia(g)* to the separation between the diets in PC space.

Figure 2.4: Microbial absolute abundances provide separation between GI locations of mice on ketogenic or control diets. Analysis of data comparing ketogenic and control diets provides changes of total microbial loads, separation of microbial communities by GI location and by diet in principal component analysis, and the top taxa driving the separation of samples along the principal components. (a) Overview of experimental setup and sample-collection protocol. Gastrointestinal tract (GIT) samples were collected from the following regions: stomach, upper small intestine (SI), lower SI, cecum, colon, and stool. (b) Comparison of total microbial loads between ketogenic and control diets in lumenal (top) and mucosal (bottom) samples collected after 10 days on



each diet. The body of the box plot goes from the first to third quartiles of the distribution and the center line is at the median. The whiskers extend from the quartiles to the minimum and maximum data point within 1.5× interquartile range, with outliers beyond. (c) Principal component analysis (PCA) on the centered log-ratio transformed absolute abundances of microbial taxa shows separation by GI location (Upper SI, light blue;

Lower SI, dark blue; Stomach, yellow; Cecum, red; Colon+Stool, orange) and diet (Ketogenic, circles and triangles; Control, X's and crosses). (d) Ranked order of the eigenvector coefficients scaled by the square root of the corresponding eigenvalue (feature loadings) for the top two principal components. The two most positive and most negative taxa are shown.

PCA is only an exploratory data-analysis technique, so we next used a non-parametric statistical test to test for differentially abundant taxa in stool samples from mice on control and ketogenic diets (Fig. 2.5c).[46] We performed separate analyses of the relative and absolute abundance data. We plotted the -log10 *P*-value for each taxon's relative abundances against the corresponding -log10 *P*-value for that taxon's absolute abundances. Points along the diagonal indicate congruence between the predictions from the relative and absolute abundance data. Points in the upper left corner indicate taxa that differed between the diets in the analysis of relative-abundance but not in the analysis of absolute abundance. Conversely, points in the lower right corner indicate taxa that do not differ between diets in the analysis of relative abundance but do differ in the analysis of absolute abundance. *Akkermansia(g)* is an example of a microbe that appears to differ ($P = 6.49 \times 10^{-3}$, Kruskal-Wallis) between mice on the two diets in the relative-abundance analysis but not in the absolute-abundance analysis ($P = 3.37 \times 10^{-1}$, Kruskal-Wallis). *Lachnospiraceae(f)* showed the opposite trend; in the relative-abundance analysis it appears unchanged ($P = 6.31 \times 10^{-1}$, Kruskal-Wallis) but in the absolute-abundance analysis it differs ($P = 3.95 \times 10^{-3}$, Kruskal-Wallis) between the two diets. Neither of these analyses is wrong, they are simply asking two different questions**:** with relative data, the question is whether the percentage of that microbe is different between two conditions whereas with absolute data, the question is whether the abundance of that microbe is different between two conditions.

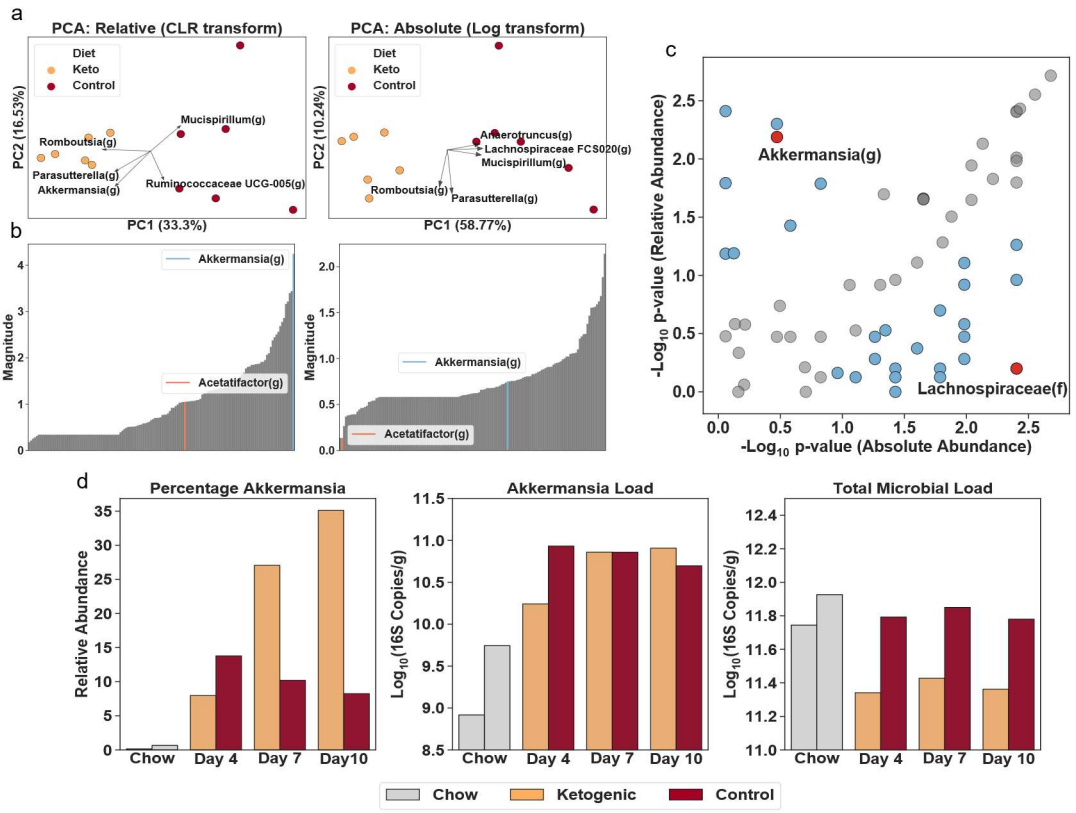


Figure 2.5: Analyses of relative and absolute microbial abundances from the same dataset result in different conclusions. (a) PCA on centered log-ratio transformed relative abundance data and log transformed absolute-abundance data (only the vectors of the five features with the largest magnitude are shown). (b) The impact of each taxon in the principal-component space (see text for details), with two taxa indicated to illustrate the comparison. (c) A comparison of the taxa determined to be significantly different between diets using relative versus absolute quantification (N = 6 mice per diet). *P*-values were determined by Kruskal-Wallis. Each point represents a single taxon; blue points indicate taxa with the absolute value of *P*-value ratios greater than 2.5; red points indicate two taxa that disagreed significantly between the relative and absolute analyses. (d) For illustrative purposes, a comparison of *Akkermansia(g)* relative abundance (percentage of Akkermansia), absolute abundance (Akkermansia load), and total microbial load between stool samples from one mouse on each diet (Ketogenic, orange; Control, red). Grey bars indicate loads prior to the diet switch when all mice were on the chow diet.

To explore one example of how different interpretations of how taxa differ between study conditions occur when using relative versus absolute abundance, we analyzed *Akkermansia(g)* in stool across each of the three time points on experimental diets (days 4, 7, and 10) and day 0 on chow diet. For simplicity in this illustration, we compared data from one mouse on each diet, but the trends hold on average between all mice on the two diets (Supplementary Figure 2.7). Analysis of relative microbial abundance demonstrated ~3X higher abundance of *Akkermansia(g)* in samples from the ketogenic compared with the control diet on days 7 and 10. However, when analyzing the difference in absolute abundance, more nuanced conclusions emerged. The rise in *Akkermansia(g)* results from switching mice from chow to experimental diets. The resulting *Akkermansia(g)* loads are similar in the two diets on days 7 and 10. However, the ketogenic diet reduces the total microbial load relative to both chow and control diets, therefore leading to the observed higher % of *Akkermansia(g)* in samples from mice on ketogenic diet.

Absolute abundances allow for quantitative differential taxon analysis

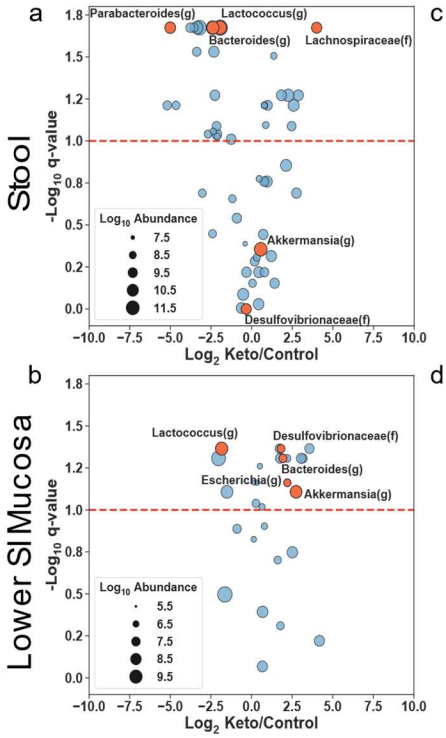
We next analyzed the absolute microbiota abundances in stool and lower small intestinal mucosa samples from day 10. A volcano plot, akin to those used in gene expression studies, was used to represent the overall changes in taxa abundances between the two diets, and the absolute abundance of each taxon was indicated by the size of its symbol (Fig. 2.6a). *P*-values from the Kruskal-Wallis tests were corrected for multiple hypothesis testing with the Benjamini–Hochberg method, resulting in q-values.[46, 47] A false discovery rate (FDR) of 10% was labeled on the volcano plot and q-values < 0.1 were used as a cutoff for designating differential taxa for downstream analyses. Comparisons between the two GI locations showed substantial differences in microbial response to diet by location. In stool, approximately 66% of the differential taxa were lower on the ketogenic diet vs the control diet whereas in the lower SI mucosa, > 80% of the differential taxa were more abundant in the ketogenic diet than control diet (Supplementary Table 2.4, Supplementary Table 2.5).

Next, we highlighted several specific differential taxa that were discordant between stool and lower SI mucosa. (1) *Bacteroides(g)* was lower on ketogenic diet in stool and higher on ketogenic diet in lower SI mucosa. This type of result could lead researchers who analyze stool samples to believe that lower levels of *Bacteroides(g)* may be associated with a phenotype when it could be the opposite if the phenotype is driven by the SI mucosal microbiota. (2) *Parabacteroides(g)* and *Lachnospiraceae GCA-900066575(g)* showed the highest fold changes (in opposite directions) in stool but were not detected in the lower SI mucosa. The opposite was observed for *Escherichia(g)*, which was more abundant in the ketogenic diet than the control diet in the lower SI mucosa but was not detected in stool. (3) *Akkermansia(g)* and *Desulfovibrionaceae(f)* were more abundant in the ketogenic diet than the control diet in the lower SI mucosa but were similar between the two diets in stool. Such microbes could have a relationship with phenotype through the small intestine but would be missed if only stool samples are analyzed.

A further breakdown of the differential taxa, using our quantitative limits of sequencing accuracy (defined earlier), allowed us to categorize four distinct scenarios that describe how microbes differed between GI locations of mice on the two diets. We refer to these four scenarios as "quantification classes" (Fig. 2.6b). First, there were microbes that were present in one diet and absent in the other ("presence/absence" class). For example, *Dorea(g),* in stool, and *Escherichia(g),* in SI mucosa, were absent from the control diet but present in the ketogenic diet. Second, there were microbes above the detection limit but below the quantitative limit in both diets ("no quant" class). For example, in stool, *Candidatus Soleaferrea(g)*, ranges in relative abundance from 0.002% to 0.025%, well below the 30% CV quantification threshold of 0.04% (as defined in Fig. 2.2d). Thus, we cannot quantitatively define the difference of this microbe between mice on the two diets. Third, microbes were above the detection limit in both diets but only above the quantitative limit in one of the diets ("semi-quant" class). For example, *Desulfovibrionaceae(f)* in the lower small-intestine mucosa was above the detection limit in mice on both diets but only above the quantitative limit in mice on the ketogenic-diet, so although we can be confident that a difference between the diets exists, we cannot be confident in our measurement of the magnitude of that difference. Fourth, microbes were found above the quantitative limits in

both diets ("quant" class). For example, for *Parabacteroides(g)* in stool, we can be confident in both the difference between the diets (it was more abundant in the control diet) and in the magnitude of that difference (a 32.2-fold difference). We have the lowest confidence in the measured absolute fold change of a taxon that is classified in the presence/absence class, and the greatest confidence in a taxon in the quant class.

Figure 2.6: Incorporating quantification limits enhances differential taxon analysis as shown in stool and SI mucosa. A quantitative framework that explicitly incorporates limits of quantification separates differential



c

Stool Differential Taxa (n=32, q-value < 0.1)	Fold Change (+ Keto, - Control)	Abundance (16S copies/g)	Quantification Class
Lachnospiraceae NK4A136(g)	6.0	$5.34x10^{9}$	Quant
Bacteroides(g)	-3.8	$3.40x10^{10}$	Quant
Faecalibaculum(g)	-3.8	$3.87x10^{11}$	Quant
Bifidobacterium(g)	-4.4	$1.42x10^{9}$	Quant
Lactococcus(g)	-5.2	$1.74x10^{10}$	Quant
Lactobacillus(g)	-9.3	$1.35x10^{11}$	Quant
Butyricimonas(g)	-9.8	$4.39x10^{9}$	Quant
Parabacteroides(g)	-32.2	$3.3x10^{9}$	Quant
Acetatifactor(g)	~ 5.5	$5.2x10^{8}$	Semi-Quant
Christensenellaceae(g)	~ -11.7	$3.69x10^{8}$	Semi-Quant
GCA-900066575(g)	~ 16	$1.94x10^{9}$	Semi-Quant
Dorea(g)	+ inf	$7.8x10^{7}$	Presence/Absence
cTPY-13(g)	- inf	$3.71x10^{8}$	Presence/Absence
Candidatus Soleaferrea(g)	--	~$6x10^{7}$	No Quant
Anaerotruncus(g)	--	~$2x10^{8}$	No Quant

d

Lower SI Mucosa Differential Taxa (n=18, q-value < 0.1)	Fold Change (+ Keto, - Control)	Abundance (16S copies/g)	Quantification Class
Akkermansia(g)	6.7	$2.57x10^{8}$	Quant
Muribaculaceae(g)	-2.8	$3.48x10^{8}$	Quant
Lactococcus(g)	-3.6	$4.31x10^{8}$	Quant
Lactobacillus(g)	-4.1	$3.70x10^{9}$	Quant
Lachnoclostridium(g)	~ 11.9	$1.64x10^{7}$	Semi-Quant
Lachnospiraceae(g)	~ 9	$9.60x10^{6}$	Semi-Quant
Bacteroides(g)	~ 3.8	$5.51x10^{6}$	Semi-Quant
Desulfovibrionaceae(f)	~ 3.5	$4.29x10^{6}$	Semi-Quant
Escherichia(g)	+ inf	$2.96x10^{6}$	Presence/Absence
Dorea(g)	+ inf	$2.94x10^{6}$	Presence/Absence
Ruminococcaceae(f)	--	~$8.49x10^{5}$	No Quant

microbial taxa into four classes, and for each GI location identifies a distinct set of differential taxa, including taxa with opposite patterns in stool and SI mucosa. (a-b) Microbial taxa in stool (a) or lower small-intestine (b) mucosa in mice on ketogenic (N = 6) and control (N = 6) diets. The fold change on the x-axis is the Log_2 ratio of the average absolute loads of taxon loads in each diet. Negative values indicate lower loads in ketogenic diet compared to control diet. The q-value for a taxon indicates the significance of the difference in absolute abundances between the two diets and were obtained by Kruskal-Wallis with a Benjamini–Hochberg correction for multiple hypothesis testing. The Log_{10} absolute abundance of each taxon is indicated by circle size. Orange circles indicate taxa discussed in the main text including taxa that show discordant fold changes between stool and lower SI mucosa. The red dashed line is shown at a q-value representing a 10% false-discovery rate. (c-d) A subset of taxa from stool (c) and lower SI mucosa (d) that were significantly different between diets (q-values < 0.1) and their corresponding fold change, absolute abundance (larger of the average absolute abundances between the two diets), and quantification class. Quantification class is determined by whether one or both measurements were above or below the lower limit of quantification and the limit of detection.

Discussion

In this study, we have shown that this technology performs across biogeographically diverse samples with microbial loads spanning over 6 orders of magnitude. Our lower limits of quantification for total microbial load from lumenal (e.g., stool, cecum contents) and mucosal samples were 4.2 x 10^5 16S rRNA gene copies/g and 1.0 x 10^7 16S rRNA gene copies/g

respectively. These lower limits were mainly restricted by the column-based extractions used which require < 200 mg of sample input for lumenal contents and < 8 mg of input for mucosal samples. This sample input is limited by the high concentration of PCR inhibitors and host DNA in these samples. New sample-processing methods that deplete host DNA before extraction (e.g. the use of propidium monoazide (PMA) or saponin with DNase)[48, 49] could help improve the quantitative limits in samples with high levels of host DNA (e.g., mucosa) by removing non-microbial DNA before extraction. Such host-depletion methods could also improve performance of other current or future methods of quantitative sequencing. Before these methods are introduced into quantitative sequencing protocols, they will require extensive validation to understand the impacts host DNA depletion has on the microbial load and composition of these samples, which will affect the accuracy of any absolute-abundance technique. We showed that the precision of any individual taxon's abundance can be defined as a function of that taxon's relative abundance and the sequencing depth. These accuracy thresholds generally state that all taxa with relative abundance > 0.01% have a maximum %CV of 30%. We did not quite reach the theoretical limit of Poisson precision (Fig. 2.2c), which might be explained by slight differences in PCR amplification between high- and low-abundance microbes, and could potentially be corrected with single-molecule counting techniques utilizing unique molecular identifiers (UMIs).[50, 51] Interestingly, the precision of these abundance measurements did not differ between high input DNA samples (1.2×10^7 16S rRNA gene copies) and low-input DNA samples (1.2×10^4 16S rRNA gene copies), even though the low-input sample required 10 additional PCR cycles. The lack of an increase in observed chimeric sequences in the low-input sample indicates that PCR bias from chimera generation may occur mainly during over-amplification; thus, we suggest monitoring library-prep amplification reactions with qPCR and stopping reactions during the late exponential phase.

Our quantitative sequencing method, as validated, is subject to some of the same limitations of general 16S rRNA gene amplicon sequencing. Primarily, the accuracy of any given taxon's abundance is believed to be impacted by amplification bias. We showed that the abundances of *Akkermansia muciniphila(s)*, *Lachnospiraceae(f)*, *Bacteroidales(o)*, and *Lactobacillaceae(f)* could be quantified with similar precision (2X), but different accuracy, i.e. *Akkermansia muciniphila(s)* abundance was ~2.5X higher in the quantitative-sequencing estimate compared with the estimate from dPCR with taxa-specific primers. This offset was consistent between samples, indicating that it may be related to differences in primer coverage between the taxon-specific primer set and the universal primer set used in this study. Nevertheless, such offsets should be similar if the same library-prep conditions are used, so one can reliably compare taxa among groups or studies and the use of UMIs may further eliminate any potential amplification biases. We note that dPCR-based total microbial load measurements should be more robust to amplification biases of individual taxa. Additionally, the total microbial load measurement will be affected by the 16S rRNA primer set chosen and its respective taxonomic coverage. The primers in this study were chosen to have broad coverage and also to limit amplification of host mitochondrial DNA,[31-33] to ensure proper quantification of mucosal and small-intestine samples with high host DNA loads. Finally, to take full advantage of the power of this quantitative framework, study designs must incorporate proper sampling techniques to address spatiotemporal variation in microbial abundances.[22]

A method-specific limitation is the requirement of an additional step, dPCR total microbial load quantification, which consumes a portion of the extracted DNA sample. This limitation is minor because dPCR generally requires at least 100 copies for a measurement with a ~10% Poisson error, which is much less than the roughly 10,000 copies required for sequencing. Additionally, the absolute abundances are reported in 16S rRNA gene copies/g and require conversion to number of cells/g, which has standard limitations (e.g., the completeness of rRNA databases and copy-number variation among similar species). However, when comparing taxa across study groups, the 16S rRNA gene copies per taxonomic group should be similar. Finally, this method was only validated for 16S rRNA gene amplicon sequencing; thus, further validation would be required for applying this method to converting metagenomic sequencing from relative to absolute quantification. We were not able to directly compare our measurements to other absolute abundance techniques discussed in the Introduction because these techniques have not been validated on the broad range of sample types and microbial loads tested here (Supplementary Table 2.2). A fair side-by-side comparison would require re-optimization of the other techniques for complex sample types, like those with high host DNA levels and low microbial biomass (e.g., mucosa).

We applied the quantitative framework to a murine ketogenic-diet study to identify how microbial taxa at several GI locations respond to diet. Because total microbial loads were lower in the ketogenic diet compared to the control, analysis of absolute abundance was required to correctly identify differential taxa. The lower load observed on the ketogenic diet can likely be explained by its lower fiber and carbohydrate content, as these dietary components are main substrates for many gut microbes.[52] Many factors (including diet) that induce changes in relative microbial abundances can also impact total microbial load.[25, 53] Even among healthy mice on the same (chow) diet, total microbial loads in stool can differ by 10 times.[25] Such variation in total microbial load likely contributes to the noise in microbiome studies. Another insight of this study was that we found different patterns in the microbial communities at each GI sampling site. For example, *Akkermansia(g)* loads did not differ between diets in stool, but they were significantly greater in the small-intestine mucosa in the ketogenic diet compared with the control. *Bacteroides(g)* load was lower in stool and greater in the small-intestine mucosa in the ketogenic relative to the control diet. Clearly, differential taxa at one GI location cannot be used as a proxy for measuring differential taxa at another GI location. To our knowledge, this is the first microbiome study to show that microbial taxa in the small intestine and the stool can change in different directions and by different magnitudes in response to diet. Furthermore, for each taxon, this method enables a comparison of absolute microbial abundance to limits of detection and quantification. This comparison separates differential taxa into four classes (Quant, Semi-Quant, No Quant, Presence/Absence) which provide a convenient shortcut for more quantitative interpretation of microbiome studies. It should be noted that the absence of a microbe in a dataset is a factor of the sequencing depth, and just because a microbe is not found in the sequencing data does not mean it is not in the sample. However, with absolute anchoring, one can confidently say that when a microbe is not found, that microbe is below a given abundance.

We have not focused on correlations among taxa in this dataset. However, the absolute abundance measurements acquired using our method should help overcome many of the limitations of correlation-based analyses on relative abundances[54, 55] and enable analyses

using standard methodologies like Spearman's rank correlation (Supplementary Figure 2.8). However, further work will be required to properly address the impact that correlations between total microbial loads will have on taxon-based correlation networks. In addition, new statistical and/or experimental design methods may be required for interpreting the correlations between a taxon's presence and/or total load and observed phenotypes.

This method overcomes three bottlenecks to wider adoption of absolute quantitative measurements in microbiome analysis: (i) performance across samples with a wide range of microbial loads; (ii) performance across biogeographically diverse sample types (iii) explicit evaluation of limits of quantification of the method. This method will be useful in other areas that benefit from quantitative analysis, such as monitoring microbial communities during manufacturing of complex probiotic mixtures[56] and monitoring changes of host-associated microbial communities over time (e.g. in health, aging and development, disease progression, and during probiotic or other treatments). Applying absolute quantification[19-21, 23-28, 32, 33] of microbial taxa to biogeographically relevant GI locations will provide researchers with new insights in how microbial communities affect host phenotypes.

