Date: September 28th, 2020

Cancer Immune monitoring and Analysis Center
Department of Translational Molecular Pathology
The University of Texas at MD Anderson Cancer Center
Contact PI: Ignacio Ivan Wistuba

Ignacio Ivan Wistuba, MD, Professor, and Chair, Translational Molecular Pathology
Robert Jenq, M.D., Director Microbiome Core, Assistant Professor, and Deputy Chair Genomic Medicine
Samuel Shelburne, M.D., Ph.D., Co-Director Microbiome Core, Professor and Deputy Chair, Infectious Diseases
J.Jack Lee M.S., Ph.D., Associate VP, Quantitative Sciences, Biostatistics
Chia-Chi (Tina) Chang, Ph.D., Research Scientist, Microbiome Core Facility Manager
Jiexin Zhang, M.C.S., M.S., Principal Bioinformatician, Bioinformatics & Computational Biology

16S V4 ribosomal RNA (rRNA) Sequencing Analytical Validation

Version 1.0

This report describes the analytical validation parameters for 16S V4 ribosomal RNA (rRNA) sequencing assay performed at MD Anderson Cancer Center on frozen stool samples from patients and healthy donors.

16S rRNA gene sequencing assay performance	
Processing/QC	Nuclease-free water and ZymoBIOMICS standard mock community were co-extracted along with fecal samples as control. Nuclease-free water control did not amplify after PCR amplification.
Intra-Assay Reproducibility	The intra-assay reproducibility was performed using eight healthy donors aliquoted five times and sequenced within the same sequencing run. The mean reads are 158736. Principal coordinate analyses (PCoA), pairwise analysis at the genus level showed high intra-assay reproducibility between aliquots despite inter-sample variability with high correlation value, indicating high reproducibility and high consistency each aliquot within the same sample. The Spearman's correlation for genus data from all healthy donors > 0.8. All samples had values within 2-SD distribution, suggesting good precision.
Inter-Assay Reproducibility	The precision and reproducibility of the assay were determined with three ZymoBIOMICS standard aliquots and analyzed in three independent sequencing runs. High reproducibility and consistency and low standard deviation in relative abundance of taxonomy were observed in specimens across three separate sequencing runs. The Spearman's correlation was calculated for genus data from each patient. The average read counts were 74552. Spearman's correlation coefficient rho > 0.7 and SD distribution for all genera <1.5
Analytical Accuracy	Accuracy was determined with three ZymoBIOMICS standard mock community (known composition of microbes). Eight major genera were detected with significant similarity to provided bacterial composition from Zymo Research.
Patient Normal Comparison	The average counts for patients were 143560, while healthy donors were 158736. Spearman's correlation for genus data for all healthy donors and three patients had rho > 0.8; two patients had rho> 0.7. All samples had values within 2-SD distribution, suggesting good precision. Patient samples showed more outliers with higher SD compared to healthy donors.
Performance characteristics	All of the required instrument has obtained annual contracts with regular calibration and maintenance performed by vendors as part of quality control.

Introduction

Microbiome analysis was performed on 25 frozen stool samples obtained from healthy donors and 3 Zymo mock microbial communities. Bacterial DNA was extracted, and the V4 hypervariable region of the bacterial 16S rRNA gene was amplified. The pooled libraries were co-sequenced with PhiX control on the Illumina platform[1]. Sequencing data from paired-end reads were de-multiplexed and analyzed to determine the alpha diversity, beta-diversity, and taxonomic classification. Statistical analysis was performed for the data.

1. Analytes

1) 15 samples from 3 healthy donors (5 aliquots for each sample) were received from the Icahn School of Medicine at Mount Sinai for harmonization analysis (Table 1).

2) 25 samples from 5 healthy donors (5 aliquots for each sample) and 25 samples from 5 patients (5 aliquots for each sample) who underwent stem cell transplant. Patient #1, #4 and #5 have received antibiotics (Levofloxacin) during stool collection. Patient #5 has received two additional, vancomycin and cefepime, during stool collection. These 25 samples were collected at the University of MD Anderson Cancer Center (Table 1).

3) 3 aliquots of Zymo Mock Microbial Community standard were purchased from Zymo Research.