Methods

Mice

All animal husbandry and experiments were approved by the Caltech Institutional Animal Care and Use Committee (IACUC protocols #1646 and #1769). Male and female germ free (GF) C57BL/6J mice were bred in the Animal Research Facility at Caltech, and 4-week-old female specific-pathogen-free (SPF) Swiss Webster mice were obtained from Taconic Farms (Germantown, NY, USA). Mice were housed on heat-treated hardwood chip bedding (Aspen Chip Bedding, Northeastern Products, Warrensburg, NY, USA) and provided with tissue paper (Kleenex, Kimberly-Clark, Irving, TX, USA) nesting material. Experimental animals were fed standard chow (Lab Diet 5010), 6:1 ketogenic diet (Envigo TD.07797, Indianapolis, IN, USA; Supplementary Table 2.3) or vitamin- and mineral-matched control diet (Envigo TD.150300; Supplementary Table 2.3). Diet design and experimental setup were taken from a recently published study.[40] To minimize cage effects, mice were housed two per cage with three cages per diet group. Custom feeders, tin containers approximately 2.5 inches tall with a 1-inch diameter hole in the top, were used for the ketogenic diet as it is a paste at room temperature. Autoclaved water was provided ad libitum and cages were subjected to a daily 13:11 light:dark cycle throughout the study. Mice were euthanized via CO_2 inhalation as approved by the Caltech IACUC in accordance with the American Veterinary Medical Association Guidelines on Euthanasia.[57]

Microbial Samples

The mock microbial community (Zymobiomics Microbial Community Standard; D6300) was obtained from Zymo Research (Irvine, CA, USA). This community is stored in DNA/RNA Shield, which could interfere with extraction efficiency at high concentrations. We found that a 100 µL input of a 10X dilution of the microbial community stock is the maximum input that the Qiagen DNeasy Powersoil Pro Kit can handle without recovery losses. Negative control blanks were also used which included 100 µL of nuclease free water instead of mock community.

Fresh stool samples were collected immediately after defecation from individual mice and all collection occurred at approximately the same time of day. For intestinal samples, the GIT was excised from the stomach to the anus. Contents from each region of the intestine (stomach, upper half of SI, lower half of SI, cecum, and colon) were collected by longitudinally opening each segment with a scalpel and removing the content with forceps. Terminal colonic pellets are referred to as stool. After contents were removed the intestinal tissue was washed by vigorously shaking in cold sterile saline. The washed tissue was placed in a sterile petri dish and then dabbed dry with a Kimwipe (VWR, Brisbane, CA, USA) before scraping the surface of the tissue with a sterile glass slide. These scrapings were collected as the mucosa samples. All samples were stored at -80 °C after cleaning and before extraction of DNA.

DNA Extraction

DNA was extracted from all samples by following the Qiagen DNeasy Powersoil Pro Kit protocol (Qiagen; Valencia, CA, USA). Bead-beating was performed with a Mini-BeadBeater (BioSpec, Bartlesville, OK, USA) for 4 min. To ensure extraction columns were not overloaded, we used ~10 mg of scrapings and ~50 mg of contents. Half of the lysed volume was loaded onto the column and elution volume was 100 µL. Nanodrop (NanoDrop 2000, ThermoFisher Scientific) measurements were performed with 2 µL of extracted DNA to ensure concentrations were not close to the extraction column maximum binding capacity (20 µg).

Absolute Abundance

The concentration of total 16S rRNA gene copies per sample was measured using the Bio-Rad QX200 droplet dPCR system (Bio-Rad Laboratories, Hercules, CA, USA). The concentration of the components in the dPCR mix used in this study were as follows: 1x EvaGreen Droplet Generation Mix (Bio-Rad), 500 nM forward primer, and 500 nM reverse primer. Universal primers to calculate the total 16S rRNA gene concentrations were a modification to the standard 515F-806R primers[4] to reduce host mitochondrial rRNA gene amplification in mucosal and small-intestine samples (Supplementary Table 2.6).[31-33] Thermocycling for universal primers was performed as follows: 95 °C for 5 min, 40 cycles of 95 °C for 30 s, 52 °C for 30 s, and 68 °C for 30 s, with a dye stabilization step of 4 °C for 5 min and 90 °C for 5 min. All ramp rates were 2 °C per second. The concentration of taxon-specific gene copies per sample was measured using a similar dPCR protocol, except with different annealing temperatures. Annealing temperatures during thermocycling for taxa-specific primers can be found in Supplementary Table 2.6. The concentration of the components in the qPCR mix used in this study were as follows: 1x SsoFast EvaGreen Supermix (BioRad), 500 nM forward primer, and 500 nM reverse primer. Thermocycling was performed as follows: 95°C for 3 min, 40 cycles of 95 °C for 15 s, 52 °C for 30 s, and 68 °C for 30 s. All dPCR measurements are single replicates.

Concentrations of 16S rRNA gene per microliter of extraction were corrected for elution volume and losses during extraction before normalizing to the input sample mass (Equation 1).

$$\text{Microbial Load} = \text{dPCR concentration} * \text{elution volume} * \frac{\text{dead volume}}{\text{extraction volume}} * \frac{1}{\text{sample mass}} \quad (1)$$

Absolute abundance of individual taxa was calculated either by dPCR with taxa-specific primers or multiplying the total microbial load from Equation 1 by the relative abundance from 16S rRNA gene amplicon sequencing.

16S rRNA Gene Amplicon Sequencing

Extracted DNA was amplified and sequenced using barcoded universal primers and protocol modified to reduce amplification of host DNA[31-33]. The variable 4 (V4) region of the 16S rRNA gene was amplified in triplicate with the following PCR reaction components: 1X 5Prime Hotstart mastermix, 1X Evagreen, 500 nM forward and reverse primers. Input template concentration varied. Amplification was monitored in a CFX96 RT-PCR machine (Bio-Rad) and samples were removed once fluorescence measurements reached ~10,000 RFU (late exponential phase). Cycling conditions were as follows: 94 °C for 3 min, up to 40 cycles of 94 °C for 45 s, 54 °C for 60 s, and 72 °C for 90 s. Triplicate reactions that amplified were pooled together and quantified with Kapa library quantification kit (Kapa Biosystems, KK4824, Wilmington, MA, USA) before equimolar sample mixing. Libraries were concentrated and cleaned using AMPureXP beads (Beckman Coulter, Brea, CA, USA). The final library was quantified using a High Sensitivity D1000 Tapestation Chip. Sequencing was performed by Fulgent Genetics (Temple City, CA, USA) using the Illumina MiSeq platform and 2x300bp reagent kit for paired-end sequencing.

16S rRNA Gene Amplicon Data Processing

Processing of all sequencing data was performed using QIIME 2 2019.1.[58] Raw sequence data were demultiplexed and quality filtered using the q2-demux plugin followed by denoising with DADA2.[59] Chimeric read count estimates were estimated using DADA2. Beta-diversity metrics (Aitchison distance,[9] Bray-Curtis Dissimilarity) were estimated using the q2-diversity plugin after samples were rarefied to the maximum number of sequences in each of the relevant samples. Rarefaction was used to force zeros in the dataset to have the same probability (across samples) of arising from the taxon being at an abundance below the limit of detection. Although rarefaction may lower the statistical power of a dataset[60] it helps decrease biases caused by different sequencing depths across samples.[12] Taxonomy was assigned to amplicon sequence variants (ASVs) using the q2-feature-classifier[61] *classify-sklearn* naïve Bayes taxonomy classifier against the Silva[62] 132 99% OTUs references from the 515F/806R region. All datasets were collapsed to the genus level before downstream analyses. All downstream analyses were performed in IPython primarily through use of the *Pandas*, *Numpy* and *Scikit-learn* libraries.

Data Transforms and Dimensionality Reduction

For dimensionality reduction techniques requiring a log transform, a pseudo-count of 1 read was added to all taxa. With relative abundance data, the centered log-ratio transform was used (Equation 2) to handle compositional effects whereas a log transform was applied to the absolute-abundance data to handle heteroscedasticity in the data.

$$x_{\text{clr}} = \left[\log\left(\frac{x_1}{G(x)}\right), \log\left(\frac{x_2}{G(X)}\right), \ldots, \log\left(\frac{x_D}{G(X)}\right)\right] \quad where \quad G(X) = \sqrt[D]{x_1 * x_2 * \ldots * x_D} \qquad (2)$$

For comparative purposes, principal co-ordinates analysis (PCoA) was also performed using the Bray-Curtis dissimilarity metric. Principal component analysis (PCA) and PCoA were performed using *scikit-learn* decomposition methods. Feature loadings for each principal component were calculated by multiplying each eigenvector by the square root of its corresponding eigenvalue. All data were visualized using matplotlib and seaborn.

Taxa Limits of Quantification

Poisson confidence intervals were calculated by bootstrapping Poisson samples for rate parameters across the percentage abundance range (0–1) corresponding to either the input DNA copies or number of reads. We took 10^4 bootstrap replicates with a Poisson sample size of 4 to match the number of replicates we sequenced. The %CV for each replicate was calculated and the middle 95th percentile was shown as the confidence interval.

Thresholds for percentage abundance were calculated by first fitting a negative exponential curve $y = ax^{-b}$ to the plot of %CV versus percentage abundance using SciPy. Then the percentage abundance at a given %CV threshold was determined. This process was repeated after subsampling the data at decreasing read depths to find the relationship between percent abundance accuracy limits at sequencing depth.

Measurement Uncertainty

When measuring the absolute abundance of a given taxon in a sample, many factors contribute to the uncertainty of the measurement. Two primary factors, extraction efficiency and average amplification efficiency for each taxon, should be equivalent for each taxon across samples processed under identical conditions and thus neither should impact the discovery of differential taxa. However, other factors contributing to the uncertainty of an absolute-abundance measurement vary among samples and can impact the discovery of differential taxa. At least six independent errors can contribute to the overall uncertainty of a taxon's absolute abundance: (i) extraction error (ii) the Poisson sampling error of dPCR, (iii) the Poisson sampling error of sample input into an amplification reaction to make a sequencing library, (iv) the uncertainty in the amplification rates among sequences, (v) the Poisson sampling error of the sequencing machine, and (vi) the uncertainty in taxonomic assignment resulting from different software programs that differ in how they convert raw sequencing reads to a table of read counts per taxon.

To measure the total error in our absolute-abundance measurements, we compared the true absolute load value of four "representative" taxa (taxa that are common gut flora from different taxonomic ranks) as measured by taxa-specific dPCR, with the value obtained from our method of quantitative sequencing with dPCR anchoring (Fig. 2.3b) and then analyzed the relative error in these measurements, defined as the $\log_2$ of the observed taxon load over the true taxon load. We constructed a quantile-quantile (Q–Q) plot (Supplementary Figure 2.9) of the mean-centered $\log_2$ relative errors and found that the errors appeared normally distributed. We confirmed this by running a Shapiro–Wilk test (P-value = 0.272) on the mean-centered $\log_2$ relative errors, which uses a null hypothesis that the dataset comes from a normal distribution. The standard deviation of the mean-centered $\log_2$ relative errors was 0.48, which results in a 95% confidence interval of ~(-1,1), indicating a 2x precision on each individual measurement. However, as seen with *Akkermansia(g)* (Fig. 2.3b), accuracy offsets may exist for specific taxa. It is important to note that all samples used in this analysis had relative abundances above the 50% CV threshold defined in Fig. 2.2d and thus we do not

make any conclusions about the precision of absolute abundance measurements for taxa with relative abundances below the 50% CV threshold.

Biological Uncertainty and Statistical Inference Methods

When measuring the absolute abundance of a taxon from a defined population (e.g., healthy adults, mice on a ketogenic diet) it is unlikely this abundance comes from a well-defined statistical distribution. Given this inherent limitation, we used non-parametric statistical tests, which do not rely on distributional assumptions, for our differential abundance analyses.

Statistical comparisons between diet groups were analyzed using the Kruskal–Wallis[46] rank sums test with Benjamini–Hochberg[47] multiple hypothesis testing correction. All statistical tests were implemented using *SciPy.stats Kruskal* function and *statsmodels.stats.multitest multipletests* function with the *fdr_bh* option for Benjamini-Hochberg multiple-testing correction. When calculating differentially abundant taxa, only taxa present in at least 4 out of 6 mice in a group were considered to remove fold-change outliers when plotting (Fig. 2.6a-b).

Correlation Analysis

Samples were separated by diet (ketogenic and control) and only stool samples were used (days 4, 7, and 10). The total microbial load and top 30 taxa with the highest average absolute abundance were selected for analysis. Spearman's rank correlation coefficient and corresponding *P*-values were calculated for all pairwise interactions using the *scipy.stats.spearmanr* function. Benjamini–Hochberg procedure was to calculate q-values, which account for multiple hypothesis testing. A heatmap of the diagonal correlation matrix was plotted (Supplementary Figure 2.8) for q-values <10% FDR.

Data Availability

The complete sequencing data generated during this study are available in the National Center for Biotechnology Information Sequence Read Archive repository under study accession number PRJNA575097. Raw data for all figures available through CaltechDATA: https://data.caltech.edu/records/1371. Raw data for all figures is also provided as source data files.

Acknowledgements

This work was supported in part by the Kenneth Rainin Foundation (2018-1207), the Army Research Office (ARO) Multidisciplinary University Research Initiative (MURI #W911NF-17-1-0402), and a National Institutes of Health Biotechnology Leadership Pre-doctoral Training Program (BLP) fellowship from Caltech's Donna and Benjamin M. Rosen Bioengineering Center (T32GM112592, to J.T.B.). We thank Elaine Hsiao and Christine Olson for helpful discussions and input on the experimental design and diets; we thank the Caltech Bioinformatics Resource Center for assistance with statistical analyses; we acknowledge the Caltech animal facility for experimental resources; we thank the Caltech Office of Laboratory Animal Resources and the veterinary technicians at Caltech for technical support; and Natasha Shelby for contributions to writing and editing this manuscript.

References

1. Schirmer, M., Denson, L., Vlamakis, H., Franzosa, E.A., Thomas, S., Gotman, N.M., Rufo, P., Baker, S.S., Sauer, C., Markowitz, J., Pfefferkorn, M., Oliva-Hemker, M., Rosh, J., Otley, A., Boyle, B., Mack, D., Baldassano, R., Keljo, D., LeLeiko, N., Heyman, M., Griffiths, A., Patel, A.S., Noe, J., Kugathasan, S., Walters, T., Huttenhower, C., Hyams, J. & Xavier, R.J. Compositional and Temporal Changes in the Gut Microbiome of Pediatric Ulcerative Colitis Patients Are Linked to Disease Course. *Cell Host Microbe* **24**, 600-610.e604 (2018).

2. Routy, B., Le Chatelier, E., Derosa, L., Duong, C.P.M., Alou, M.T., et al. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91-97 (2018).

3. Sharon, G., Cruz, N.J., Kang, D.W., Gandal, M.J., Wang, B., et al. Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell* **177**, 1600-1618.e1617 (2019).

4. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *P. Natl. Acad. Sci. USA* **108**, 4516 (2011).

5. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200 (2013).

6. Xia, Y. & Sun, J. Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases* **4**, 138-148 (2017).

7. Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410-422 (2018).

8. Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. & Egozcue, J.J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8** (2017).

9. Aitchison, J. The Statistical Analysis of Compositional Data. *J. Roy. Stat. Soc. B Met.* **44**, 139-160 (1982).

10. Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K. & Knight, R. Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* **10**, 2719 (2019).

11. Martino, C., Morton, J.T., Marotz, C.A., Thompson, L.R., Tripathi, A., Knight, R. & Zengler, K. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **4**, e00016-00019 (2019).

12. Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).

13. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210-221 (2017).

14. Tsilimigras, M.C. & Fodor, A.A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* **26**, 330-335 (2016).

15. Gloor, G.B. & Reid, G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* **62**, 692-703 (2016).

16. Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G. & Gloor, G.B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS One* **8**, e67019 (2013).

17. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R. & Peddada, S.D. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).

18. Morton, J.T., Sanders, J., Quinn, R.A., McDonald, D., Gonzalez, A., et al. Balance Trees Reveal Microbial Niche Differentiation. *mSystems* **2**, e00162-00116 (2017).

19. Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A. & Spang, R. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, 28 (2016).

20. Tkacz, A., Hortala, M. & Poole, P.S. Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* **6**, 110 (2018).

21. Tourlousse, D.M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N. & Sekiguchi, Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* **45**, e23-e23 (2016).

22. Ji, B.W., Sheth, R.U., Dixit, P.D., Wang, H.H. & Vitkup, D. Quantifying spatiotemporal dynamics and noise in absolute microbiota abundances using replicate sampling. *Nat. Methods*, **16**,pages731–736 (2019).

23. Hardwick, S.A., Chen, W.Y., Wong, T., Kanakamedala, B.S., Deveson, I.W., Ongley, S.E., Santini, N.S., Marcellin, E., Smith, M.A., Nielsen, L.K., Lovelock, C.E., Neilan, B.A. & Mercer, T.R. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nature Commun.* **9**, 3096 (2018).

24. Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G. & Raes, J. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507 (2017).

25. Contijoch, E.J., Britton, G.J., Yang, C., Mogno, I., Li, Z., et al. Gut microbiota density influences host physiology and is shaped by host and microbial factors. *eLife* **8**, e40553 (2019).

26. Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiome profiling. *PLoS One*, **15**, e0227285(2020).

27. Lou, J., Yang, L., Wang, H., Wu, L. & Xu, J. Assessing soil bacterial community and dynamics by integrated high-throughput absolute abundance quantification. *PeerJ* **6**, e4514-e4514 (2018).

28. Kleyer, H., Tecon, R. & Or, D. Resolving Species Level Changes in a Representative Soil Bacterial Community Using Microfluidic Quantitative PCR. *Front. Microbiol.* **8**, 2017 (2017).

29. Brankatschk, R., Bodenhausen, N., Zeyer, J. & Burgmann, H. Simple absolute quantification method correcting for quantitative PCR efficiency variations for microbial community samples. *Appl. Environ. Microbiol.* **78**, 4481-4489 (2012).

30. Suzuki, M.T. & Giovannoni, S.J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625-630 (1996).

31. Bogatyrev, S.R. & Ismagilov, R.F. Quantitative microbiome profiling in lumenal and tissue samples with broad coverage and dynamic range via a single-step 16S rRNA gene DNA copy quantification and amplicon barcoding. *Preprint at:* https://www.biorxiv.org/content/10.1101/2020.01.22.914705v1 (2020).

32. Bogatyrev, S.R. Development of Analytical Tools and Animal Models for Studies of Small-Intestine Dysbiosis. *Dissertation (Ph.D.), California Institute of Technology*, doi:10.7907/VJDZ-7B52 (2020).

33. Bogatyrev, S.R., Rolando, J.C. & Ismagilov, R.F. Self-reinoculation with fecal flora changes microbiota density and composition leading to an altered bile-acid profile in the mouse small intestine. *Microbiome*, **8**. doi: 10.1186/s40168-020-0785-4 (2020).

34. Shen, F., Du, W., Kreutz, J.E., Fok, A. & Ismagilov, R.F. Digital PCR on a SlipChip. *Lab Chip* **10**, 2666-2672 (2010).

35. Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., et al. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. *Anal. Chem.* **83**, 8604-8610 (2011).

36. Sanders, R., Huggett, J.F., Bushell, C.A., Cowen, S., Scott, D.J. & Foy, C.A. Evaluation of Digital PCR for Absolute DNA Quantification. *Anal. Chem.* **83**, 6474-6484 (2011).

37. Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B. & Chiodini, R.J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 24 (2016).

38. Caruso, V., Song, X., Asquith, M. & Karstens, L. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems* **4**, e00163-00118 (2019).

39. Wen, C., Wu, L., Qin, Y., Van Nostrand, J.D., Ning, D., Sun, B., Xue, K., Liu, F., Deng, Y., Liang, Y. & Zhou, J. Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS One* **12**, e0176716 (2017).

40. Olson, C.A., Vuong, H.E., Yano, J.M., Liang, Q.Y., Nusbaum, D.J. & Hsiao, E.Y. The Gut Microbiota Mediates the Anti-Seizure Effects of the Ketogenic Diet. *Cell* **173**, 1728-1741.e1713 (2018).

41. Newell, C., Bomhof, M.R., Reimer, R.A., Hittel, D.S., Rho, J.M. & Shearer, J. Ketogenic diet modifies the gut microbiota in a murine model of autism spectrum disorder. *Mol. Autism* **7**, 37 (2016).

42. Klein, M.S., Newell, C., Bomhof, M.R., Reimer, R.A., Hittel, D.S., Rho, J.M., Vogel, H.J. & Shearer, J. Metabolomic Modeling To Monitor Host Responsiveness to Gut Microbiota Manipulation in the BTBRT+tf/j Mouse. *J. Proteome Res.* **15**, 1143-1150 (2016).

43. Polz, M.F. & Cavanaugh, C.M. Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64**, 3724-3730 (1998).

44. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M.F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966-8969 (2005).

45. Donaldson, G.P., Lee, S.M. & Mazmanian, S.K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20-32 (2016).

46. Kruskal, W.H. & Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583-621 (1952).

47. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B Met.* **57**, 289-300 (1995).

48. Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R. & Zengler, K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42-42 (2018).

49. Zelenin, S., Hansson, J., Ardabili, S., Ramachandraiah, H., Brismar, H. & Russom, A. Microfluidic-based isolation of bacteria from whole blood for sepsis diagnostics. *Biotechnol. Lett.* **37**, 825-830 (2015).

50. Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., Rosenbaum, M. & Gordon, J.I. The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).

51. Hoshino, T. & Inagaki, F. Application of Stochastic Labeling with Random-Sequence Barcodes for Simultaneous Quantification and Sequencing of Environmental 16S rRNA Genes. *PLoS One* **12**, e0169431 (2017).

52. Carmody, Rachel N., Gerber, Georg K., Luevano, Jesus M., Gatti, Daniel M., Somes, L., Svenson, Karen L. & Turnbaugh, Peter J. Diet Dominates Host Genotype in Shaping the Murine Gut Microbiota. *Cell Host Microbe* **17**, 72-84 (2015).

53. Faith, J.J., McNulty, N.P., Rey, F.E. & Gordon, J.I. Predicting a Human Gut Microbiota's Response to Diet in Gnotobiotic Mice. *Science* **333**, 101 (2011).

54. Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J. & Bonneau, R.A. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Comp. Biol.* **11**, e1004226 (2015).

55. Friedman, J. & Alm, E.J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comp. Biol.* **8**, e1002687 (2012).

56. Tanoue, T., Morita, S., Plichta, D.R., Skelly, A.N., Suda, W., Sugiura, Y., Narushima, S., Vlamakis, H., Motoo, I., Sugita, K., Shiota, A., Takeshita, K., Yasuma-Mitobe, K., Riethmacher, D., Kaisho, T., Norman, J.M., Mucida, D., Suematsu, M., Yaguchi, T., Bucci, V., Inoue, T., Kawakami, Y., Olle, B., Roberts, B., Hattori, M., Xavier, R.J., Atarashi, K. & Honda, K. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600-605 (2019).

57. Leary, S., Underwood, W., Anthony, R. & Cartner, S. AVMA Guidelines for the Euthanasia of Animals. (2013).

58. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *Nat. Biotechnol.* **37**, pages852–857 (2019).

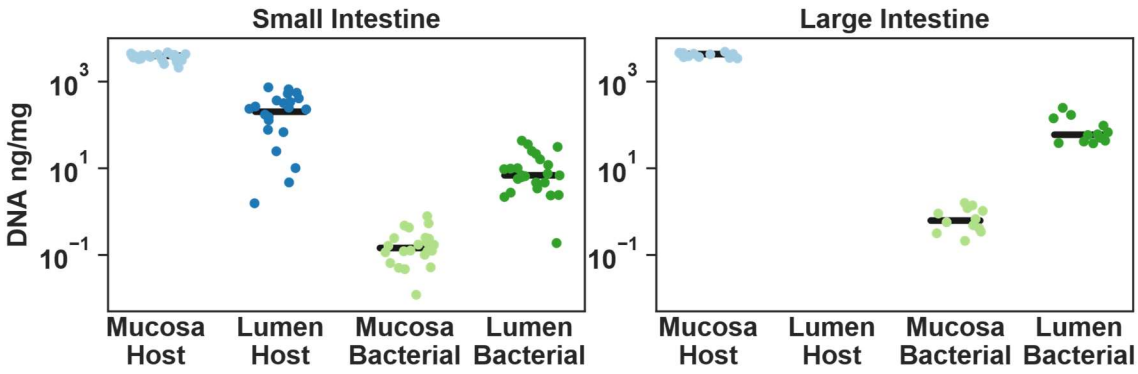
59. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581-583 (2016).

60. McMurdie, P.J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comp. Biol.* **10**, e1003531 (2014).

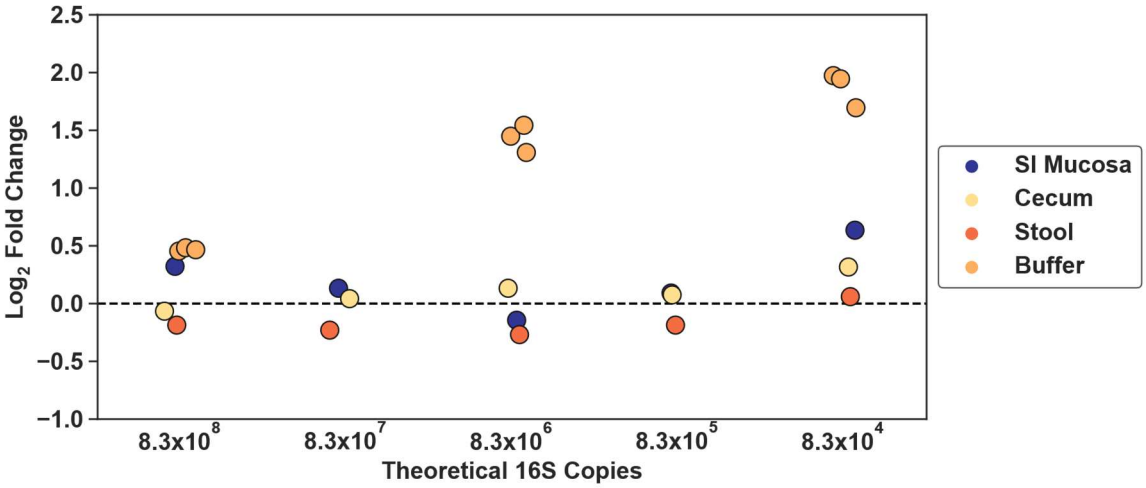
61. Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A. & Gregory Caporaso, J. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90-90 (2018).

62. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F.O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-D596 (2013).

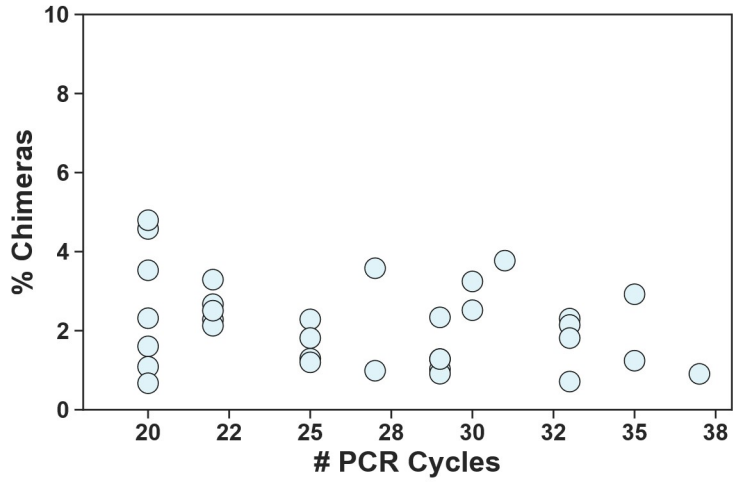
Supplementary Information



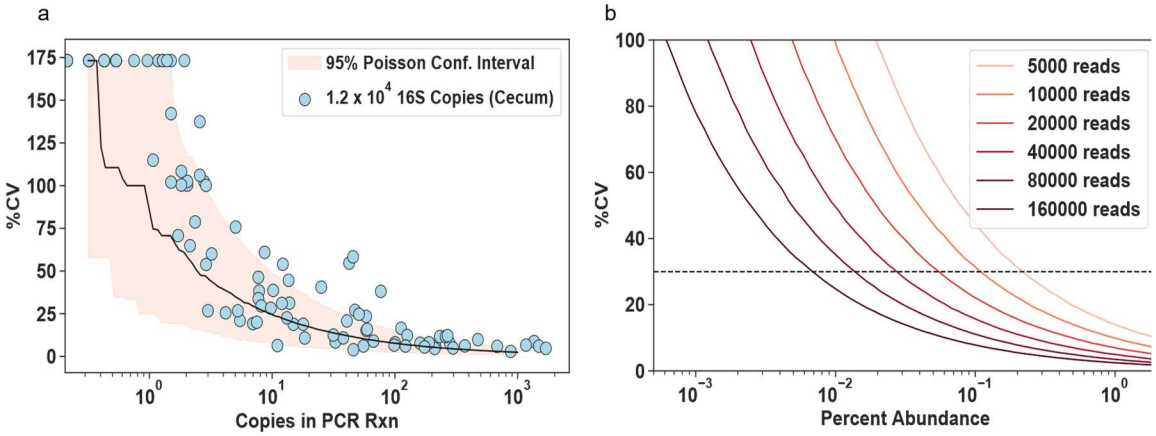
Supplementary Figure 2.1: Total DNA loads in small intestine and large intestine mucosa and lumen. Extracted DNA samples from mice in the ketogenic-diet group were measured by Nanodrop (total DNA) and digital PCR (microbial DNA). The horizontal lines represent the means and the points represent individual biological replicates (N = 24 for small intestine; N = 12 for large intestine).



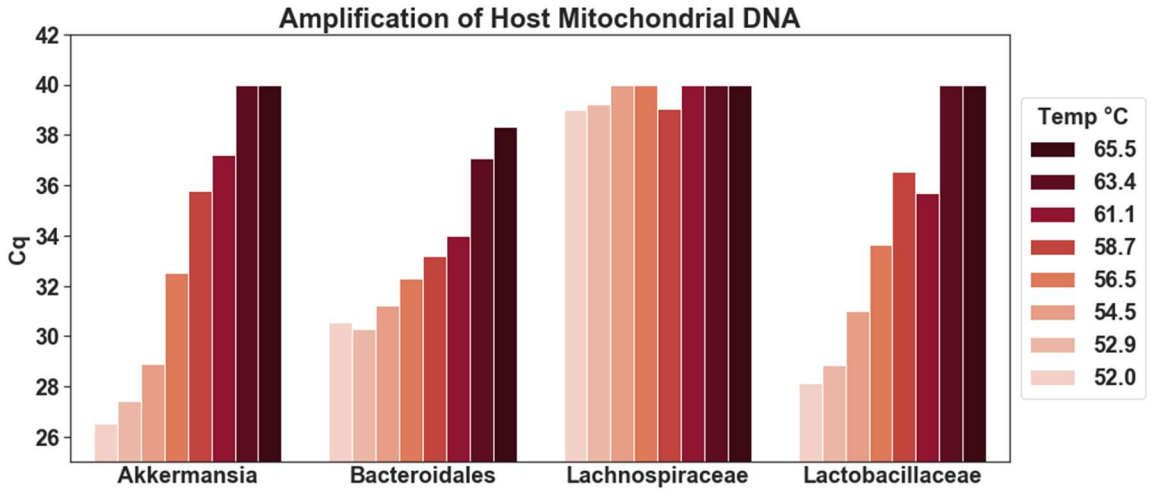
Supplementary Figure 2.2: Extraction and total DNA measurement accuracy of an eight-member mock microbial community dilutions spiked into extraction buffer or small-intestine mucosa, cecum, or stool from germ free mice. Log_2 fold change between theoretical and dPCR measured copies of 16S rRNA gene after extraction with varying input levels. Three technical replicates for buffer extractions are shown. All other sample types shown are N = 1 to illustrate the biological noise among sample types.



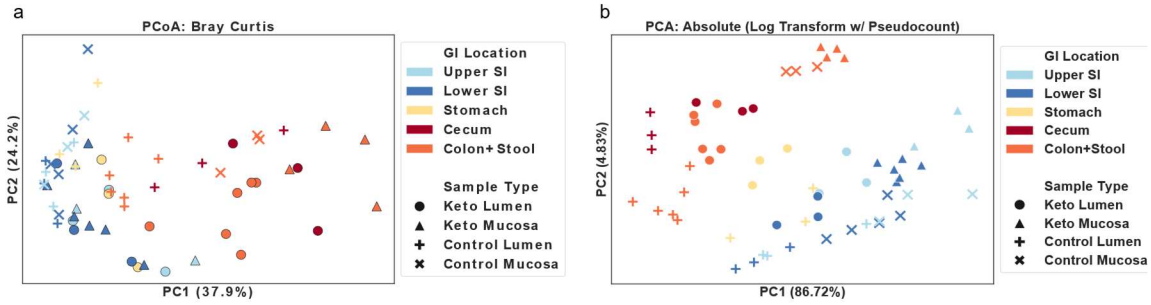
Supplementary Figure 2.3: Chimeric sequence prevalence is not determined by the number of PCR cycles. Relationship between the number of PCR cycles during the amplification reaction for library prep and the percentage of chimeric sequences detected by Divisive Amplicon Denoising Algorithm 2 (DADA2).[1] N = 33 samples that were sequenced from mice in the ketogenic-diet group.



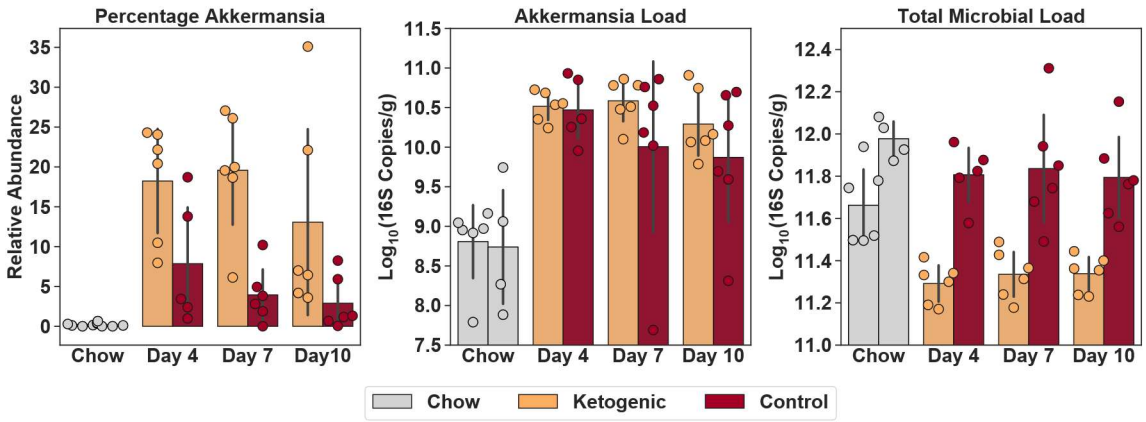
Supplementary Figure 2.4: Poisson limits of sequencing accuracy. (a) Relationship between the relative abundance of each taxon and % coefficient of variation (CV) using four technical (sequencing) replicates of a mouse cecum sample with an initial template input of 1.2 x 10^4 16S rRNA gene copies. The red shading indicates the bootstrapped (B = 10^4) Poisson sampling confidence interval of the input 16S rRNA gene copies. (b) Bootstrapped Poisson sampling relationship between %CV and percentage abundance as a function of read depth.



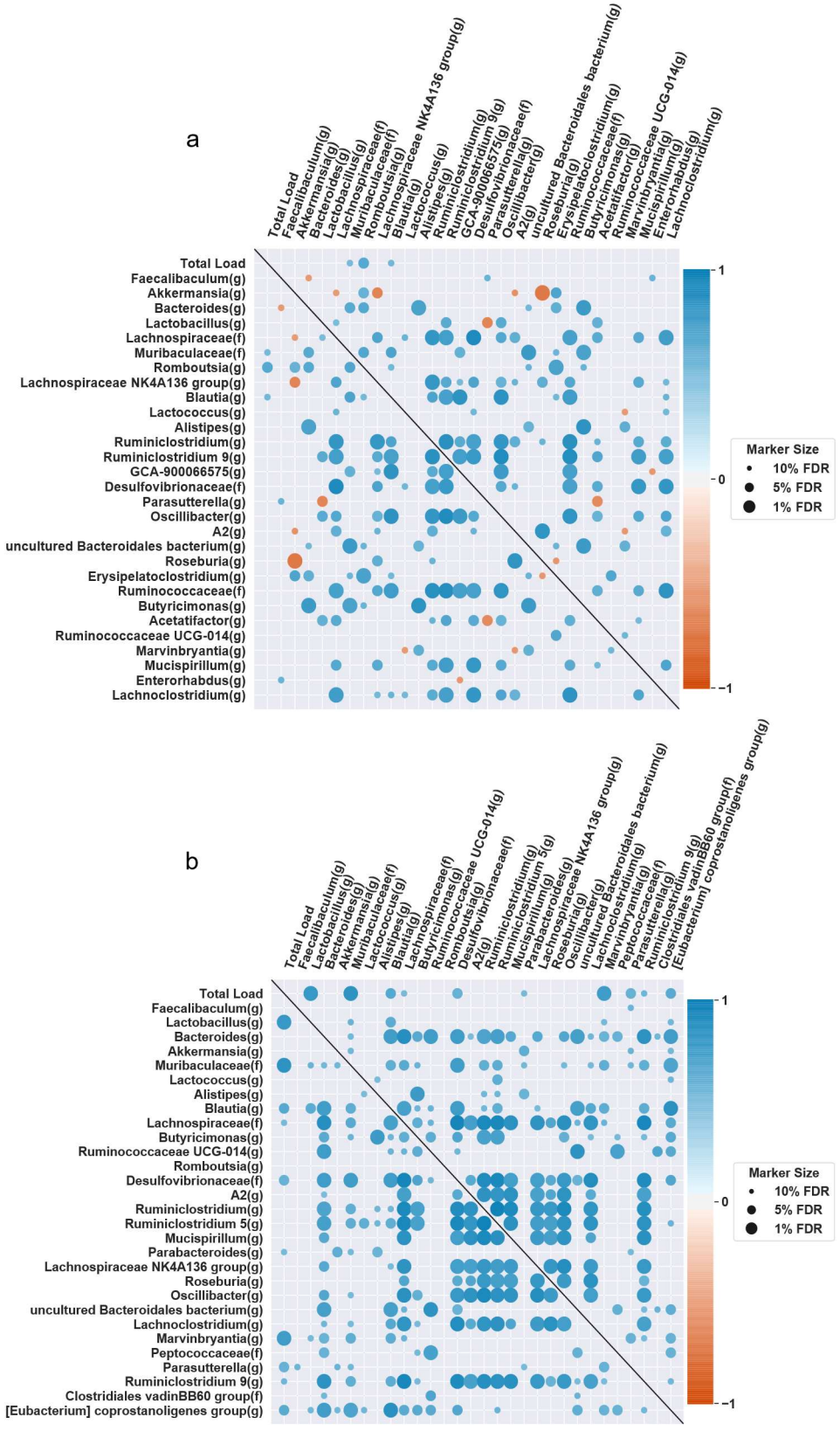
Supplementary Figure 2.5: Optimization of group-specific primers to eliminate amplification of host DNA. Relative abundance of non-specific product amplified from 20 ng/µL small-intestine mucosa sample from a germ-free mouse measured by qPCR. Lower Cq values indicate more amplification. Each color represents a different annealing temperature used during the cycling process. Samples were run in singlet at each temperature.



Supplementary Figure 2.6: Impact of ordination method on data visualization. (a) Principal coordinates analysis (PCoA) plot using Bray–Curtis dissimilarity metric of all samples collected 10 days after the diet switch. (b) Principal component analysis (PCA) plot using log-transform of absolute abundance data after adding a pseudocount of 1 read to all taxa.

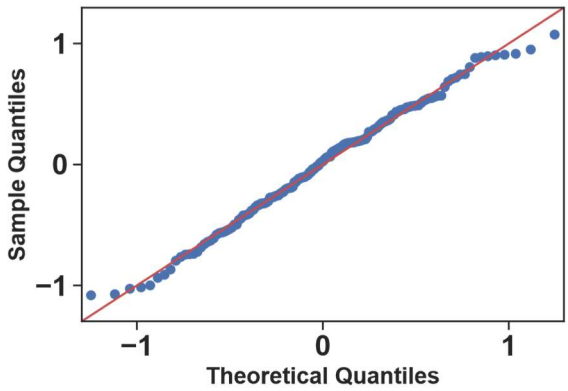


Supplementary Figure 2.7: Comparison of relative and absolute abundance quantification of *Akkermansia(g)* between mice on ketogenic and control diet. Average *Akkermansia(g)* load from stool of N = 6 mice on control diet (red) and N = 6 mice on ketogenic diet (orange). Grey points and bars indicate loads prior to the diet switch when all mice were on the chow diet. Data points from mice without *Akkermansia(g)* are not shown. Bar plots show mean plus or minus the standard deviation. Individual data points are overlaid on the bar plots.



Supplementary Figure 2.8: Absolute-abundance measurements enable unbiased determination of correlation structure in microbiome datasets. Correlation matrices, using Spearman's rank, for the total

microbial load and the top 30 most abundant taxa in stool samples from mice on either a ketogenic diet (a) or control diet (b). The color of each marker is based on the correlation coefficient (orange indicates negative correlations, blue indicates positive correlations) and the size is determined by the q-value of the correlation after Benjamini–Hochberg multiple testing correction. False-discovery rates (FDR) indicate the q-value at which the correlation was deemed significant: 1%, 5%, 10%. Abbreviations: (f), family; (g), genus; (o), order.



Supplementary Figure 2.9: The uncertainty in taxon absolute-abundance measures approximately follows a normal distribution. The quantile-quantile (Q–Q) plot of the mean-centered $\log_2$ relative error of absolute taxon abundances. The relative error is calculated as the ratio of the absolute taxon loads measured by our method of quantitative sequencing with dPCR anchoring over the absolute loads measured by taxon-specific primers in dPCR (data are from Fig. 3b). The x-axis represents the theoretical quantiles from a normal distribution while the y-axis is the actual quantiles of the mean-centered $\log_2$ relative errors.

Supplementary Table 2.1: Contaminant taxa with greater than 1% abundance in negative-control extraction.

Contaminant Taxa	Percentage Abundance
Acinetobacter(g)	31.38
Pseudomonas(g)	24.12
Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium(g)	9.77
Brevundimonas(g)	5.86
Massilia(g)	2.84
Delftia(g)	2.52
Dietzia(g)	2.33
Corynebacterium 1(g)	2.08
Xanthomononadaceae(f)	2.06
Anaerococcus(g)	1.95
Nubsella(g)	1.94
Lysobacter(g)	1.91
Comamonas(g)	1.82
Janthinobacterium(g)	1.30
Shinella(g)	1.29
Novosphingobium(g)	1.23
Sphingobium(g)	1.15
Taibaiella(g)	1.02

(f), family; (g), genus

Supplementary Table 2.2: Comparison between digital PCR anchoring method for absolute abundance measurements and other published absolute abundance methods[2-5].

Approach	Major Improvement to the Field	Demonstrated Limit of Quantification	Demonstrated Limit of Detection	Demonstrated Precision	Validated Sampling Locations	Validation Against PCR Amplification Bias	Bias-Free Validation with High Host DNA Loads	References
Flow Cytometry	Showed importance of quantifying absolute abundance in clinical samples	Not Discussed	Not Discussed	Not Discussed	Stool	Not Applicable	Not Shown	Vandeputte *et al.* 2017[2]
Sequencing Spike-ins	Generated a variety of spike-in standards that can be used. Provided comprehensive analysis of detection limits and accuracy	Not Discussed	Dependent on spike-in amount (~100 copies/reaction)	1.5-1.7X with mock communities	Sludge, Soil	Show that it may skew total load measurement	Not Shown	Tourlousse *et al.* 2016[3]
qPCR Anchoring	Provided a simple and easy method for absolute quantification	Not Discussed	Not Discussed	High correlation at high DNA input levels	Stool	Not Discussed	Not Shown	Jian *et al.* 2018[4]
Total DNA	Provided a simple method for absolute quantification. Showed dramatic variability in loads across animal kingdom and clinical scenarios	Not Discussed	~100 pg of DNA	Not Discussed	Stool	Not Applicable	Not Applicable for stool	Contijoch *et al.* 2019[5]
Digital PCR Anchoring	Quantitative assessment of accuracy and precision of absolute abundances in complex gut samples and their impact on differential taxon analyses	$4.2x10^5$ 16S copies/g Stool $1.0x10^7$ 16S copies/g Mucosa	$4.2x10^4$ 16S copies/g Stool $1.0x10^6$ 16S copies/g Mucosa	2X across 6 orders of magnitude with low and high host DNA load	Stool, Mucosa, Small Intestine, Cecum, Stomach	Yes	Yes	This paper

Supplementary Table 2.3: Composition of ketogenic and control diets used in this study were based on previously reported diets (Envigo, Indianapolis, IN, USA).[6]

	TD.150300	**TD.07797.PWD**
	Control Diet (g/kg)	**Ketogenic Diet (g/kg)**
Casein	200	121
Crisco	61.25	605
Corn Oil	8.75	86.2
Cellulose	50	112.95
Corn Starch	389	0
Maltodextrin	100	0
Sucrose	150	0
DL-Methionine	3	1.56
Vitamin Mix, Teklad (40060)	10	17.8
Choline Bitartrate	0	2.5
TBHQ, antioxidant	0.07	0.14
Mineral Mix, Ca-P Deficient (79055)	13.37	23.8
Calcium Phosphate, dibasic	7.5	24.3
Calcium Carbonate	6.85	4.4
Magnesium Oxide	0.2	0.35

Supplementary Table 2.4: Absolute abundance, relative abundance, fold change and quantification class for each differentially abundant taxon in the stool 10 days after diet switch.

Taxon	Absolute Abundance Ketogenic Diet (16S copies/g)	Absolute Abundance Control Diet (16S copies/g)	log_2 Fold Change (Keto/Control)	Relative Abundance Ketogenic Diet (%)	Relative Abundance Control Diet (%)	Quantification Class
GCA-900066575(g)	1.94E+09	8.71E+07	4.00	0.799	0.014	Semi-Quant
Ruminococcaceae(f)	2.02E+09	2.43E+08	2.87	0.909	0.033	Semi-Quant
Lachnospiraceae NK4A136 group(g)	5.34E+09	8.57E+08	2.58	2.362	0.135	Quant
Acetatifactor(g)	5.20E+08	6.19E+07	2.46	0.256	0.009	Semi-Quant
Lachnospiraceae(f)	8.29E+09	1.71E+09	2.25	3.708	0.226	Quant
Ruminiclostridium 9(g)	1.91E+09	5.01E+08	1.84	0.863	0.059	Quant
Dorea(g)	7.79E+07	0.00E+00	1.36	0.032	0.000	Presence/Absence
Enterorhabdus(g)	7.55E+08	3.51E+08	0.99	0.349	0.052	Quant
[Eubacterium] xylanophilum group(g)	8.69E+07	1.91E+07	0.87	0.037	0.003	No Quant
Peptococcus(g)	8.83E+07	2.29E+07	0.79	0.040	0.004	Semi-Quant
Candidatus Soleaferrea(g)	5.91E+07	6.21E+06	0.78	0.026	0.002	No Quant
Marvinbryantia(g)	5.67E+08	1.35E+09	-1.26	0.226	0.218	Quant
Bacteroides(g)	8.81E+09	3.37E+10	-1.93	3.990	5.578	Quant
Faecalibaculum(g)	1.01E+11	3.87E+11	-1.93	46.724	54.268	Quant
Prevotellaceae UCG-001(g)	1.50E+07	7.87E+07	-2.12	0.006	0.015	No Quant
Bifidobacterium(g)	3.18E+08	1.42E+09	-2.15	0.153	0.100	Quant
Muribaculaceae(f)	1.25E+08	5.74E+08	-2.16	0.056	0.091	Quant
Ruminiclostridium 5(g)	2.85E+08	1.38E+09	-2.26	0.127	0.202	Quant
Ruminococcaceae UCG-014(g)	4.85E+08	2.45E+09	-2.32	0.209	0.427	Quant
Ruminococcaceae NK4A214 group(g)	8.66E+06	6.64E+07	-2.36	0.003	0.009	No Quant
Lactococcus(g)	3.34E+09	1.74E+10	-2.38	1.528	2.715	Quant
Muribaculaceae(f)	8.04E+09	4.26E+10	-2.40	3.520	6.506	Quant
Anaerotruncus(g)	2.23E+07	1.79E+08	-2.68	0.010	0.031	No Quant
Lactobacillus(g)	1.45E+10	1.35E+11	-3.22	6.632	19.295	Quant
Butyricimonas(g)	4.38E+08	4.39E+09	-3.30	0.203	0.755	Quant
Alistipes(g)	1.11E+09	1.15E+10	-3.36	0.526	2.018	Quant
Mollicutes RF39(o)	3.06E+07	3.99E+08	-3.38	0.014	0.074	Semi-Quant
Christensenellaceae(f)	2.35E+07	3.69E+08	-3.55	0.009	0.057	Semi-Quant
Clostridiales vadinBB60 group(f)	3.31E+07	5.68E+08	-3.77	0.017	0.108	Semi-Quant
ASF356(g)	0.00E+00	2.44E+08	-4.65	0.000	0.029	Presence/Absence
Parabacteroides(g)	9.26E+07	3.30E+09	-5.01	0.040	0.457	Quant
Gram-negative bacterium cTPY-13(g)	0.00E+00	3.71E+08	-5.20	0.000	0.050	Presence/Absence

Supplementary Table 2.5: Absolute abundance, relative abundance, fold change, and quantification class for each differentially abundant taxon in the lower small-intestine mucosa 10 days after diet switch.

Taxon	Absolute Abundance Ketogenic Diet (16S copies/g)	Absolute Abundance Control Diet (16S copies/g)	log_2 Fold Change (Keto/Control)	Relative Abundance Ketogenic Diet (%)	Relative Abundance Control Diet (%)	Quantification Class
Lachnoclostridium(g)	1.64E+07	7.07E+05	3.57	0.293	0.006	Semi-Quant
Lachnospiraceae(f)	9.60E+06	4.05E+05	3.16	0.171	0.006	Semi-Quant
A2(g)	2.29E+07	2.08E+06	3.06	0.441	0.025	Semi-Quant
Akkermansia(g)	2.57E+08	3.77E+07	2.74	5.576	0.419	Quant
Escherichia-Shigella(g)	2.96E+06	0.00E+00	2.21	0.059	0.000	Presence/Absence
Dorea(g)	2.94E+06	0.00E+00	2.20	0.062	0.000	Presence/Absence
Bacteroides(g)	5.51E+06	8.09E+05	1.94	0.112	0.009	Semi-Quant
Desulfovibrionaceae(f)	4.29E+06	5.91E+05	1.82	0.097	0.005	Semi-Quant
uncultured Bacteroidales bacterium(g)	1.55E+07	3.98E+06	1.76	0.365	0.041	Quant
Enterorhabdus(g)	1.12E+07	2.74E+06	1.73	0.245	0.022	Semi-Quant
Lachnospiraceae NK4A136 group(g)	1.38E+06	3.38E+05	0.65	0.031	0.005	No Quant
Ruminococcaceae(f)	8.48E+05	6.76E+04	0.51	0.017	0.001	No Quant
uncultured Lachnospiraceae bacterium(g)	6.46E+05	0.00E+00	0.35	0.013	0.000	Presence/Absence
Marvinbryantia(g)	4.62E+06	3.33E+06	0.27	0.127	0.037	Semi-Quant
Ruminiclostridium(g)	5.41E+05	0.00E+00	0.17	0.011	0.000	Presence/Absence
Muribaculaceae(f)	1.22E+08	3.48E+08	-1.51	2.585	3.041	Quant
Lactococcus(g)	1.20E+08	4.31E+08	-1.84	2.533	4.582	Quant
Lactobacillus(g)	9.05E+08	3.70E+09	-2.03	16.956	35.988	Quant

(f), family; (g), genus; (o), order

Supplementary Table 2.6: Primers used in this study, relevant conditions, and specificity. All primers were tested *in silico* for coverage of their desired taxonomic group and specificity.[7-13]

	Akkermansia muciniphila	**Bacteroidales**	***Lachnospiraceae***	***Lactobacillaceae***	**519F-806R**
Forward Primer	CAGCACGTGAAGGTGGGGAC	GGTGTCGGCTTAAGTGCCAT	CGGTACCTGACTAAGAAGC	GCAGCAGTAGGGAATCTTCCA	CAGCMGCCGCGGTAA
Reverse Primer	CCTTGCGGTTGGCTTCAGAT	CGGAYGTAAGGGCCGTGC	AGTTTYATTCTTGCGAACG	CACCGCTACACATGGAG	GGACTACHVGGGTWTCTAAT
Taxonomy Level	Species	Order	Family	Family	Kingdom
Annealing Temp (°C)	65	65	55	60	52
Concentration (nM)	500	500	500	500	500
Coverage (n=1 mismatch)	100%	75%	86%	91%	94% Bacteria, 95% Archaea
Potential Undetected Taxa	None	Rikenellaceae(f); Alistipes(g)	UCG-010(g)	None	None
Potential non-specific interactions (n=1 mismatch)	None	None	None	Leuconostocaceae(o)	None
Citation	Collado *et al*. (2007)[8]	Rinttilä *et al.* (2004)[9]	Kennedy *et al*. (2014)[10]	Castillo *et al*. (2006)[11]	Bogatyrev & Ismagilov (2020)[14] Bogatyrev *et al*. (2020)[13] Bogatyrev (2020)[12]

(f), family; (g), genus; (o), order

SUPPLEMENTARY REFERENCES

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

2. Vandeputte, D., Kathagen, G., D'hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., Vermeire, S., Falony, G. & Raes, J. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507 (2017).

3. Tourlousse, D.M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N. & Sekiguchi, Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* **45**, e23-e23 (2016).

4. Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiome profiling. *PLoS One*, **15**, e0227285 (2020).

5. Contijoch, E.J., Britton, G.J., Yang, C., Mogno, I., Li, Z., Ng, R., et alw. Gut microbiota density influences host physiology and is shaped by host and microbial factors. *eLife* **8**, e40553 (2019).

6. Olson, C.A., Vuong, H.E., Yano, J.M., Liang, Q.Y., Nusbaum, D.J. & Hsiao, E.Y. The Gut Microbiota Mediates the Anti-Seizure Effects of the Ketogenic Diet. *Cell* **173**, 1728-1741.e1713 (2018).

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

8. Collado, M.C., Derrien, M., Isolauri, E., de Vos, W.M. & Salminen, S. Intestinal integrity and Akkermansia muciniphila, a mucin-degrading member of the intestinal microbiota present in infants, adults, and the elderly. *Appl. Environ. Microbiol.* **73**, 7767-7770 (2007).

9. Rinttilä, T., Kassinen, A., Malinen, E., Krogius, L. & Palva, A. Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. *J. Appl. Microbiol.* **97**, 1166-1177 (2004).

10. Kennedy, N.A., Walker, A.W., Berry, S.H., Duncan, S.H., Farquarson, F.M., Louis, P., Thomson, J.M., Satsangi, J., Flint, H.J., Parkhill, J., Lees, C.W. & Hold, G.L. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9**, e88982 (2014).

11. Castillo, M., Martín-Orúe, S.M., Manzanilla, E.G., Badiola, I., Martín, M. & Gasa, J. Quantification of total bacteria, enterobacteria and lactobacilli populations in pig digesta by real-time PCR. *Vet. Microbiol.* **114**, 165-170 (2006).

12. Bogatyrev, S.R. Development of Analytical Tools and Animal Models for Studies of Small-Intestine Dysbiosis. *Dissertation (Ph.D.), California Institute of Technology*, doi:10.7907/VJDZ-7B52 (2020).

13. Bogatyrev, S.R., Rolando, J.C. & Ismagilov, R.F. Self-reinoculation with fecal flora changes microbiota density and composition leading to an altered bile-acid profile in the mouse small intestine. *Microbiome*, **8**. doi: 10.1186/s40168-020-0785-4 (2020).

14. Bogatyrev, S.R. & Ismagilov, R.F. Quantitative microbiome profiling in lumenal and tissue samples with broad coverage and dynamic range via a single-step 16S rRNA gene DNA copy quantification and amplicon barcoding. *Preprint available at:* https://biorxiv.org/cgi/content/short/2020.01.22.914705v1 (2020).

Chapter III

QUANTITATIVE SEQUENCING CLARIFIES THE ROLE OF DISRUPTOR TAXA, ORAL MICROBIOTA, AND STRICT ANAEROBES IN THE HUMAN SMALL-INTESTINE MICROBIOME

1. This chapter was originally published in: **Barlow J.*,** Leite G.*, Romano A., Sedighi R., Chang C., Celly S., Rezaie A., Mathur R., Pimentel M., and Ismagilov R. "Quantitative sequencing clarifies the role of disruptor taxa, oral microbiota, and strict anaerobes in the human small-intestine microbiome." *Microbiome;* doi:10.1186/s40168-021-01162-2.

Abstract

Background

Upper gastrointestinal (GI) disorders and abdominal pain afflict between 12-30% of the worldwide population and research suggests these conditions are linked to the gut microbiome. Although large-intestine microbiota have been linked to several GI diseases, the microbiota of the human small intestine and its relation to human disease has been understudied. The small intestine is the major site for immune surveillance in the gut, and compared with the large intestine, it has greater than 100 times the surface area and a thinner and more permeable mucus layer.

Results

Using quantitative sequencing, we evaluated total and taxon-specific absolute microbial loads from 250 duodenal-aspirate samples and 21 paired duodenum-saliva samples from participants in the REIMAGINE study. Log-transformed total microbial loads spanned 5 logs and were normally distributed. Paired saliva-duodenum samples suggested potential transmission of oral microbes to the duodenum, including organisms from the HACEK group. Several taxa, including *Klebsiella, Escherichia*, *Enterococcus*, and *Clostridium,* seemed to displace strict anaerobes common in the duodenum, so we refer to these taxa as disruptors. Disruptor taxa were enriched in samples with high total microbial loads and in individuals with small intestinal bacterial overgrowth (SIBO). Absolute loads of disruptors were associated with more severe GI symptoms, highlighting the value of absolute taxon quantification when studying small-intestine health and function.

Conclusion

This study provides the largest dataset of the absolute abundance of microbiota from the human duodenum to date. The results reveal a clear relationship between the oral microbiota and the duodenal microbiota and suggest an association between the

absolute abundance of disruptor taxa, SIBO, and the prevalence of severe GI symptoms.

Background

Hundreds of studies have linked the human microbiome to specific diseases. In metabolic diseases or gastrointestinal (GI) disorders (e.g., irritable bowel syndrome [IBS], Crohn's disease, malabsorption) that can cause GI symptoms, such as pain, bloating, and diarrhea, the small intestine instead of the colon may be the primary site of microbial interactions related to disease. Studies have focused on stool primarily for its ease of access and the fact that it has the highest density of microbes out of any human sample type[1]. The stool microbiome has been shown to be a good proxy for the large-intestine microbiome, but is known to differ substantially from the small-intestine microbiome[2, 3]. Compared with the large intestine, the small intestine has several physiological differences that indicate its potential relevance for microbial interactions. The surface area of the small intestine is greater than 100 times that of the large intestine, underlining its role in nutrient absorption. Additionally, the mucus layer of the small intestine is much thinner and more diffuse[4], potentially allowing closer interactions between microbes and the host. Finally, the small intestine is the main site for intestinal immune surveillance by lamina propria dendritic cells[5] and Peyer's patches[6], contributing to the body's response to both commensal and pathogenic microbes.

Although mouse studies have been an insightful proxy for understanding the large-intestine microbiome of humans, the coprophagic behavior of mice[7] and many other animal models results in a substantially different small-intestine microbiome compared with humans[8]. For example, the total microbial load of the human small intestine is generally thought to be low, around 10^2–10^6 CFU/mL[1], whereas microbial loads in laboratory mice are nearly 10^9 CFU/mL[8, 9]. In humans, culturable levels above 10^3–10^5 CFU/mL from duodenal aspirates are used as the clinical determination of small intestinal bacterial overgrowth (SIBO)[10]. SIBO has been shown to correlate with IBS and GI symptoms such as bloating, constipation, and diarrhea[11, 12]. Physiologically, SIBO has also been linked to slow intestinal transit[13], higher body mass index (BMI)[14], and reduced stomach-acid levels[15]. Standard-of-care treatments for SIBO often include antibiotics and diets designed to reduce the amount of rapidly fermentable products in the small intestine[16]. However, reoccurrence of symptoms after antibiotics is common and adherence to strict diets is often difficult for patients[17]. Only recently has a connection between the relative abundance of specific microbial taxa, generally from the *Enterobacteriaceae* family, and SIBO begun to be uncovered[18].

The difficult nature of sampling most of the gastrointestinal tract has resulted in a limited number of studies analyzing the microbial composition of the human small intestine. Several studies have relied on sampling from ileostomy bags[19, 20], but such sampling will not be fully representative of the small-intestine microbiome[21]. More recent studies sample directly from the intact small intestine through an endoscopic procedure and have begun to unravel unique relationships between small-intestine

microbes and disease[18, 22-25]. An added challenge when quantifying individual microbial taxa from samples of low total microbial biomass is that it can be difficult to distinguish true small-intestine microbes from contamination (e.g. from the oral cavity while sampling or from reagents during sample processing). Additionally, the wide range of total microbial loads in the small intestine across individuals highlights the value of using absolute rather than relative microbial loads when investigating potential associations between small-intestine microbes and physiological factors[9, 26, 27].

In this study, we selected a cohort of 250 individuals from the REIMAGINE study[3] to assess the absolute microbial loads in the human duodenum and their potential relationship with factors related to health and disease. We also surveyed the oral microbiome in a subset of 21 individuals from this cohort to understand the relationship between microbial taxa at these two body sites. We utilized our recently developed digital PCR anchored 16S rRNA gene amplicon sequencing method to provide absolute taxon abundances and filter out contaminants in samples with low microbial abundance[9]. We also used our optimized sample-collection procedure with a custom double lumen sterile closed catheter system and optimized processing steps to minimize oral, gastric and dead microbial contamination[28]. We hypothesized that by capturing the absolute microbial abundances of the human duodenal and oral microbiome we would be able to better understand the makeup of the human duodenal microbiome, improve the understanding of the underlying community structure of SIBO, and determine how microbial load and composition correlate with upper GI symptoms.

Results

We studied the microbiome of the duodenum and its potential relationship with health and disease in a cohort of 250 patients enrolled in the REIMAGINE study at Cedars-Sinai Medical Center. All patients undergoing esophagogastroduodenoscopy (EGD) without colonoscopy preparation as standard of care were eligible to enroll, resulting in patients with a wide range of GI conditions. We grouped the reason for endoscopy into 11 broad categories (Table S3.1). The most common (45% of the patient population) reasons for endoscopy were to rule out cancer/polyps and GERD/dyspepsia workup. No healthy controls are currently approved to be included in the study due to the risks associated with the EGD procedure. Summary statistics for patient demographic data and selected metadata categories from the enrollment questionnaire are included in Table S3.1.

Total microbial load of the duodenum across patients with GI symptoms is log-normally distributed

A digital PCR-based determination of total microbial load[9, 29] from 250 human duodenal aspirates revealed samples that spanned loads from our detection limit of ~$5x10^3$ rRNA gene copies/mL up to nearly 10^9 copies/mL. The overall distribution of total loads was log-normal with mean = 6.13 Log_{10} copies/mL and standard deviation = 1.12 Log_{10} copies/mL (Figure 3.1 A, B). A quantile-quantile (QQ) plot

was constructed to compare the sample distribution to a log-normal distribution (Figure 3.1B). Data from our samples aligning with the y = x line on a QQ-plot indicate a high similarity between the sample distribution and a theoretical log-normal distribution[30]. Neither age nor gender significantly correlated with total microbial load (Figure S3.1). Total microbial load also did not correlate with patient reported intake of probiotics supplements or yogurts, smoking, or usage of proton pump inhibitors (Figure S3.2, Table S3.2). Current antibiotic usage appeared to lower the average total microbial load, but antibiotic usage in the previous 6 months had no impact (Figure S3.2, Table S3.2).

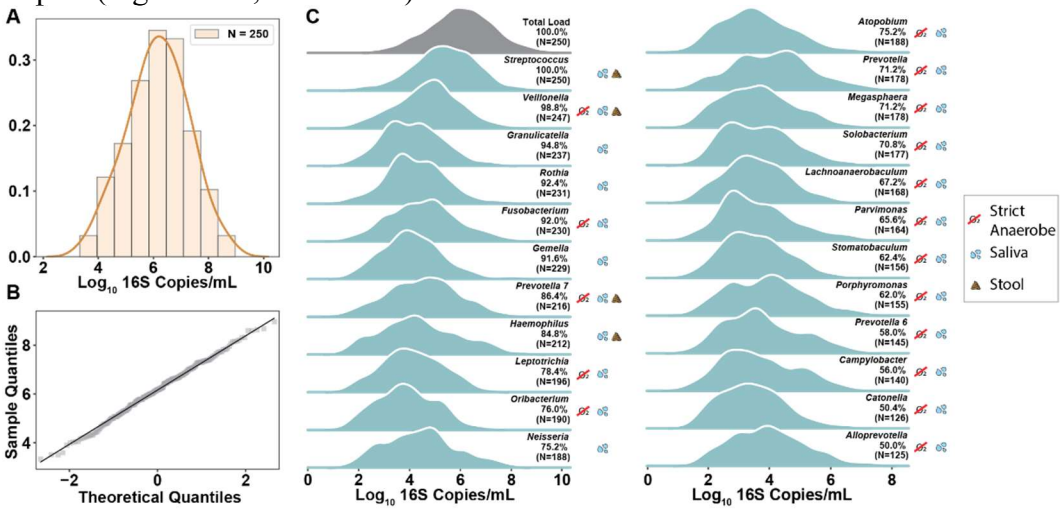


Figure 3.1: Microbial load distribution across 250 human duodenal aspirate samples. (A) Histogram of the total microbial load in 250 duodenal aspirate samples overlaid with a kernel-density estimate. (B) Quantile-quantile plot comparing the sample distribution of the $\log_{10}$-transformed total microbial load in duodenal aspirate samples to a normal distribution. (C) Kernel-density estimate plots showing the absolute abundance distribution for the taxa with greater than 50% prevalence in duodenal aspirates. Prevalence (defined as a taxon's frequency of occurrence in our dataset) and number of samples with each genus are labeled next to the distribution. A legend indicates strict anaerobes (red line through O_2) and the location each genus is commonly found (saliva and/or stool)[31, 32]. Classification of taxa as common in stool or saliva was determined by prevalence of ≥50% (stool data are not included in this study) in the 16 participants for whom we had paired samples.

Digital PCR anchored 16S rRNA gene amplicon sequencing[9] (hereafter quantitative sequencing) provided absolute taxon abundances in each sample and a statistical framework for differentiation between real and contaminant taxa (Methods). We first compared the culture counts from aerobic (MacConkey agar) and anaerobic (blood agar) plates to the total load of microbes expected to grow on these plates (Figure S3.3). For aerobic plating, we observed a bimodal distribution of combined *Escherichia-Shigella*, *Enterobacteriaceae*, *Enterococcus*, and *Aeromonas* bacterial load from quantitative sequencing and culture and a high correlation between the two measurements (Spearman, 0.61, $P < 0.001$, N=244). For anaerobic plating, we observed lower concordance (Spearman, 0.35, $P < 0.001$, N=244) between quantitative sequencing and culture. This discrepancy could reflect the difficulty in

culturing many intestinal microbes[33], especially anaerobes that are initially collected and processed in aerobic environments.

Next, we analyzed the log-transformed absolute-abundance distributions for the most prevalent genera in our dataset (Figure 3.1C). We define prevalence as a taxon's frequency of occurrence in our dataset. *Streptococcus* was present in all 250 samples and followed an approximately log-normal distribution with a mean load that was half an order of magnitude below that of the mean total microbial load and an equal standard deviation. Other genera showed wide-ranging distributions that deviated from normality. For example, *Porphyromonas* appears bimodal with two local maxima whereas *Haemophilus* exhibits a long tail towards higher microbial loads. The 23 most prevalent genera in this study are also commonly found in the oral microbiota[31]. A subset of these genera (*Streptococcus*, *Veillonella*, *Prevotella 7*, *Haemophilus*) are also commonly found in stool samples, indicating possible survival of these genera throughout the entire GI tract[32]. The majority of prevalent genera are either strict or facultative anaerobes, indicating that parts of the duodenal environment are likely anoxic in this patient population.

Direct transmission of microbes from saliva to duodenum

To investigate whether many of the taxa found in the duodenum originated from the oral cavity we analyzed a subset of 21 patients for whom we had paired saliva and duodenum samples that were collected during the same hospital visit. Digital PCR revealed that the total microbial load in saliva was roughly 2.5 orders of magnitude higher than the total load in the duodenum (Kruskal-Wallis, $P < 0.001$).

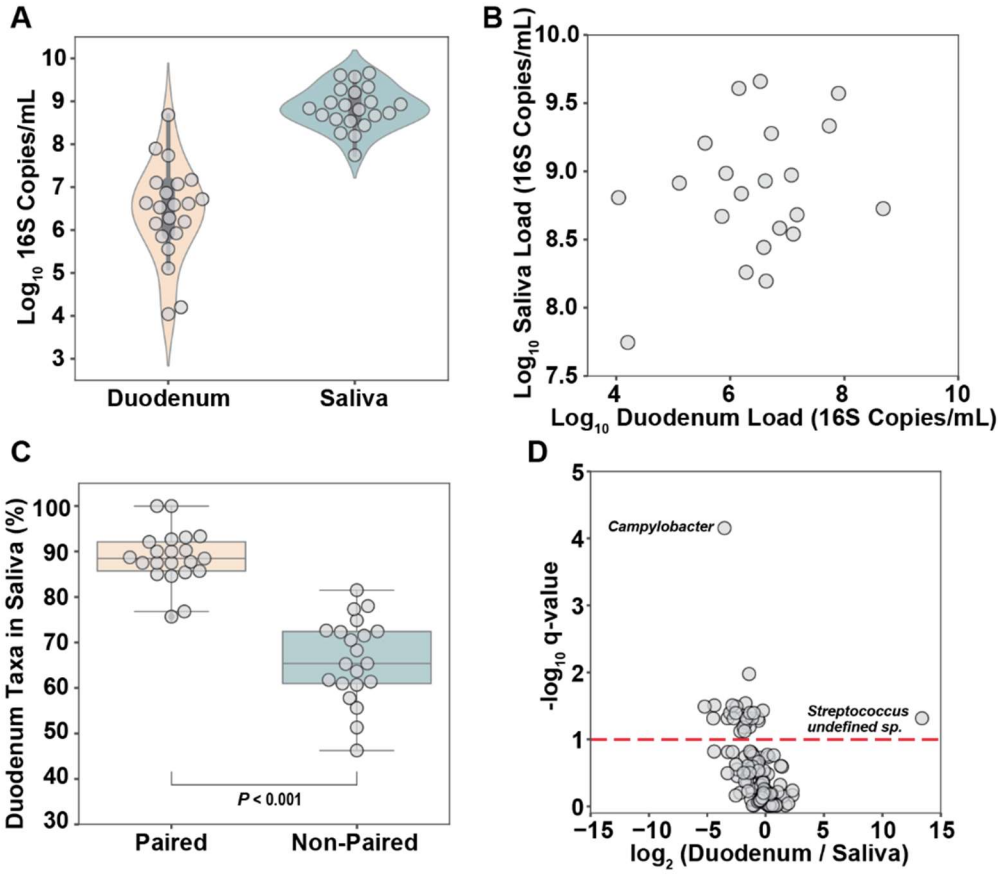


Figure 3.2: Relationship between saliva and duodenal aspirate microbiomes. (A) Total microbial load of 21 paired duodenal aspirate and saliva samples. (B) No significant correlation between the total microbial load of 21 paired duodenal aspirate and saliva samples. (C) Percentage of taxa in duodenal aspirate samples also present in paired (same patient) vs the average of all non-paired saliva samples (Kruskal-Wallis, $P<0.001$). (D) Volcano plot showing the ratio of relative abundances of species in duodenum vs saliva samples. The red dashed line indicates a significance threshold at q=0.1 (Kruskal-Wallis with Benjamini-Hochberg correction). Undefined *Streptococcus sp*. classified as *S. pneumoniae* with 80% confidence and one base pair mismatch to common *Streptococcus* taxon found in all samples.

Further, the range in saliva total loads was 3 orders of magnitude smaller than the range in total loads of the duodenum samples (Figure 3.2A). No significant correlation was observed between the total microbial loads in paired saliva and duodenum samples (Figure 3.2B). In this study, all samples were collected with a custom double-sheathed catheter via endoscope (see Methods) that moves beyond the outer sheath before aspirating duodenal fluid. This custom catheter should limit oral microbiota contamination of the duodenum during the procedure. Additionally, the optimized sample-processing protocol (see Methods) should eliminate extracellular DNA from swallowed dead bacteria.

To evaluate the direct transmission of microbes from saliva to duodenum, we compared the shared taxa between paired (same patient) and randomly paired samples from the same dataset. On average, 89% (±6% S.D.) of the taxa in the

duodenum were also found in the paired saliva sample, whereas only 66% (±9% S.D.) were found in the average of all non-paired comparisons (Figure 3.2C, Kruskal-Wallis, $P < 0.001$), suggesting direct transmission of oral taxa to the duodenum. We then looked for genera that were proportionally enriched in either saliva or duodenum samples. *Campylobacter* was present in 21/21 saliva samples but only 10/21 duodenum samples. The absence of *Campylobacter* in about half of the paired duodenum samples indicates the oral cavity may be the preferred niche of *Campylobacter* or that *Campylobacter* has a high sensitivity to the antibacterial properties of the stomach and small intestine[34] (Figure 3.2D). In contrast, an undefined species of *Streptococcus* was only found in duodenum samples (6/21) (Figure 3.2D). A breakdown of the difference between duodenal and saliva abundance of all taxa is provided in Table S3.3. These differences in the relative abundance of specific taxa of microbes between paired saliva and duodenum samples also provide evidence against oral contamination in the duodenal samples.

Taxa co-correlations reveal disruptor taxa

We assumed that the taxa with the highest absolute abundance would have the highest potential for impacting the host. Thus, we began by analyzing the relationships between the top 20 most abundant genera. A co-correlation heatmap of these taxa revealed several distinct motifs (Figure 3.3A): (1) Taxa whose absolute loads had a high correlation with total load, (2) taxa whose absolute loads had a higher co-correlation with another taxon's absolute load than with total microbial load, (3) taxa with a mutually exclusive relationship with almost all other abundant taxa. Examples of the first motif are in the first column/row of the co-correlation heatmap in Figure 3.3A. Correlation with total load was often an indicator of a prevalent taxon because the variance in total microbial load was larger than the variance in relative abundance. When two taxa have a higher co-correlation with each other than with total load (motif 2) it potentially indicates these taxa share preferred environmental factors or directly cooperate. One group of these co-correlating taxa that included several *Prevotella* species and a species of *Porphyromonas* matches a known shared metabolic niche in the oral cavity[35, 36] (Table S3.4).

Several genera stood out as having no significant correlation with almost all other abundant taxa (motif 3): *Enterobacteriaceae*, *Escherichia-Shigella*, *Clostridium sensu stricto 1*, and *Lactobacillus* (Figure 3.3A). For clarification, throughout the manuscript our references to *Enterobacteriaceae* and *Escherichia-Shigella* refer to unique sequence variants from the *Enterobacteriaceae* family, but only *Escherichia-Shigella* could be classified at the genus level. Based on evidence from a previous study[18] using the REIMAGINE cohort that found *Klebsiella* in several samples, we decided to measure the abundance of *Klebsiella* via qPCR in all samples containing a high abundance (at least 10^5 16S rRNA gene copies/mL) of *Enterobacteriaceae*. We found that the majority (16/22) of the samples with a high abundance of *Enterobacteriaceae* contained *Klebsiella* (Figure S3.4A). Furthermore, in the samples containing *Klebsiella,* there was a high correlation (Pearson, 0.88, $P < 0.001$) between *Klebsiella* load and *Enterobacteriaceae* load (Figure S3.4B). These taxa appeared to disrupt the commonly observed microbial structure (i.e., the prevalent

taxa that generally co-correlate with one another) of the duodenal microbiome. This pattern of mutual exclusivity can be represented algorithmically by sorting all taxa by the difference between their maximum abundance and their mean abundance. Practically, this means that these disruptors are relatively rare (i.e., present in a small fraction of samples), but when they are present they usually dominate, excluding other common taxa. A clustered heatmap of the top 16 taxa as ranked by the difference in their maximum and mean abundances reveals two taxonomic signatures (Figure 3.3B). The first signature in the top left of the heatmap contained the mutually-exclusive taxa from the co-correlation heatmap, along with *Enterococcus*, *Romboutsia*, *Aeromonas*, and *Bacteroides*. The second signature contained taxa that were generally found in lower abundance, many of which are from the HACEK (*Haemophilus*, *Aggregatibacter*, *Cardiobacterium*, *Eikenella*, *Kingella*) group of organisms associated with infective endocarditis[34]. However, the second group also clustered with more common taxa in this dataset, such as *Prevotella* and *Fusobacterium*. Thus, we initially labelled all eight of the taxa in the first taxonomic signature as “disruptors” (Figure 3.3B, bolded taxa) because their presence appeared to be mutually exclusive with many other common taxa.

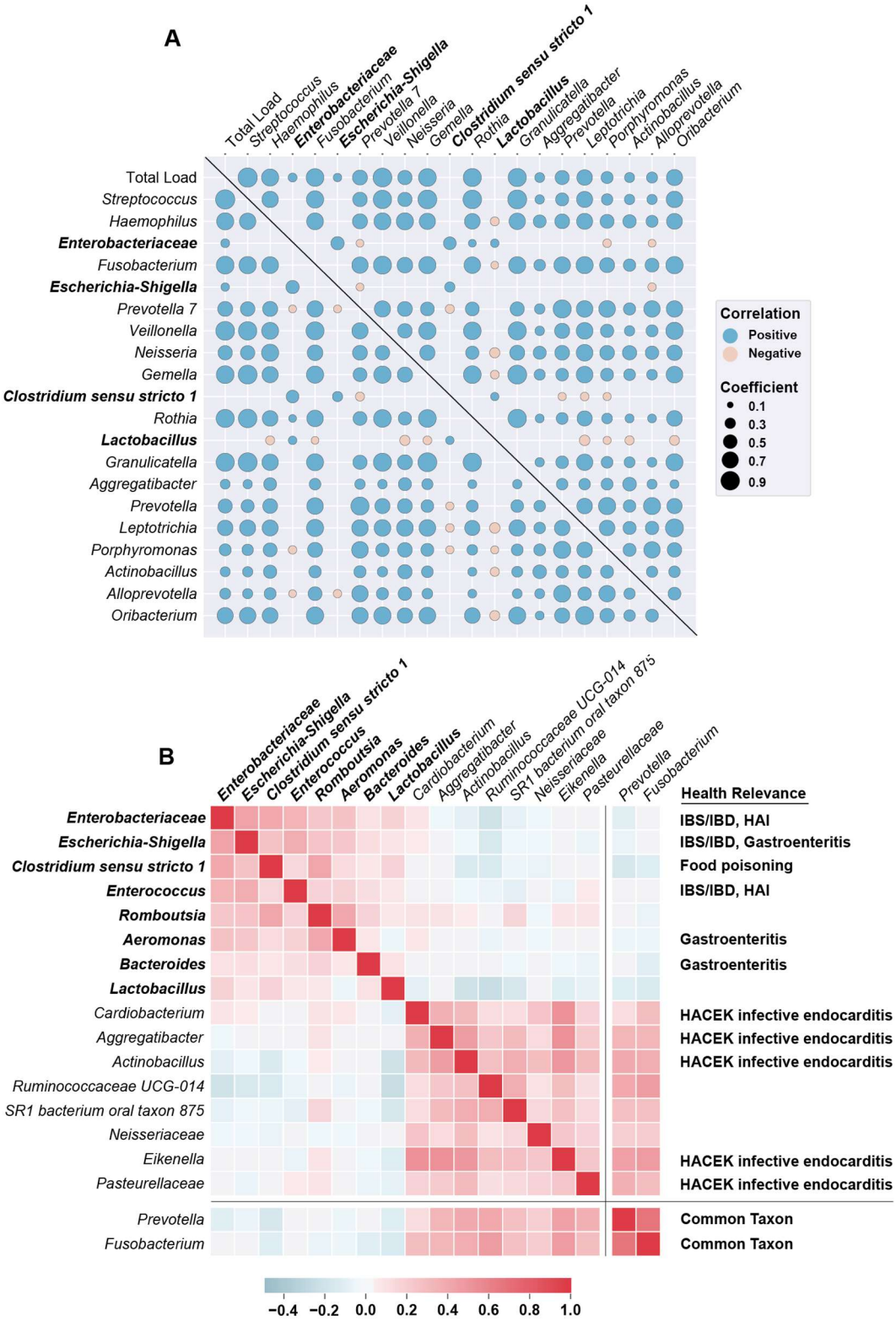


Figure 3.3: Co-correlations reveal which taxa co-occur in high abundance and which can be considered disruptor taxa. (A) Co-correlation matrix of the top 20 most abundant genera and total microbial load. Only significant correlations (q<0.1, Benjamini–Hochberg correction) are shown.

Color of each marker is determined by the sign of the Spearman's correlation coefficient and size of each marker is determined by the magnitude of the coefficient. Disruptor taxa labels are bolded. (B) Clustered co-correlation matrix of the top 16 genera ranked by the difference between their maximum abundance and mean abundance. Two common genera in the dataset are shown at the bottom for reference. The color of each square indicates the Spearman correlation coefficient from negative (blue) to positive (red). Disruptor taxa labels are bolded. Taxa with known relevance to human health are indicated. *Enterobacteriaceae* and *Escherichia-Shigella* are unique sequence variants from the *Enterobacteriaceae* family but only *Escherichia-Shigella* could be classified at the genus level. HAI=hospital acquired infection; IBS, irritable bowel syndrome; IBD, inflammatory bowel disease; HACEK, *Haemophilus*, *Aggregatibacter*, *Cardiobacterium*, *Eikenella*, *Kingella*.

Aerobic disruptor taxa displace strict anaerobes and decrease diversity

After performing the co-correlation analysis, we ran a principal component analysis (PCA) on the absolute taxon abundances to investigate the drivers of variance in the dataset (Figure 3.4A). Total loads spanned 5 orders of magnitude, accounting for most of the variance. Total load cleanly separated samples along the PC1 axis. The second most explanatory axis, PC2, strongly correlated with the Shannon diversity index of samples (Spearman, 0.74, $P < 0.001$, N=250). Ranked feature loadings for PC2 (Figure 3.4B) indicated that many of the disruptor taxa (dark blue) are the main drivers of separation in the positive direction of PC2 whereas the five taxa driving most of the separation in the negative direction (light blue) of PC2 consisted of four strict anaerobes (*Porphyromonas*, *Leptotrichia*, *Prevotella*, *Prevotella 7*) and one obligate aerobe (*Neisseria*). It should be noted that many more taxa were strongly associated with the negative direction of PC2 than the positive direction. This separation matches well with the mutual exclusivity seen between the disruptor taxa and other organisms in the co-correlation analysis. The two disruptor taxa with the highest loads are aerobic pathogens from the *Enterobacteriaceae* family and the taxa most associated with the negative direction of PC2 were strict anaerobes, so we next took a closer look at the composition of strict vs facultative anaerobes in each sample. We found a nearly 1:1 correlation between the strict and facultative anaerobe loads across all samples (Figure 3.4C). Additionally, the fraction of strict anaerobes in a sample was strongly correlated (Pearson, 0.71, $P < 0.001$, N=250) with Shannon diversity (Figure 3.4D), indicating that the disruptor taxa appear to be mutually exclusive with strict anaerobes and the "bloom" of absolute abundance of disruptors decreases Shannon diversity. Furthermore, in half of the samples containing the two most common disruptor taxa (*Enterobacteriaceae* and *Escherichia-Shigella*), the total microbial loads were greater than 10^7 16S rRNA gene copies/mL, indicating a clear enrichment of disruptor taxa in samples with higher than average total microbial loads (Figure 3.4E). This signature of higher than average total microbial loads and mutual exclusivity with other microbes has been observed in some pathogenic microbial species[37, 38].

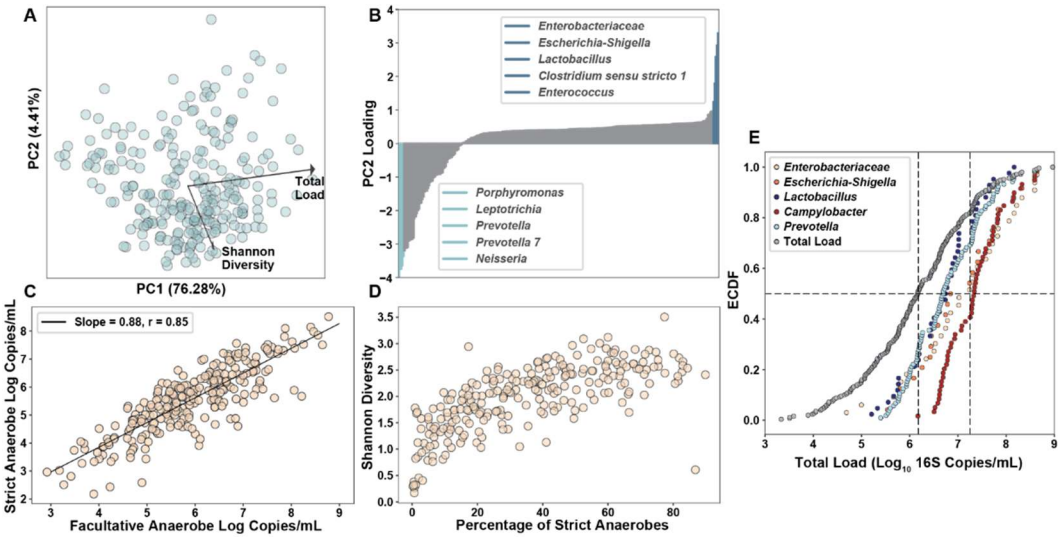


Figure 3.4: Strict anaerobes and disruptor taxa control diversity. (A) PCA plot of absolute microbial abundances at the genus level with the top two correlated metadata variables overlaid. (B) Feature loadings for principal component 2. Top five value-ranked genera in each direction (positive and negative) are highlighted and labeled. (C) Correlation between the strict anaerobic microbial load and facultative anaerobic microbial load. (D) Relationship between the percentage abundance of strict anaerobes and Shannon diversity index. (E) Empiric cumulative distribution function (ECDF) plot for *Enterobacteriaceae* (N=33), *Escherichia-Shigella* (N=24), *Campylobacter* (N=59), *Lactobacillus* (N=42) and the common taxa *Prevotella* (N=104).

Absolute load of disruptor taxa correlates with SIBO and GI symptoms

To determine whether disruptor taxa are associated with disease or GI symptoms we began by looking at patients with and without SIBO (SIBO classification was made based on aerobic culture results, $\geq 10^3$ CFU/mL of duodenal aspirate[10]). Coloring the PCA plot by SIBO classification indicates a clear enrichment of patients with SIBO in the positive direction of the disruptor taxa axis (Figure 3.5A). We observed slightly but not significantly higher total microbial loads in samples from patients with SIBO vs without SIBO (Figure 3.5B). However, comparing the absolute abundance of specific taxa between the SIBO and non-SIBO samples by Kruskal-Wallis showed that the three taxa whose abundances differed the most between SIBO and non-SIBO (*Enterobacteriaceae*, *Escherichia-Shigella*, and a *Clostridium* which, based on the V4 region of the 16S rRNA gene, was classified as *Clostridium perfringens*) were also the three most common disruptor taxa in all samples (Figure 3.5C). This enrichment of disruptor taxa, but not total microbial load, in SIBO samples indicates that overgrowth of specific taxa drives the current clinical classification of SIBO. Additionally, using disruptor taxa load as the criterion for SIBO classification agreed well (80%) with the classification by the gold-standard method, aerobic aspirate culture (Figure S3.5). *Lactobacillus* abundance was similar in SIBO and non-SIBO samples (Figure 3.5C) even though it co-correlated with many of the disruptor taxa (Figure 3.3B). Most of the non-SIBO samples that clustered with SIBO samples on the upper part of the PC plot contained *Lactobacillus* (Figure 3.5A). *Lactobacillus* does not grow on the aerobic (MacConkey agar) plates used for SIBO classification,

which could explain why these samples cluster together by sequencing but are not classified as SIBO by culture.

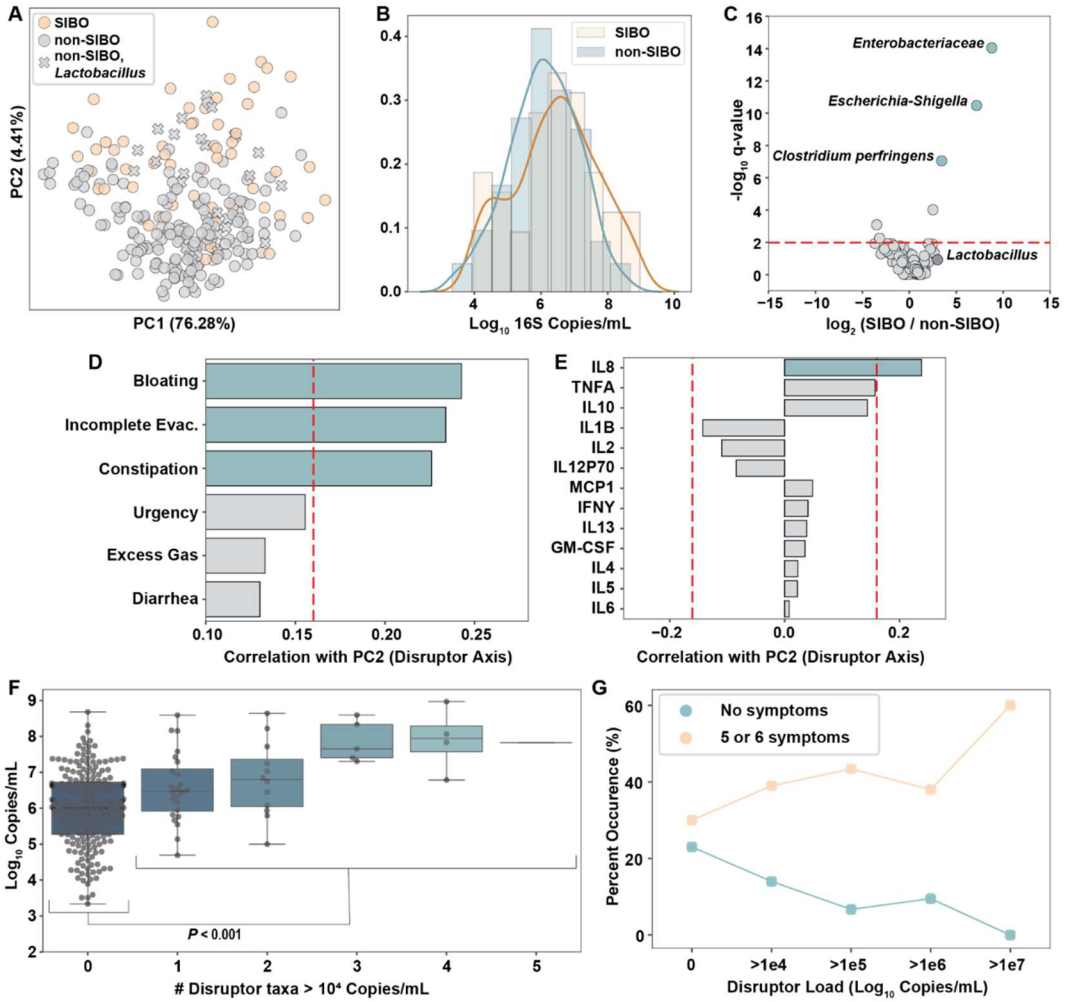


Figure 3.5: Disruptor species are dominant in SIBO samples and correlate with GI symptoms and the inflammatory cytokine IL8. (A) Principal component analysis (PCA) of absolute microbial abundances at the genus level. Colors indicate non-SIBO (grey) or SIBO (orange) participants as determined by culture. "X" markers indicate samples from non-SIBO participants that contained *Lactobacillus*. The PC1 axis correlates with total load and the PC2 axis correlates with the abundance of disruptor taxa. (B) Histogram with overlaid kernel-density estimate of the total microbial loads in samples from SIBO and non-SIBO participants. (C) Volcano plot indicating the taxa that differed between SIBO and non-SIBO samples. The red dashed line indicates the significance threshold at q=0.01. (D) Correlation between PC2 (disruptor axis) and patient-reported symptom scores (on a 0-100 scale). The red dashed line represents significance threshold at q=0.05. (E) Correlation between PC2 and patient serum cytokine levels. The red dashed lines represent the significance thresholds at q=0.05. (F) Boxplot indicating increasing average total microbial load with increasing number of disruptor taxa with loads greater than 10^4 rRNA gene copies/mL (not including *Lactobacillus*). A significant difference between total load in samples with zero disruptor taxa and total load in samples with at least 1 disruptor taxa was observed (P<0.001). (G) Percentage of samples from patients with either 0 symptoms or 5-6 symptoms (out of 6 categories) for individuals with varying loads of disruptor taxa (not including *Lactobacillus*).

Patient-reported GI symptom scores (on a 0-100 scale) were correlated with the disruptor taxa axis (PC2). Bloating, incomplete evacuation, and constipation had the highest correlation with the disruptor taxa axis, whereas correlations between urgency, excess gas, or diarrhea and the disruptor taxa axis were much weaker (Figure 3.5D). There was a weak positive correlation between the disruptor taxa axis and serum interleukin 8 (IL8) levels (Spearman, 0.24, $P < 0.001$, N=232), indicating a potential neutrophil-related response (Figure 3.5E). However, none of the symptoms or cytokines had a significant correlation with the total load axis (PC1). One taxon, which based on the V4 region of the 16S rRNA gene was classified as *C. perfringens,* was the only one that, when present in patients, coincided with a significant increase (Kruskal-Wallis, *P=0.039*) in serum IL8 levels (Figure S3.6). However, there were only 9/250 samples with *C. perfringens,* limiting our ability to draw conclusions about this relationship. Although the two disruptor taxa with the highest absolute abundance (*Enterobacteriaceae* and *Escherichia-Shigella*) were enriched in high total microbial load samples, *Lactobacillus* did not follow this trend. *Lactobacillus* was found in samples with total microbial loads that were similar to the total loads of samples containing common taxa like *Prevotella* (Figure 3.4E). Additionally, in patients with high disruptor taxa loads (after excluding *Lactobacillus* load) the presence of *Lactobacillus* at greater than $5x10^4$ copies/mL negatively correlated with bloating symptoms (Figure S3.7). These two facts led us to believe *Lactobacillus* likely has a more nuanced relationship with the host than the other taxa we classified as disruptors. Thus, we removed *Lactobacillus* from our list of disruptor taxa in our analyses of the association of disruptors with total load (Figure 3.5F) and GI symptoms (Figure 3.5G). When multiple disruptor taxa were present, there was a significant increase in total microbial load (Kruskal-Wallis, $P < 0.001$; Figure 3.5F). Patient-reported symptom scores are inherently qualitative, so to test whether disruptor taxa loads were correlated with more severe GI symptoms, we turned the 0-100 scores into a binary yes/no variable, representing a severe symptom, by drawing a threshold at the median score reported for each symptom (Figure S3.8). We then calculated the percentage of patients with zero severe symptoms and the percentage of patients with many severe symptoms (people reporting severe symptoms in 5-6 of the 6 symptom categories) as a function of disruptor taxa loads (Figure 3.5G). We made three observations. First, at higher disruptor loads, patients were more likely to have more severe GI symptoms. Second, none of the patients with disruptor loads greater than 10^7 copies/mL (N=10) had zero symptoms whereas 60% of them had 5 or 6 symptoms. Of the patients without disruptor taxa (N=153), 23% had zero symptoms and 30% had 5 or 6 symptoms. Disruptor loads may also be higher as a function of age, all but one of the individuals with disruptor loads greater than 10^6 copies/mL (N=23) were older than 50 (Figure S3.9). The absolute and relative abundances of disruptor taxa did not correlate (Figure S3.10), preventing the clear connection between abundant symptoms and high absolute loads of disruptor taxa from being observed when analyzing only relative abundances.

Discussion

In this study, we utilized quantitative sequencing to determine the total and taxon-specific loads from the duodenum of 250 patients undergoing EGD as standard of care. We showed that the total microbial load in the duodenum of these patients spans 5 orders of magnitude and follows a log-normal distribution. Paired saliva-duodenum samples revealed that on average 89% of the taxa in the duodenum were also present in paired saliva samples, suggesting potential transmission of taxa from the oral cavity. Co-correlation analysis of the most abundant taxa revealed a distinct taxonomic motif of "disruptor" taxa that, when present, dominate over other taxa. The most common of these disruptor taxa were aerobic pathogens from the *Enterobacteriaceae* family and were negatively correlated with the presence of strict anaerobes and diversity. In addition to the apparent community disruption, disruptor taxa were enriched in many patients classified as having SIBO and high loads of disruptors correlated with a high prevalence of severe GI symptoms.

Human vs mouse small-intestine microbiome

Several findings from this study emphasize how different the small-intestine microbiome is between mice and humans. Our previous study revealed that the coprophagic nature of mice resulted in total microbial loads spanning approximately one order of magnitude from $5x10^8$–$5x10^9$ 16S rRNA gene copies/mL[8] in the small intestine while our human cohort spanned 5 orders of magnitude with a median of 10^6 copies/mL (Figure S3.11). Additionally, neither the most common disruptor family, *Enterobacteriaceae*, nor any of the taxa with at least 50% prevalence in this study were commonly found in our previous study examining microbial loads in the mouse small intestine[8]. Instead, in that study we found that the mouse small intestine was dominated by *Lactobacillus* and, as a result of coprophagy, several stool microbes[8]. The total microbial load of stool is similar between mice and humans[39] and they both share several common taxonomic groups[40]. These differences should be considered when using mice to model human health or disease impacted by the small intestine.

Value of quantitative analysis

The nearly 5 orders of magnitude spread in total microbial loads in the duodenum of these patients revealed the value of utilizing an absolute abundance measurement technique when analyzing microbial communities. Analyzing absolute abundances of individual taxa also let us filter out likely contaminants using Poisson loading statistics, which is critical for samples with low microbial abundance, such as those sometimes found in the human small intestine[41, 42]. The range of total loads in saliva and in stool each appear to be smaller than in the duodenum, closer to two orders of magnitude, which likely relates to differences in residence times, nutrient availability, and host defenses at these two sites compared with the small intestine[39]. Another benefit of using absolute rather than relative abundance measurements is the improved accuracy of correlations between microbes and host phenotype. For example, the 10 patient samples with the greatest disruptor loads had the highest prevalence of severe GI symptoms, but these samples had relative abundances of disruptor taxa that ranged from 8-97%. This wide range of relative abundances made

samples with high disruptor loads indistinguishable from samples with intermediate disruptor loads when analyzing relative abundances.

Microbial connection between oral cavity and small intestine

The majority (89%) of identified microbial taxa in the paired duodenum samples were also present in the paired saliva samples. Our data supports the hypothesis of oral-duodenal transmission of microbes but a larger paired study utilizing shotgun metagenomic sequencing techniques would provide stronger evidence for this claim. Survival of microbes after ingestion is likely dependent on many host factors, including stomach-acid levels, bile secretions, antimicrobial-peptide production, and GI motility. The bimodal taxon abundance distributions (Figure 3.1C) observed for some taxa, including *Prevotella,* may indicate two subsets of patients with distinct stomach and/or duodenal environments that allow for differential abundance of specific taxa. For example, *Campylobacter concisus*, one of the most common oral *Campylobacter* species, is known to be sensitive to both stomach and bile acids[34]. Therefore, one could hypothesize that if a patient had low levels of stomach or bile acids some *C. concisus* may survive ingestion. Low-acid conditions could also allow many other bacteria to survive transit to the duodenum, resulting in higher total microbial loads in the small intestine. We suspect we observed something similar in our samples; the *Campylobacter* genus was only found in samples with greater than average total microbial loads (Figure 3.4E). However, we did not observe a relationship between total microbial loads in the duodenum and the patients' use of proton pump inhibitors (PPI), which are known to reduce acid production. PPI impact on survival of microbes between the oral cavity and duodenum may be dependent on how recently the PPI was taken, however this information was not collected from patients in the REIMAGINE study. A conclusive comparison of the relative importance of various factors affecting bacterial survival in the duodenum would require additional information on small-intestine secretions of bile acids and antimicrobial peptides in these patients.

Several common oral microbes have been implicated in GI diseases when present in stool[31, 43]. A high microbial load in the small intestine could increase the likelihood of these microbes surviving all the way down the GI tract. The shared taxa between the small intestine and oral microbiota in our paired saliva-duodenum samples provides evidence that blooms of opportunistic pathogens in the mouth could also lead to colonization in the SI[31]. In this study, only 1 of the 21 paired duodenum-saliva samples contained disruptor taxa in the duodenum, but these taxa were not present in the corresponding saliva sample. Several *Enterobacteriaceae* species have been identified in oral samples[44] but usually at a low frequency in healthy populations. Many *Enterobacteriaceae* species are introduced into the gut from contaminated food and water sources[45] which would likely result in only transient oral residence. However, persistent oral *Enterobacteriaceae* species have been linked to the use of dentures and the presence of periodontal disease[46]. All the taxa we classified as disruptors in this study are more frequently found in stool than in the small intestine or oral cavity[31, 32]. Further studies should be performed to determine the source of disruptor taxa in the upper GI tract.

A number of taxonomic groups we identified in the duodenum have members known to be opportunistic pathogens. Beyond disruptor taxa, several taxa from the HACEK group of organisms[47] associated with infective endocarditis were found in high abundance in the duodenum. The route that these and other opportunistic pathogens take to reach the blood stream is not clear but our data show that the HACEK organisms are not limited to the oral cavity. The same traits that allow them to colonize the mouth and heart (biofilm production[48], and general resistance to most host secretions) likely contribute to their ability to survive in the small intestine. Additionally, in mouse models, the transmission of opportunistic pathogens, like *Klebsiella*, from the oral cavity to the intestine has been shown to induce inflammation[31]. The oral cavity presents a potential reservoir for a wide range of opportunistic pathogens that have been linked to GI disorders.

Potential relationship between oxygen and disruptor taxa

Several colonic GI disorders are linked to increased oxygen levels in the lumen resulting from decreased epithelial integrity and inflammation[49]. However, the barrier properties of the small intestine, an absorptive organ, are different from those of the colon. To our knowledge, shifts in absolute abundance of microbes capable of aerobic respiration and anaerobes has not been quantitatively studied previously in the human small intestine. The highly correlated abundance of both strict and facultative anaerobes that we observed could be a function of the oxygen gradients in the gut from the epithelial surface to the center of the lumen[50]. In our study, when diversity collapsed and disruptor taxa bloomed, the microbial composition shifted away from strict anaerobes to taxa capable of aerobic respiration. One clear outlier was a *Clostridium* classified as *C .perfringens,* which is a strict anaerobe but was highly correlated with the *Enterobacteriaceae* genera classified as disruptors. Previous mutualistic relationships between aerobic and anaerobic species that could help facilitate colonization have been observed in other studies with *Bacteroides fragilis* and either *Klebsiella pneumonia* or *Escherichia coli*[51, 52]. We have previously hypothesized that the surprising coexistence of aerobe-anaerobe communities can occur in multi-stable systems, and that these communities can persist due to hysteresis[51]. Although multi-stability and hysteresis have not yet been documented in the gut microbiome, this phenomenon could explain the unexpected coexistence and persistence of aerobe-anaerobe communities in the small intestine.

Disruptor taxa predict SIBO classification and likelihood of GI symptoms

Clinically, SIBO is classified by culture of duodenal aspirates on aerobic MacConkey agar or measurement of exhaled hydrogen and methane after intake of a fermentable sugar solution[10, 53]. The main disruptor taxa (*Enterobacteriaceae*) grow well on MacConkey agar plates, which may explain the high correlation between SIBO classification and samples with disruptor taxa. It is commonly hypothesized that overgrowth of these taxa in the small intestine is responsible for the gas production detected during a breath test, and our study further supports this understanding because we found a correlation between bloating symptoms (attributable to gas production) and disruptor taxa. Future studies should determine whether individuals

with and without high loads of disruptor taxa yield positive breath test results. Our findings support a strong relationship between overgrowth of specific disruptor taxa and GI symptoms in subjects with SIBO. High total microbial load alone in the small intestine was not associated with GI symptoms usually observed in subjects with SIBO and other GI conditions and diseases. Microbial culture is never perfect and will not capture all taxa associated with SIBO and GI conditions. However, our data suggest that SIBO diagnosis via microbial culture should focus on quantification of a specific group of disruptor taxa (*Enterobacteriaceae*) rather than total microbial load. Additionally, SIBO diagnosis via quantitative sequencing should focus on the absolute abundance of the seven disruptor taxa identified in this study.

Lactobacillus seemed to be an exception among the disruptor taxa in several ways. It commonly co-occurred with other disruptors; however, it was also present in many "normal" samples at low abundance. Additionally, when present at high total loads in the presence of other disruptor taxa, *Lactobacillus* load had a negative correlation with bloating score. However, *Lactobacillus* also dominated several samples that had no other disruptor taxa but had high symptom scores. It should also be noted that individuals taking probiotics (N=49) did not have increased prevalence or abundance of *Lactobacillus* in the duodenum. Overall, finer taxonomic resolution may be required to decipher the role of different *Lactobacillus* species and strains. Their impact on human health is likely also dependent on the overall microbial community and host environment.

Although most patients in this study have various GI complications that could result in abdominal symptoms independent of a microbial component, patient samples with high loads of disruptor taxa had a substantially higher likelihood of having many severe GI symptoms. However, total microbial load alone did not associate with GI symptoms. Of the 13 cytokines and chemokines measured, only IL8 levels were significantly higher in the serum of patients with disruptor taxa, potentially indicating an associated local inflammatory process. Future studies that analyze biopsy transcriptomes would be needed to determine whether there is an associated host response, such as immune infiltration or epithelial stress responses in regions with disruptor taxa and/or high total microbial loads.

We initiated this study with four expectations, only one of which was supported by our data. Because mice are coprophagic and humans are not, we expected to see a dramatic difference between mouse and human small-intestine microbiomes. We indeed observed large quantitative and qualitative differences between the two. However, we were more surprised and educated by the three expectations that were shown to be incorrect. First, we expected microbial load in the human duodenum to have a bimodal distribution, with low microbial loads for non-SIBO patients and much higher load for SIBO patients, which our findings did not support (Figure 3.5B). Second, because stomach acid and bile acid secretions isolate the duodenum from the upper GI tract and because the unidirectional flow of digesta and the ileocecal valve isolate the small intestine from the colon, we expected to find a unique population of microbes in the duodenum. We were surprised by the extent to which the oral microbiota appeared to influence the small-intestine microbiota (Figure 3.1C, Figure 3.2). Third, we expected to see microbiomes dominated by taxa generally

thought of as commensals like *Lactobacillus* and *Bifidobacterium*. We were surprised by the prevalence and abundance of taxa known to be human pathogens (Figure 3.1C, Figure 3.3B), especially given that the small intestine is an immune-rich, absorptive organ with a loose mucus structure that likely permits substantial exposure to microbial cells and microbial-associated proinflammatory molecules.

Limitations

An acknowledged limitation of the study is that there are no healthy controls. All participants had some GI condition warranting the EGD procedure, which could bias our dataset and mask our ability to perceive relationships between microbial abundances and patient symptoms. New sampling techniques may be required to reduce the procedural risk involved with sampling healthy controls. Additionally, all collected samples in this study were from the lumenal contents of the duodenum. Distal regions of the small intestine may reveal further insights, and mucosal biopsies could be more indicative of mucosa-associated microbes that interact closely with the host. Although short amplicon sequencing allowed for more samples to be included in this study, utilizing shotgun sequencing approaches to reveal species- and strain-level resolution could provide additional insights, especially with regard to disruptor taxa and potential transmission of taxa from saliva to the duodenum. Additionally, DNA-based analyses can only inform which microbes are in a sample, not whether they are actively performing a function. RNA-based analyses, either 16S rRNA or meta-transcriptomics, may shed additional light on which microbes are resident vs transient members of the duodenum and what functions they are performing. Finally, to truly unravel the connection between oral-to-small intestine microbial transmission and small-intestine microbe-host interactions, a more extensive characterization of the host is needed. Specifically, studies are needed to establish how variations in stomach acid levels, bile secretions, and GI motility impact the abundance and composition of small-intestine microbiota and in turn how the abundance and composition of small-intestine microbiota impacts immune and epithelial cell responses.

Conclusions

This study, with its acknowledged limitations, provides the largest dataset of the absolute abundance of microbiota from the human duodenum to date. We show a clear relationship between the human oral microbiota and that of the duodenum. Furthermore, absolute taxon abundances in the duodenum reveal a distinct subset of disruptor taxa, associated with human pathogens, that appear to displace common strict anaerobes. These same disruptor taxa are enriched in some individuals classified with SIBO and the absolute abundance of these disruptor taxa were associated with more severe GI symptoms. Future studies are needed to establish the host factors that control total microbial load in the duodenum, the mechanism of appearance and persistence of disruptor taxa, and how these disruptor taxa interact with the host.

Methods

Study population and design

The REIMAGINE (Revealing the Entire Intestinal Microbiota and its Associations with the Genetic, Immunologic, and Neuroendocrine Ecosystem) study was conceived to explore the relationships between the small-intestine microbial populations and different conditions and diseases[3]. Male and female subjects aged 18-80 years undergoing standard-of-care upper endoscopy (esophagogastroduodenoscopy, EGD) without colon preparation were prospectively recruited. All subjects were required to fast (from both solids and liquids, including water) starting at midnight the night before the procedure. The study protocol was approved by the Institutional Review Board (IRB) at Cedars-Sinai Medical Center, and subjects provided written informed consent prior to participation (IRB Protocol: 00035192). Data presented here represents a retrospective analysis of this prospectively collected information.

Questionnaires

Prior to EGD, all subjects completed a study questionnaire documenting demographic information and family and medical history, including GI disease and bowel symptoms, medication use, use of alcohol and recreational drugs, travel history, and dietary habits and changes. Subjects also reported any known underlying conditions, such as GI diseases and disorders, neurologic disease, hematologic disease, autoimmune disease, kidney disease, heart disease and cancer. All medical information provided by subjects was verified through audits of medical records. All data were de-identified prior to analysis.

Blood collection and analysis

After completing the study questionnaire, fasting blood samples were collected in BD Vacutainer SST tubes (Becton Dickson, Franklin Lakes, NJ, USA). Levels of circulating pro- and anti-inflammatory cytokines and chemokines were analyzed on a Luminex FlexMap 3D (Luminex Corp., Austin, TX, USA) using a bead-based multiplex panel that included: GM-CSF, IFNγ, IL10, IL12P70, IL13, IL1B, IL2, IL4, IL5, IL6, IL8, MCP1 and TNFα (EMD Millipore Corp., Billerica, MA, USA, cat. #HCYTOMAG-60K).

Saliva and small-intestine lumenal sample collection

Prior to EGD procedure, saliva was collected in a sterile 5 mL tube. During the EGD procedure, samples of duodenal lumenal fluid were procured using a custom-designed sterile aspiration double-lumen catheter (Hobbs Medical, Inc.)[28]. Duodenal aspirates (DA) were collected using a custom-designed sterile inner catheter which was pushed through a sterile bone wax cap only after the endoscopist entered the second portion of the duodenum, in order to reduce contamination from the mouth, esophagus, and stomach. After collection, samples were immediately placed on ice and transferred to the laboratory for further analysis.

Aspirate processing and microbial culture

Prior to microbial culture, an equal volume of sterile 6.5 mM dithiothreitol (DTT) prepared with RNase and DNase PCR-grade sterile water was added at a 1:1 ratio to each saliva and duodenal aspirate (~1mL) and the samples were vortexed until fully liquified (~30 sec) as described previously[28]. A 100 µl aliquot of each duodenal sample (DA+DTT) was then serially diluted with 900 µL sterile 1x PBS and plated on MacConkey agar (Becton Dickinson), and on blood agar (Becton Dickinson). Plates were incubated at 37 °C for 16-18 h under aerobic (MacConkey) or anaerobic (blood agar) conditions. Plates without bacterial growth after 18 h were re-incubated for an additional 18 hours. Colony forming units (CFU) were then counted electronically using a Scan 500 (Interscience, Paris, France). Saliva+DTT and the remainder of each DA+DTT were centrifuged at maximum speed (>13000 RPM) for 5 min. The supernatant was removed, and 1 mL of sterile Allprotect reagent (Qiagen, Hilden, Germany) was added to the microbial pellet and then stored at -80°C.

DNA isolation

On the day of the DNA isolation, DA pellets were thawed on ice and processed as described previously[28]. Microbial DNA was isolated using the MagAttract PowerSoil DNA KF Kit (Qiagen) on a KingFisher Duo (Thermo Fisher Scientific, Waltham, MA, USA), and quantified using Qubit dsDNA high sensitivity and Qubit dsDNA BR Assay kits (Invitrogen by Thermo Fisher Scientific) on a Qubit 4 Fluorometer (Invitrogen, Carlsbad, CA, USA).

16S rRNA gene sequencing

Extracted DNA was amplified, barcoded and sequenced as described previously[8, 9, 29]. Briefly, amplification of the variable 4 (V4) region of the 16S rRNA gene was performed in 20 µL duplicate reactions with: 8 µL of 2.5X 5Prime Hotstart Mastermix (VWR, Radnor, PA, USA), 1 µL of 20X Evagreen (VWR), 2 µL each of 5 µM forward and reverse primers (519F, barcoded 806R, IDT, CoralVille, IA, USA), 3.5 µL of water, and 3.5 µL of extracted DNA template. A CFX96 RT-PCR machine (Bio-Rad Laboratories, Hercules, CA, USA) was used to monitor amplification reactions and all samples were removed in late exponential phase (~10,000 FRU) to minimize chimera formation and non-specific amplification[9, 54, 55]. Amplification was performed under the following cycling conditions: 94 °C for 3 min, up to 50 cycles of 94 °C for 45 s, 54 °C for 60 s, and 72 °C for 90 s. Several samples were rerun after diluting the template as they showed non-exponential amplification in the undiluted sample, a sign of PCR inhibition. Amplified duplicates were pooled together and quantified with KAPA library quantification kit (Roche, Basel, Switzerland) and then all samples were pooled at equimolar concentrations with up to 96 samples per library. AMPureXP beads (Beckman Coulter, Brea, CA, USA) were used to clean up and concentrate libraries before final library quantification with a High Sensitivity D1000 Tapestation Chip (Agilent, Santa Clara, CA, USA). Illumina MiSeq sequencing was performed with a 2x300bp reagent kit by Fulgent Genetics (Temple City, CA, USA).

Raw reads were demultiplexed by Fulgent Genetics. Demultiplexed forward and reverse reads were processed with QIIME 2 2020.2[56]. Loading of sequence data was

performed with the demux plugin followed by quality filtering and denoising with the dada2 plugin[57]. Dada2 trimming parameters were set to the base pair where the average quality score dropped below thirty. All samples were rarefied to the lowest read depth present in all samples (45,386 reads) to decrease biases from varying sequencing depth between samples[58]. The q2-feature-classifier[59] was then used to assign taxonomy to amplicon sequence variants (ASV) with the Silva[60] 132 99% OTUs references. Resulting read count tables were used for downstream analyses in IPython notebooks (see Data Availability Section).

***Klebsiella* specific qPCR**

Primers specific for the *Klebsiella gltA* gene[61] (F: 5'-CAGGCCGAATATGACGAATTC-3' , R: 5'-CGGGTGATCTGCTCATGAA-3') were first informatically evaluated for coverage across *Klebsiella pneumoniae*, *Klebsiella oxytoca,* and *Klebsiella aerogenes* via Primer-BLAST[62]. This primer set was found to have a perfect match against strains from all three tested *Klebsiella* species. These primers were also evaluated in the lab for specificity against *Escherichia coli*. No amplification after 40 cycles was observed with a DNA equivalent of ~10^6 *E. coli* cells from the Zymo microbial community DNA standard (Zymo Research, Irvine, CA, USA). *Klebsiella* qPCR was performed in 10 µL reactions with: 5 µL of Ssofast Evagreen Supermix (Bio-Rad Laboraties), 0.5 µL of 10 µM *gltA* primers, and 3.5 µL of water. A CFX96 RT-PCR machine (Bio-Rad Laboratories) was used for amplification with the following cycling conditions: 95 °C for 3 min, 40 cycles of 95 °C for 15 s, 62 °C for 30 s, and 68 °C for 30 s. Estimated conversion of cycle threshold (Cq) to copies/µL was performed where a Cq of 22.4 equals 1000 copies/uL. *Klebsiella* load was then calculated by adjusting for dilutions and normalizing to the collected sample volume.

Absolute abundance

The total microbial load (bacteria and archaea) of each sample and the absolute abundance of each taxon in individual samples was determined as described previously[9, 29]. Briefly, the Bio-Rad QX200 droplet dPCR system (Bio-Rad Laboratories) was utilized to measure the 16S concentration in each sample with the following reaction components: 1X QX200 EvaGreen Supermix (Bio-Rad), 500 nM forward primer, and 500 nM reverse primer (519F, 806R) and thermocycling conditions: 95 °C for 5 min, 40 cycles of 95 °C for 30 s, 52 °C for 30 s, and 68 °C for 30 s, followed by a dye stabilization step of 4 °C for 5 min and 90 °C for 5 min. The final concentration of 16S rRNA gene copies in each sample was corrected for dilutions and normalized to the extracted sample volume.

For each sample, the input-volume-normalized total microbial load from dPCR was multiplied by each amplicon sequence variant's (ASV) relative abundance to determine the absolute abundance of each ASV. No correlation between collected sample volume and measured bacterial load was observed. The average of all sample volumes for a specific sample type was used for a few samples (11 duodenum, 10 saliva) that were missing the starting volume information. A 95% confidence interval of input volumes for duodenum samples ranged from 0.18-1.93 mL indicating that

the estimated input volume measurement would likely be up to 4X off in either direction while the total microbial load ranged 40,000X. Similarly, a 95% confidence interval of input volumes for saliva samples range from 0.36-1.28 mL indicating that the estimated input volume measurement would likely be up to 2X off in either direction while the total microbial load ranged 82X.

Poisson quality filtering

Two separate quality-filtering steps based on Poisson statistics were used to determine the statistical confidence in the measured values. First, a 95% confidence interval was calculated from the repeated measures of water blanks. Samples with a total microbial load below the upper bound of this confidence interval were removed from further analysis.

Second, the limit of detection (LOD) in terms of relative abundance was determined for each sample. Sequencing can be divided into two separate Poisson sampling steps. First, an aliquot of sample is taken from the extracted sample and input into the library amplification reaction. The LOD of the library amplification step was determined by multiplying the total microbial load from dPCR by the input volume into the library amplification reaction and then finding the relative abundance corresponding to an input of three copies. Poisson statistics tells us that the likelihood of sampling one or more copies with an average input of three copies is 95%. The second Poisson sampling step in sequencing arises from the number of reads generated from the amplified library. The accuracy of the second Poisson sampling step was previously shown[9] to follow a negative exponential curve, $LOD = 7.115 * read\ depth^{-0.115}$, between the total read depth and relative abundance at which 95% confidence of detection is observed. The minimum of the two described LODs (first determined per sample by total load, and second by sequencing depth) was then determined for each sample. For each sample, the abundance of any ASV with a relative abundance below the LOD was set to zero. After filtering, data tables for each taxonomic level were generated.

Data transforms and dimensionality reduction

For PCA, all absolute taxon abundances were log-transformed. To handle zeros, a pseudo-count of 0.1 reads was added to all taxon relative abundances before multiplying by each sample's total microbial load as determined by digital PCR. PCA was performed with the *sklearn.decomposition.PCA* function in Python. Ranked feature loadings for each taxon on a given principal component were determined by scaling the corresponding eigenvector by the maximum transformed value for that principal component axis.

Statistical analysis and correlations

Group comparisons (e.g., SIBO vs. no SIBO, saliva vs. duodenum) were analyzed using the non-parametric Kruskal-Wallis rank sums tests with Benjamini–Hochberg multiple hypothesis testing correction using *SciPy.stats Kruskal* function and *statsmodels.stats.multitest multipletests* function with the *fdr_bh* option.

Correlation coefficients were either Spearman or Pearson and corresponding *P*-values for all correlations were determined with *scipy.stats.spearmanr* or *scipy.stats.pearsonr* functions. Multiple hypothesis testing was performed for each group of correlations (e.g. taxa co-correlations, cytokine correlations) separately using the Benjamini–Hochberg procedure.

Ethic Approval and Consent to Participate

The study was reviewed and approved by the Cedars-Sinai Medical Center IRB (Protocol #00035192). All participants provided written informed consent prior to participation.

Consent for Publication

Not applicable.

Availability of Data and Material

Sequencing data generated during this study are available in the National Center for Biotechnology Information Sequence Read Archive repository under study accession number PRJNA674353. Raw data for each figure and IPython notebooks for data processing and figure generation are available through CaltechDATA: https://data.caltech.edu/records/1701.

Acknowledgements

We thank the Caltech Bioinformatics Resource Center for assistance with statistical analyses, Jenny Ji for related analyses and Natasha Shelby for contributions to writing and editing this manuscript. We acknowledge OpenMoji for use of the saliva and stool graphics in Figure 3.1. We thank Stacy Weitsman, Walter Morales and Maria Jesus Villanueva-Milan from MAST for assistance with sample processing and data curation from the REIMAGINE study. We also thank the Gastroenterology team at Cedars-Sinai Medical Center for assistance with patient recruitment and endoscopy procedures.

References

1. Donaldson, G.P., Lee, S.M. & Mazmanian, S.K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20-32 (2016).
2. Yasuda, K. et al. Biogeography of the Intestinal Mucosal and Lumenal Microbiome in the Rhesus Macaque. *Cell Host Microbe* **17**, 385-391 (2015).
3. Leite, G.G.S. et al. Mapping the Segmental Microbiomes in the Human Small Bowel in Comparison with Stool: A REIMAGINE Study. *Dig. Dis. Sci.* **65**, 2595-2604 (2020).
4. Johansson, M.E.V., Sjövall, H. & Hansson, G.C. The gastrointestinal mucus system in health and disease. *Nature reviews. Gastroenterology & hepatology* **10**, 352-361 (2013).
5. Ko, H.-J. & Chang, S.-Y. Regulation of intestinal immune system by dendritic cells. *Immune Netw.* **15**, 1-8 (2015).
6. Rios, D. et al. Antigen sampling by intestinal M cells is the principal pathway initiating mucosal IgA production to commensal enteric bacteria. *Mucosal Immunol.* **9**, 907-916 (2016).
7. Ebino, K.Y. Studies on coprophagy in experimental animals. *Jikken Dobutsu* **42**, 1-9 (1993).
8. Bogatyrev, S.R., Rolando, J.C. & Ismagilov, R.F. Self-reinoculation with fecal flora changes microbiota density and composition leading to an altered bile-acid profile in the mouse small intestine. *Microbiome* **8**, 19 (2020).
9. Barlow, J.T., Bogatyrev, S.R. & Ismagilov, R.F. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. *Nature Communications* **11**, 2590 (2020).
10. Pimentel, M., Saad, R.J., Long, M.D. & Rao, S.S.C. ACG Clinical Guideline: Small Intestinal Bacterial Overgrowth. *Official journal of the American College of Gastroenterology | ACG* **115** (2020).
11. Lupascu, A. et al. Hydrogen glucose breath test to detect small intestinal bacterial overgrowth: a prevalence case–control study in irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **22**, 1157-1160 (2005).
12. Shah, A. et al. Small Intestinal Bacterial Overgrowth in Irritable Bowel Syndrome: A Systematic Review and Meta-Analysis of Case-Control Studies. *Official journal of the American College of Gastroenterology | ACG* **115** (2020).
13. Roland, B.C. et al. Small Intestinal Transit Time Is Delayed in Small Intestinal Bacterial Overgrowth. *J. Clin. Gastroenterol.* **49** (2015).
14. Roland, B.C. et al. Obesity increases the risk of small intestinal bacterial overgrowth (SIBO). *Neurogastroenterol. Motil.* **30**, e13199 (2018).
15. Su, T., Lai, S., Lee, A., He, X. & Chen, S. Meta-analysis: proton pump inhibitors moderately increase the risk of small intestinal bacterial overgrowth. *J. Gastroenterol.* **53**, 27-36 (2018).
16. Quigley, E.M.M., Murray, J.A. & Pimentel, M. AGA Clinical Practice Update on Small Intestinal Bacterial Overgrowth: Expert Review. *Gastroenterology* **159**, 1526-1532 (2020).
17. Pimentel, M. et al. Antibiotic Treatment of Constipation-Predominant Irritable Bowel Syndrome. *Dig. Dis. Sci.* **59**, 1278-1285 (2014).
18. Leite, G. et al. The duodenal microbiome is altered in small intestinal bacterial overgrowth. *PLoS One* **15**, e0234906 (2020).
19. Jonsson, H. Segmented filamentous bacteria in human ileostomy samples after high-fiber intake. *FEMS Microbiol. Lett.* **342**, 24-29 (2013).
20. Zoetendal, E.G. et al. The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *The ISME Journal* **6**, 1415-1426 (2012).
21. Hartman, A.L. et al. Human gut microbiome adopts an alternative state following small bowel transplantation. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17187-17192 (2009).

22. Chen, Y. et al. Dysbiosis of small intestinal microbiota in liver cirrhosis and its association with etiology. *Sci. Rep.* **6**, 34055 (2016).
23. Saffouri, G.B. et al. Small intestinal microbial dysbiosis underlies symptoms associated with functional gastrointestinal disorders. *Nature Communications* **10**, 2012 (2019).
24. Zmora, N. et al. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell* **174**, 1388-1405.e1321 (2018).
25. Chen, R.Y. et al. Duodenal Microbiota in Stunted Undernourished Children with Enteropathy. *New Engl. J. Med.* **383**, 321-333 (2020).
26. Knight, R. et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology* **16**, 410-422 (2018).
27. Morton, J.T. et al. Establishing microbial composition measurement standards with reference frames. *Nature Communications* **10**, 2719 (2019).
28. Leite, G.G.S. et al. Optimizing microbiome sequencing for small intestinal aspirates: validation of novel techniques through the REIMAGINE study. *BMC Microbiol.* **19**, 239 (2019).
29. Bogatyrev, S.R. & Ismagilov, R.F. Quantitative microbiome profiling in lumenal and tissue samples with broad coverage and dynamic range via a single-step 16S rRNA gene DNA copy quantification and amplicon barcoding. *bioRxiv*, 10.1101/2020.1101.1122.914705v914701 (2020).
30. Ghasemi, A. & Zahediasl, S. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab* **10**, 486-489 (2012).
31. Atarashi, K. et al. Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. *Science* **358**, 359 (2017).
32. Schmidt, T.S.B. et al. Extensive transmission of microbes along the gastrointestinal tract. *eLife* **8**, e42693 (2019).
33. Lagkouvardos, I., Overmann, J. & Clavel, T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. *Gut Microbes* **8**, 493-503 (2017).
34. Ma, R. et al. Investigation of the effects of pH and bile on the growth of oral Campylobacter concisus strains isolated from patients with inflammatory bowel disease and controls. *J. Med. Microbiol.* **64**, 438-445 (2015).
35. Marcotte, H. & Lavoie, M.C. Oral microbial ecology and the role of salivary immunoglobulin A. *Microbiology and molecular biology reviews : MMBR* **62**, 71-109 (1998).
36. Hojo, K., Nagaoka, S., Ohshima, T. & Maeda, N. Bacterial Interactions in Dental Biofilm Development. *J. Dent. Res.* **88**, 982-990 (2009).
37. Hopkins, E.G.D., Roumeliotis, T.I., Mullineaux-Sanders, C., Choudhary, J.S. & Frankel, G. Intestinal Epithelial Cells and the Microbiome Undergo Swift Reprogramming at the Inception of Colonic Citrobacter rodentium Infection. *mBio* **10**, e00062-00019 (2019).
38. Argüello, H. et al. Early Salmonella Typhimurium infection in pigs disrupts Microbiome composition and functionality principally at the ileum mucosa. *Sci. Rep.* **8**, 7788 (2018).
39. Contijoch, E.J. et al. Gut microbiota density influences host physiology and is shaped by host and microbial factors. *eLife* **8**, e40553 (2019).
40. Nguyen, T.L.A., Vieira-Silva, S., Liston, A. & Raes, J. How informative is the mouse for human gut microbiota research? *Disease Models and Mechanisms* **8**, 1 (2015).
41. Caruso, V., Song, X., Asquith, M. & Karstens, L. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems* **4**, e00163-00118 (2019).
42. Glassing, A., Dowd, S.E., Galandiuk, S., Davis, B. & Chiodini, R.J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 24 (2016).
43. Gevers, D. et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe.* **15**, 382-392 (2014).

44. Goldberg, S., Cardash, H., Browning, H., Sahly, H. & Rosenberg, M. Isolation of Enterobacteriaceae from the Mouth and Potential Association with Malodor. *J. Dent. Res.* **76**, 1770-1775 (1997).
45. Smith, J.L. & Fratamico, P.M. in Encyclopedia of Food and Health. (eds. B. Caballero, P.M. Finglas & F. Toldrá) 539-544 (Academic Press, Oxford; 2016).
46. Gonçalves, M.O. et al. Periodontal disease as reservoir for multi-resistant and hydrolytic enterobacterial species. *Lett. Appl. Microbiol.* **44**, 488-494 (2007).
47. Sharara, S.L., Tayyar, R., Kanafani, Z.A. & Kanj, S.S. HACEK endocarditis: a review. *Expert Rev. Anti Infect. Ther.* **14**, 539-545 (2016).
48. Karched, M., Bhardwaj, R.G. & Asikainen, S.E. Coaggregation and biofilm growth of Granulicatella spp. with Fusobacterium nucleatum and Aggregatibacter actinomycetemcomitans. *BMC Microbiol.* **15**, 114 (2015).
49. Rigottier-Gois, L. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *The ISME journal* **7**, 1256-1261 (2013).
50. Litvak, Y. et al. Commensal Enterobacteriaceae Protect against Salmonella Colonization through Oxygen Competition. *Cell Host Microbe* **25**, 128-139.e125 (2019).
51. Khazaei, T. et al. Metabolic multi-stability and hysteresis in a model aerobe-anaerobe microbiome community. *Science Advances*, (provisionally accepted with minor revisions) (2020).
52. Dejea, C.M. et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. *Science* **359**, 592 (2018).
53. Rezaie, A. et al. Hydrogen and Methane-Based Breath Testing in Gastrointestinal Disorders: The North American Consensus. *The American journal of gastroenterology* **112**, 775-784 (2017).
54. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. & Polz, M.F. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl. Environ. Microbiol.* **71**, 8966-8969 (2005).
55. Suzuki, M.T. & Giovannoni, S.J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625-630 (1996).
56. Bolyen, E. et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints* **6**, e27295v27292 (2018).
57. Callahan, B.J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581 (2016).
58. Weiss, S. et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
59. Bokulich, N.A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90-90 (2018).
60. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-D596 (2013).

Supplementary Information

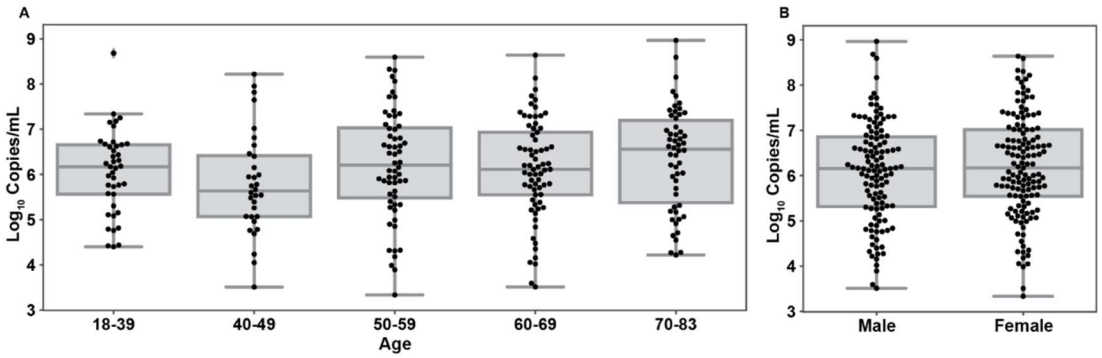


Figure S3.1: Total microbial load breakdown by age (A) and gender (B).

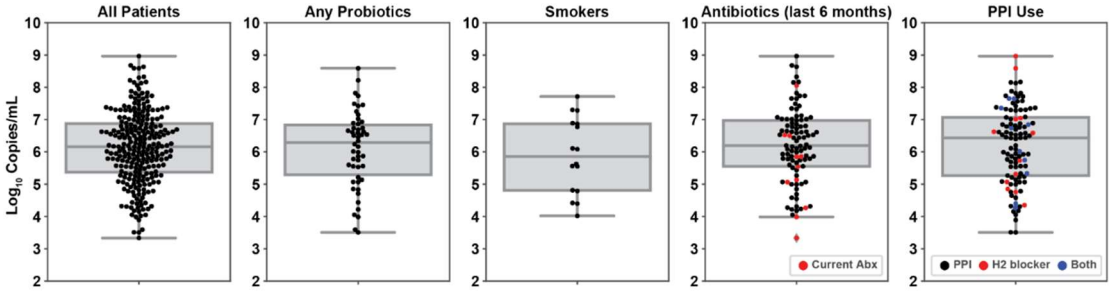


Figure S3.2: Distribution of total microbial load from subpopulations of patients: taking probiotics (N=49), active smokers (N=16), taking antibiotics in the past 6 months (N=100), or taking proton pump inhibitors (PPI, N=106).

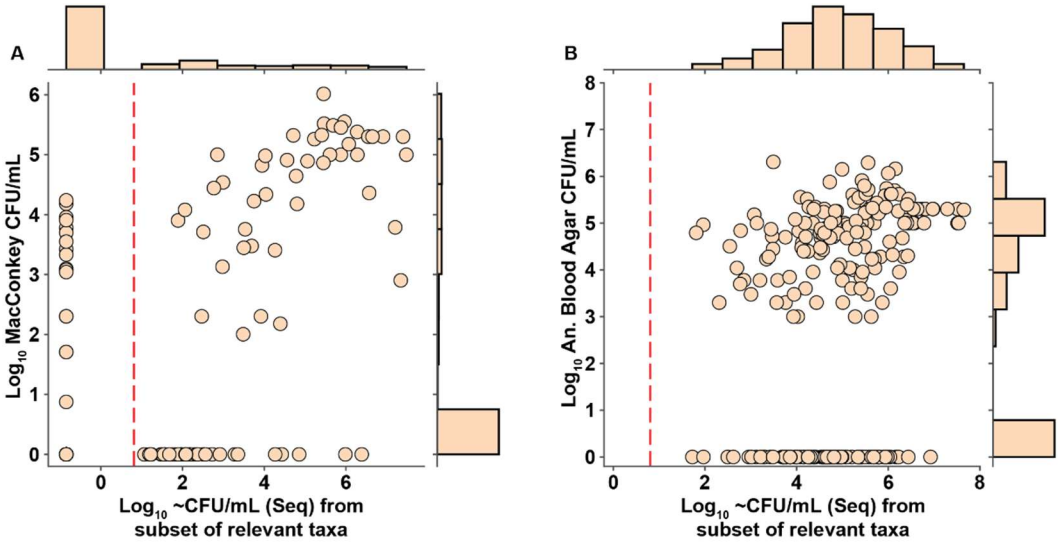


Figure S3.3: (A) Scatterplot comparing aerobic culture load from MacConkey plates to total load from 16S quantitative sequencing of only the subset of bacteria that are known to grow on MacConkey plates (*Escherichia-Shigella*, *Enterobacteriaceae*, *Enterococcus*, and *Aeromonas*)[1]. (B) Scatterplot comparing anaerobic culture load, from blood agar plates, to total load from sequencing of prevalent bacteria that are expected to grow on blood agar plates (*Prevotella*, *Streptococcus*, *Fusobacterium*, *Escherichia-Shigella*)[2]. Red dashed line indicates limit of detection of quantitative sequencing method. N = 244. (Six patients in the study were lacking culture data.)

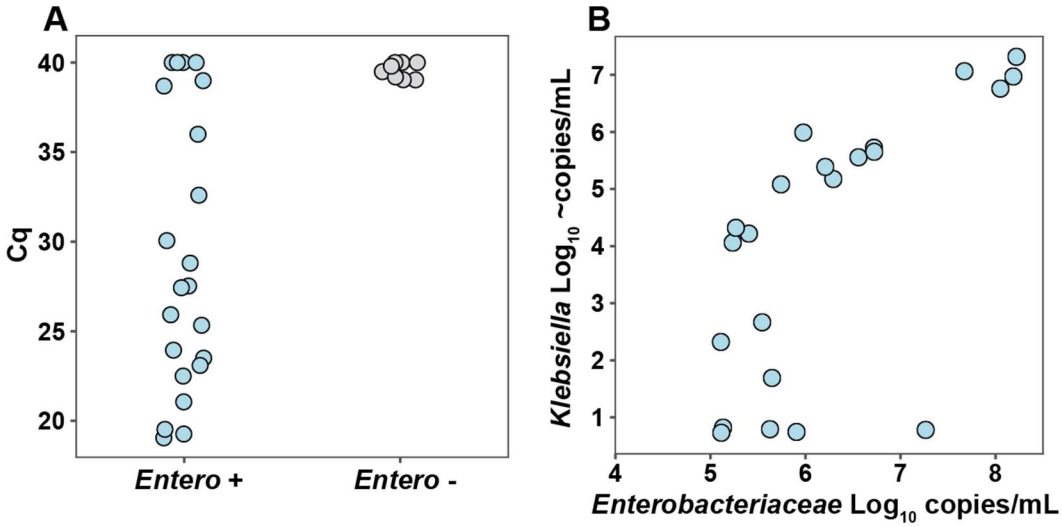


Figure S3.4: (A) Cycle threshold (Cq) values yielded by qPCR with *Klebsiella*-specific primers. Duodenum aspirate samples were classified via quantitative sequencing as containing *Enterobacteriaceae* ("*Entero* +", N=22) or not containing *Enterobacteriaceae* ("*Entero* –", N=8). (B) Total loads of *Enterobacteriaceae* (copies/mL) in duodenum aspirates as a factor of the approximate *Klebsiella load* (copies/mL). *Enterobacteriaceae* measurements are calculated based on 16S rRNA gene copies (8 copies/genome) and *Klebsiella* measurements are calculated based on the citrate synthase gene (gltA, 1 copy/genome).

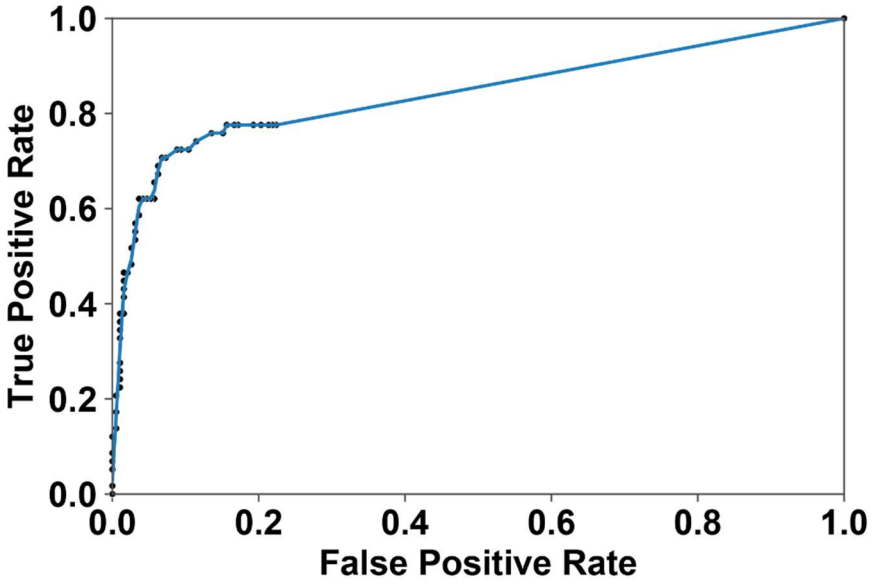


Figure S3.5: Receiver operating characteristic (ROC) curve using absolute loads of seven disruptor taxa (*Enterobacteriaceae*, *Escherichia-Shigella*, *Clostridium sensu stricto 1*, *Enterococcus*, *Romboutsia*, *Aeromonas*, *Bacteroides*) identified in the sequencing data for SIBO classification. SIBO classification was made based on gold-standard aerobic culture results, $\geq 10^3$ CFU/mL of duodenal aspirate. Data points are connected by a line between each consecutive point.

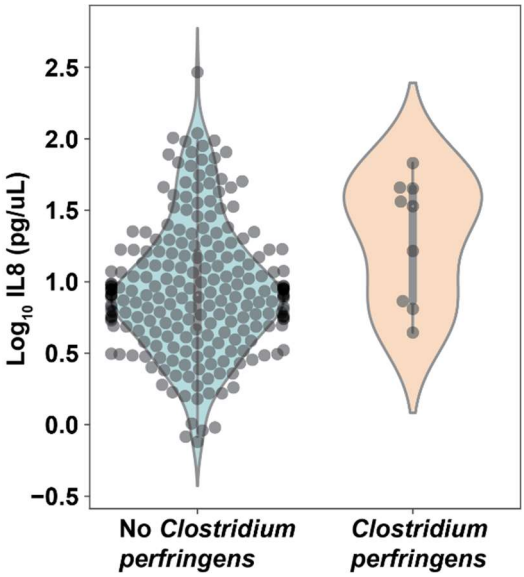


Figure S3.6: IL8 levels in samples with and without *Clostridium perfringens*.

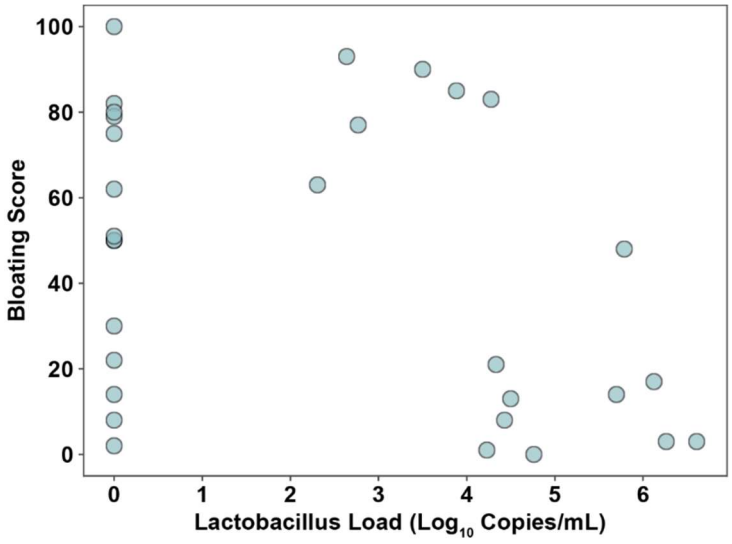


Figure S3.7: Relationship between *Lactobacillus* load and bloating symptoms in samples containing additional (non-*Lactobacillus*) disruptor taxa.

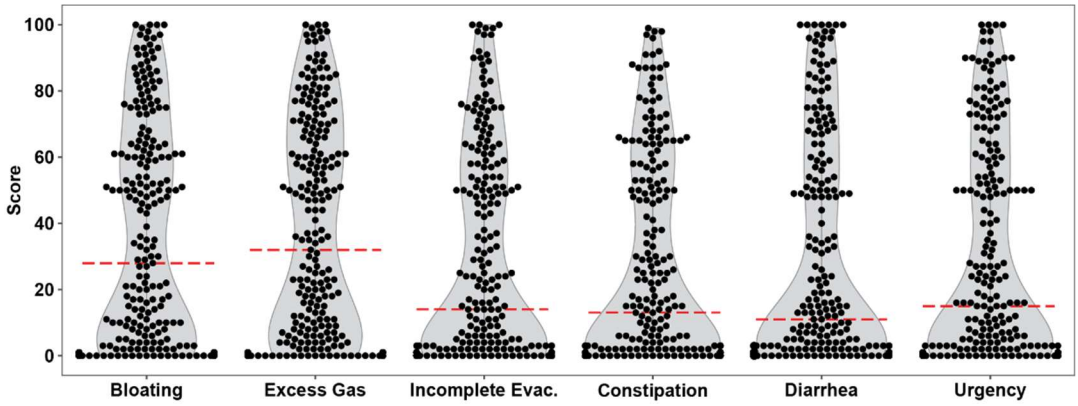


Figure S3.8: Violin plots with data points overlaid for patient-reported symptom scores. Binary threshold for determining whether severe symptoms exist was set at the median score reported of each symptom, shown by the red-dashed lines.

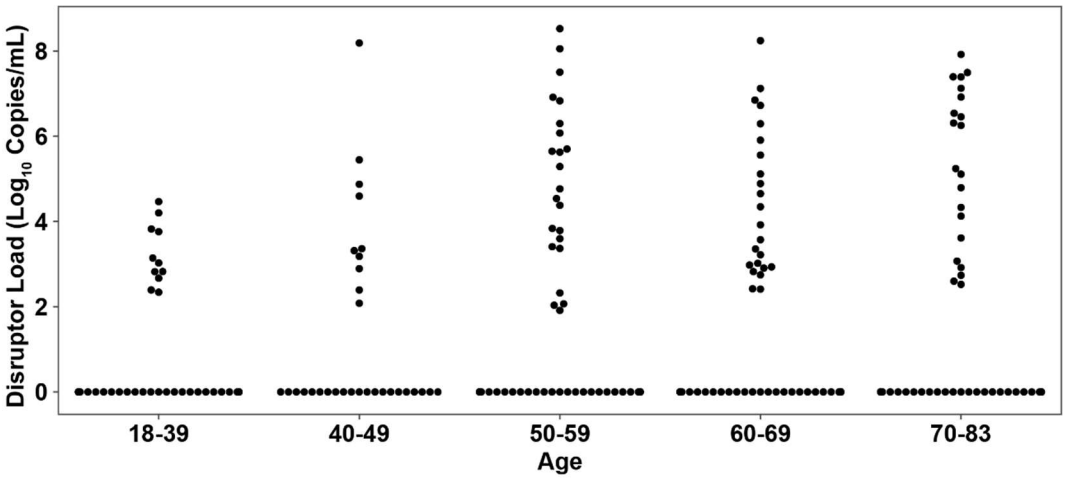


Figure S3.9: Disruptor taxa load separated by patient age: 18-39 (N=40), 40-49 (N=31), 50-59 (N=58), 60-69 (N=67), 70-83 (N=54).

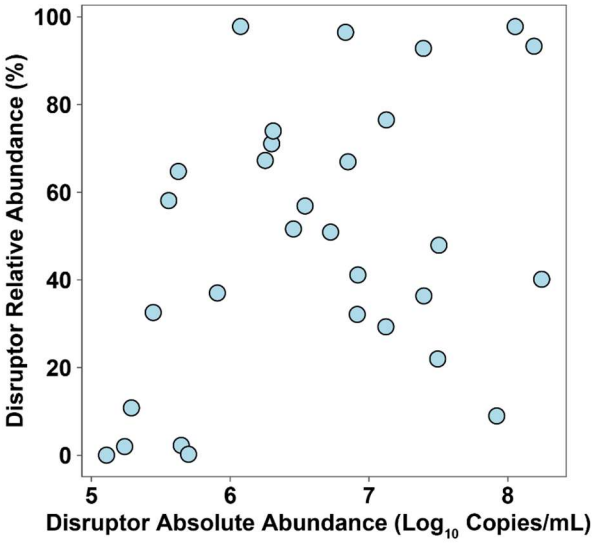


Figure S3.10: Relationship between absolute abundance (greater than 10^5 copies/mL) and relative abundance of disruptor loads (Spearman, P=0.09, not significant)

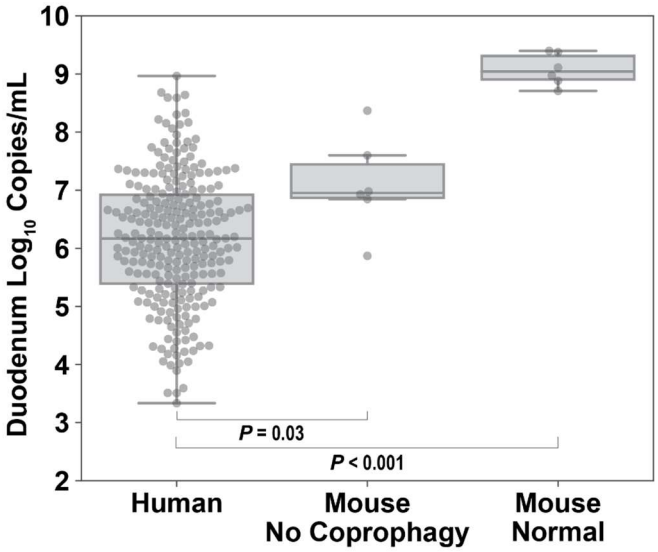


Figure S3.11: Comparison of total microbial load between human duodenum, mouse duodenum, and mouse duodenum where the mice had coprophagy prevented via tail cup. Mouse data from Bogatyrev et al. 2020[3]. Reported *P*-values are from Kruskal-Wallis test.

Table S3.1: Summary statistics for the patient cohort used in this study. All patients are from the REIMAGINE study[4].

Total subjects		**Total**	**SIBO**	**non-SIBO**
	Duodenal Aspirate	250	23% (58)	77% (192)
	Saliva	21	19% (4)	81% (17)
		Mean (Std Dev)		
		Total	**SIBO**	**non-SIBO**
Age		56.9 (14.9)	61.6 (13.6)	55.5 (15.0)
Weight (lbs)		169.4 (49.5)	166.1 (38.7)	170.4 (52.4)
		Percent (N)		
Gender		**Total**	**SIBO**	**non-SIBO**
	Male	46% (115)	38% (22)	48% (93)
	Female	54% (135)	62% (36)	52% (99)
Antibiotic usage				
	last 6 months	40% (100)	59% (34)	34% (66)

	current	4% (11)	9% (5)	3% (6)
PPI usage				
	PPI	34% (86)	36% (21)	34% (65)
	H2 blocker	4% (10)	7% (4)	3% (6)
	both	4% (10)	5% (3)	4% (7)
Any probiotic usage		20% (49)	26% (15)	18% (34)
Smokers		6% (16)	3% (2)	7% (14)
Symptom Scores				
	Bloating > 50th percentile	50%	57% (33)	45% (86)
	Constipation > 50th percentile	50%	53% (31)	46% (89)
	Incomplete Evacuation > 50th percentile	50%	55% (32)	47% (91)
	Excess Gas > 50th percentile	50%	57% (33)	45% (86)
	Diarrhea > 50th percentile	50%	55% (32)	46% (88)
Reason for Endoscopy				
	GERD/dyspepsia workup	21% (53)	16% (9)	23% (44)
	Possible bleeding/anemia workup	7% (17)	10% (6)	6% (11)
	Rule out cancer/polyp	24% (59)	29% (17)	22% (42)
	Biliary disease	13% (32)	16% (9)	12% (23)
	Dysphagia	10% (24)	9% (5)	10% (19)
	Crohn's disease	4% (10)	3% (2)	4% (8)
	Functional GI disease	5% (13)	2% (1)	6% (12)
	Rule out Celiac disease	1% (3)	3% (2)	1% (1)
	Known peptic ulcer disease	1% (3)	0% (0)	2% (3)
	G-tube management	1% (2)	2% (1)	1% (1)
	Other	3% (8)	2% (1)	4% (7)
	Missing Information	10% (26)	9% (5)	11% (21)

Table S3.2: *P*-values from significance tests (Kruskal-Wallis) comparing total microbial load between selected subgroups of individuals. Significance is indicated with an asterisk.

Comparison (N)	**p-value**
Taking probiotics (49) vs no probiotics (201)	0.97
Abx past 6 months (100) vs no Abx past 6 months (150)	0.67
Current Abx (11) vs not currently taking Abx (239)	**0.04***
Current Abx (11) vs Abx past 6 months (100)	**0.02***
Taking PPI (106) vs no PPI (144)	0.29
Current smoker (16) vs not currently smoker (234)	0.39

Table S3.3: Comparison between prevalence and relative abundance of all taxa in paired saliva and duodenum samples (N=21 participants).

# of Saliva samples taxon appears in (N=21 total)	# of Duodenum samples taxon appears in (N=21 total)	Saliva Rel. Abundance (%)	Duodenum Rel. Abundance (%)	Taxonomy	Difference between # of samples taxon is present in between Saliva and Duodenum (N=21 total)
21	8	0.22	0.20	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Flavobacteriales;D_4__Weeksellaceae;D_5__Bergeyella;D_6__uncultured bacterium	13
19	6	0.21	0.09	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;__	13
16	6	0.27	0.05	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Selenomonas 3;__	10
18	8	1.63	0.49	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Alloprevotella;D_6__uncultured Bacteroidetes bacterium	10
20	10	0.93	0.07	D_0__Bacteria;D_1__Epsilonbacteraeota;D_2__Campylobacteria;D_3__Campylobacterales;D_4__Campylobacteraceae;D_5__Campylobacter;__	10
15	5	0.32	0.05	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Flavobacteriales;D_4__Flavobacteriaceae;D_5__Capnocytophaga;__	10
12	3	0.04	0.01	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Flavobacteriales;D_4__	9

				Flavobacteriaceae;D_5__Capnocytophaga;D_6__Capnocytophaga gingivalis	
12	3	0.13	0.01	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteobacteriales;D_4__Burkholderiaceae;D_5__Lautropia;D_6__uncultured bacterium	9
15	6	0.28	0.05	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__Prevotella oris	9
13	4	0.09	0.04	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Cardiobacteriales;D_4__Cardiobacteriaceae;D_5__Cardiobacterium;D_6__uncultured bacterium	9
16	8	0.96	0.16	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteobacteriales;D_4__Neisseriaceae;D_5__Neisseria;__	8
11	3	0.56	0.07	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Alloprevotella;__	8
13	6	0.54	0.21	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptococcaceae;D_5__Streptococcus;D_6__Streptococcus mutans	7
16	9	1.73	0.37	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__Prevotella sp. oral taxon 299 str. F0039	7
14	7	0.10	0.02	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Family XIII;D_5__[Eubacterium] nodatum group;D_6__[Eubacterium] sulci	7
12	5	0.09	0.01	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Corynebacteriales;	7

				D_4__Corynebacteriaceae;D_5__Corynebacterium;__	
12	5	0.08	0.03	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;D_6__Prevotella sp. oral clone P4PB_83 P2	7
20	13	2.59	1.61	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Porphyromonadaceae;D_5__Porphyromonas;__	7
10	3	0.09	0.02	D_0__Bacteria;D_1__Epsilonbacteraeota;D_2__Campylobacteria;D_3__Campylobacterales;D_4__Campylobacteraceae;D_5__Campylobacter;D_6__Campylobacter rectus	7
8	1	0.06	0.00	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteobacteriales;D_4__Neisseriaceae;D_5__Kingella;D_6__uncultured bacterium	7
14	8	0.13	0.16	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptococcaceae;D_5__Streptococcus;D_6__Streptococcus anginosus subsp. anginosus	6
17	11	0.18	0.30	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Oribacterium;__	6
21	15	1.12	0.74	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Micrococcales;D_4__Micrococcaceae;D_5__Rothia;D_6__uncultured bacterium	6
7	1	0.06	0.03	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Alloprevotella;D_6__Alloprevotella tannerae	6
7	1	0.02	0.00	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Pre	6

				votellaceae;D_5__Prevotella;D_6__Prevotella sp. oral taxon G60	
17	11	0.30	0.20	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Corynebacteriales;D_4__Corynebacteriaceae;D_5__Corynebacterium;D_6__Corynebacterium durum	6
0	6	0.00	0.26	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptococcaceae;D_5__Streptococcus;D_6__Streptococcus pneumoniae	6
19	13	0.45	0.24	D_0__Bacteria;D_1__Actinobacteria;D_2__Coriobacteriia;D_3__Coriobacteriales;D_4__Atopobiaceae;D_5__Atopobium;D_6__uncultured bacterium	6
9	4	0.06	0.02	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Paludibacteraceae;D_5__F0058;D_6__uncultured bacterium	5
7	2	0.06	0.01	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;D_6__Prevotella sp. oral clone DO014	5
11	6	0.15	0.05	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteobacteriales;D_4__Neisseriaceae;D_5__Eikenella;D_6__uncultured bacterium	5
17	12	1.79	0.45	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 6;D_6__uncultured bacterium	5
8	3	0.04	0.02	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Tannerellaceae;D_5__Tannerella;D_6__uncultured bacterium	5

19	14	0.82	0.60	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;__	5
13	8	0.30	0.14	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptostreptococcaceae;D_5__Peptostreptococcus;D_6__uncultured organism	5
11	6	0.09	0.11	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Aerococcaceae;D_5__Abiotrophia;D_6__uncultured bacterium	5
19	14	0.51	0.33	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Stomatobaculum;D_6__uncultured bacterium	5
18	13	0.33	0.39	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Lachnoanaerobaculum;D_6__uncultured organism	5
16	11	0.20	0.05	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Family XIII;D_5__Mogibacterium;__	5
11	6	0.10	0.08	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;D_6__Prevotella denticola	5
6	2	0.05	0.01	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Bifidobacteriales;D_4__Bifidobacteriaceae;D_5__Alloscardovia;D_6__Bifidobacterium longum subsp. longum	4
5	1	0.01	0.00	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 2;D_6__unidentified	4
5	1	0.04	0.01	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteo	4

				bacteriales;D_4__Neisseriaceae;D_5__Neisseria;D_6__unidentified	
21	17	0.53	0.70	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;D_6__uncultured bacterium	4
5	1	0.02	0.00	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Flavobacteriales;D_4__Flavobacteriaceae;D_5__Capnocytophaga;D_6__Capnocytophaga granulosa	4
8	4	0.13	0.14	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Bifidobacteriales;D_4__Bifidobacteriaceae;D_5__Scardovia;D_6__unidentified	4
5	1	0.01	0.00	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 6;__	4
10	6	0.57	0.26	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Aggregatibacter;D_6__uncultured bacterium	4
5	1	0.03	0.02	D_0__Bacteria;D_1__Epsilonbacteraeota;D_2__Campylobacteria;D_3__Campylobacterales;D_4__Campylobacteraceae;D_5__Campylobacter;D_6__Campylobacter concisus	4
17	13	0.90	0.75	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Megasphaera;D_6__unidentified	4
11	7	0.05	0.04	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Family XIII;D_5__[Eubacterium] brachy group;D_6__Eubacterium brachy ATCC 33089	4

18	14	0.60	0.59	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Family XI;D_5__Parvimonas;__	4
12	8	0.11	0.23	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;D_6__Leptotrichia wadei F0279	4
19	15	0.38	0.39	D_0__Bacteria;D_1__Firmicutes;D_2__Erysipelotrichia;D_3__Erysipelotrichales;D_4__Erysipelotrichaceae;D_5__Solobacterium;__	4
19	15	0.17	0.20	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Lachnoanaerobaculum;D_6__uncultured bacterium	4
14	11	1.37	0.89	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__Prevotella pallens	3
6	3	0.18	0.03	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__Prevotella aurantiaca JCM 15754	3
18	15	9.50	4.54	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;D_6__Prevotella melaninogenica	3
10	7	0.05	0.04	D_0__Bacteria;D_1__Actinobacteria;D_2__Coriobacteriia;D_3__Coriobacteriales;D_4__Atopobiaceae;D_5__Atopobium;__	3
9	6	0.06	0.41	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Micrococcales;D_4__Micrococcaceae;D_5__Rothia;D_6__uncultured organism	3
19	16	3.57	2.22	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Betaproteo	3

				bacteriales;D_4__Neisseriaceae;D_5__Neisseria;D_6__uncultured bacterium	
21	18	2.79	4.05	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Bacillales;D_4__Family XI;D_5__Gemella;__	3
8	5	1.20	0.12	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Actinobacillus;__	3
8	5	0.02	0.07	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;__;__	3
20	17	4.01	3.57	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;__	3
6	3	0.17	0.05	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Aggregatibacter;__	3
6	4	0.04	0.01	D_0__Bacteria;D_1__Actinobacteria;D_2__Coriobacteriia;D_3__Coriobacteriales;D_4__Atopobiaceae;D_5__Atopobium;D_6__uncultured Actinomyces sp.	2
6	4	0.06	0.02	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Peptococcaceae;D_5__Peptococcus;__	2
18	16	0.31	0.73	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Oribacterium;D_6__Oribacterium sinus	2
16	14	0.09	0.28	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Catonella;D_6__uncultured bacterium	2
20	18	6.27	3.21	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptoco	2

				ccaceae;D_5__Streptococcus;D_6__Streptococcus salivarius subsp. thermophilus	
4	2	0.01	0.00	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;D_6__Leptotrichia sp. oral clone EI022	2
7	5	0.10	0.02	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;__;__	2
7	9	0.03	0.15	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Ruminococcaceae UCG-014;__	2
9	7	0.08	0.11	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Oribacterium;D_6__Oribacterium parvum ACB1	2
10	8	0.49	0.33	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;D_6__Leptotrichia sp. oral clone FP036	2
12	10	0.06	0.06	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Dialister;__	2
13	12	0.36	0.36	D_0__Bacteria;D_1__Actinobacteria;D_2__Actinobacteria;D_3__Micrococcales;D_4__Micrococcaceae;D_5__Rothia;__	1
4	5	0.01	0.01	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Shuttleworthia;__	1
6	5	0.19	0.10	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Lactobacillaceae;D_5__Lactobacillus;__	1

21	20	2.62	2.38	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Carnobacteriaceae;D_5__Granulicatella;__	1
7	6	0.02	0.02	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Selenomonas;D_6__uncultured bacterium	1
5	6	0.07	0.09	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Johnsonella;D_6__uncultured bacterium	1
19	18	4.94	2.80	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pasteurellales;D_4__Pasteurellaceae;D_5__Haemophilus;__	1
4	5	0.02	0.04	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Ruminococcaceae UCG-014;D_6__Clostridiales bacterium oral taxon 075	1
4	5	0.04	0.03	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Butyrivibrio 2;D_6__Eubacterium sp. oral clone GI038	1
3	3	0.04	0.16	D_0__Bacteria;D_1__Tenericutes;D_2__Mollicutes;D_3__Mollicutes RF39;D_4__uncultured bacterium;D_5__uncultured bacterium;D_6__uncultured bacterium	0
3	3	0.02	0.02	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Leptotrichiaceae;D_5__Leptotrichia;D_6__Leptotrichia buccalis C-1013-b	0
3	3	0.11	0.01	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;__;__	0

5	5	0.03	0.03	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae;D_5__Lachnoanaerobaculum;__	0
21	21	9.62	8.29	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Veillonella;__	0
10	10	0.05	0.19	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella;D_6__Prevotella nigrescens	0
21	21	3.92	5.89	D_0__Bacteria;D_1__Fusobacteria;D_2__Fusobacteriia;D_3__Fusobacteriales;D_4__Fusobacteriaceae;D_5__Fusobacterium;__	0
5	5	0.05	0.03	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Prevotellaceae;D_5__Prevotella 7;D_6__Prevotella veroralis DSM 19559 = JCM 6290	0
21	21	23.94	39.08	D_0__Bacteria;D_1__Firmicutes;D_2__Bacilli;D_3__Lactobacillales;D_4__Streptococcaceae;D_5__Streptococcus;__	0

Table S3.4: Two groups of taxa (light blue and dark blue) that have stronger co-correlations with another taxon than with total load. Significance values for all correlations and co-correlations were $P < 0.001$.

Taxon 1	Taxon 2	Co-Correlation	Correlation with Total Load	Difference	Biological Link
Alloprevotella	*Prevotella*	0.73	0.30	0.43	Tertiary plaque biofilm colonizers. Metabolize same byproduct of primary colonizers[5, 6].
Prevotella 6	*Prevotella 7*	0.83	0.43	0.39	
Porphyromonas	*Prevotella*	0.74	0.36	0.38	
Prevotella	*Prevotella 7*	0.82	0.50	0.32	
Megasphaera	*Solobacterium*	0.69	0.43	0.26	Not Known
Solobacterium	*Oribacterium*	0.80	0.55	0.25	
Leptotrichia	*Oribacterium*	0.81	0.60	0.20	
Atopobium	*Solobacterium*	0.78	0.61	0.17	

Supplementary References

1. Elazhary, M.A., Saheb, S.A., Roy, R.S. & Lagacé, A. A simple procedure for the preliminary identification of aerobic gram negative intestinal bacteria with special reference to the Enterobacteriaceae. *Canadian journal of comparative medicine : Revue canadienne de medecine comparee* **37**, 43-46 (1973).

2. Ruoff, K.L. Miscellaneous Catalase-Negative, Gram-Positive Cocci: Emerging Opportunists. *J. Clin. Microbiol.* **40**, 1129 (2002).

3. Bogatyrev, S.R., Rolando, J.C. & Ismagilov, R.F. Self-reinoculation with fecal flora changes microbiota density and composition leading to an altered bile-acid profile in the mouse small intestine. *Microbiome* **8**, 19 (2020).

4. Leite, G.G.S. et al. Mapping the Segmental Microbiomes in the Human Small Bowel in Comparison with Stool: A REIMAGINE Study. *Dig. Dis. Sci.* **65**, 2595-2604 (2020).

5. Marcotte, H. & Lavoie, M.C. Oral microbial ecology and the role of salivary immunoglobulin A. *Microbiology and molecular biology reviews : MMBR* **62**, 71-109 (1998).

6. Hojo, K., Nagaoka, S., Ohshima, T. & Maeda, N. Bacterial Interactions in Dental Biofilm Development. *J. Dent. Res.* **88**, 982-990 (2009).