Table 1. Sample information

Batch	Sample.ID	Sample Type	SampleID	Primer	Sample Description	Approximate amount of stools (mg)
Batch1	Zymo.mock.community.batch1	Inter	P2020.42.515rcbc1.147	515rcbc1.147	Mock community	50 µL of community stock
Batch2	Zymo.mock.community.batch2	Inter	P2019.38.515rcbc3.096	515rcbc3.096	Mock community	50 µL of community stock
Batch3	Zymo.mock.community.batch3	Inter	P2020.43.515rcbc3.059	515rcbc3.059	Mock community	50 µL of community stock
Batch1	P.MDA.1.1	Intra	P2020.42.515rcbc1.097	515rcbc1.097	Patient Sample #1	74.4
Batch1	P.MDA.1.2	Intra	P2020.42.515rcbc1.098	515rcbc1.098	Patient Sample #1	84.3
Batch1	P.MDA.1.3	Intra	P2020.42.515rcbc1.099	515rcbc1.099	Patient Sample #1	92.6
Batch1	P.MDA.1.4	Intra	P2020.42.515rcbc1.100	515rcbc1.100	Patient Sample #1	76.4
Batch1	P.MDA.1.5	Intra	P2020.42.515rcbc1.101	515rcbc1.101	Patient Sample #1	83.5
Batch1	P.MDA.2.1	Intra	P2020.42.515rcbc1.102	515rcbc1.102	Patient Sample #2	100.6
Batch1	P.MDA.2.2	Intra	P2020.42.515rcbc1.103	515rcbc1.103	Patient Sample #2	107.9
Batch1	P.MDA.2.3	Intra	P2020.42.515rcbc1.104	515rcbc1.104	Patient Sample #2	91.2
Batch1	P.MDA.2.4	Intra	P2020.42.515rcbc1.105	515rcbc1.105	Patient Sample #2	93.1
Batch1	P.MDA.2.5	Intra	P2020.42.515rcbc1.106	515rcbc1.106	Patient Sample #2	90.4
Batch1	P.MDA.3.1	Intra	P2020.42.515rcbc1.107	515rcbc1.107	Patient Sample #3	73.0
Batch1	P.MDA.3.2	Intra	P2020.42.515rcbc1.108	515rcbc1.108	Patient Sample #3	61.3
Batch1	P.MDA.3.3	Intra	P2020.42.515rcbc1.109	515rcbc1.109	Patient Sample #3	77.4
Batch1	P.MDA.3.4	Intra	P2020.42.515rcbc1.110	515rcbc1.110	Patient Sample #3	66.3
Batch1	P.MDA.3.5	Intra	P2020.42.515rcbc1.111	515rcbc1.111	Patient Sample #3	82.0
Batch1	P.MDA.4.1	Intra	P2020.42.515rcbc1.112	515rcbc1.112	Patient Sample #4	79.6
Batch1	P.MDA.4.2	Intra	P2020.42.515rcbc1.113	515rcbc1.113	Patient Sample #4	86.1
Batch1	P.MDA.4.3	Intra	P2020.42.515rcbc1.114	515rcbc1.114	Patient Sample #4	95.0
Batch1	P.MDA.4.4	Intra	P2020.42.515rcbc1.115	515rcbc1.115	Patient Sample #4	88.2
Batch1	P.MDA.4.5	Intra	P2020.42.515rcbc1.116	515rcbc1.116	Patient Sample #4	89.5
Batch1	P.MDA.5.1	Intra	P2020.42.515rcbc1.117	515rcbc1.117	Patient Sample #5	82.6
Batch1	P.MDA.5.2	Intra	P2020.42.515rcbc1.118	515rcbc1.118	Patient Sample #5	84.3
Batch1	P.MDA.5.3	Intra	P2020.42.515rcbc1.119	515rcbc1.119	Patient Sample #5	84.4

Batch1	P.MDA.5.4	Intra	P2020.42.515rcbc1.120	515rcbc1.120	Patient Sample #5	78.2
Batch1	P.MDA.5.5	Intra	P2020.42.515rcbc1.121	515rcbc1.121	Patient Sample #5	87.9
Batch1	HD.MDA.1.1	Intra	P2020.42.515rcbc1.122	515rcbc1.122	Heathy Donor #1	77.5
Batch1	HD.MDA.1.2	Intra	P2020.42.515rcbc1.123	515rcbc1.123	Heathy Donor #1	83.0
Batch1	HD.MDA.1.3	Intra	P2020.42.515rcbc1.124	515rcbc1.124	Heathy Donor #1	82.9
Batch1	HD.MDA.1.4	Intra	P2020.42.515rcbc1.125	515rcbc1.125	Heathy Donor #1	91.5
Batch1	HD.MDA.1.5	Intra	P2020.42.515rcbc1.126	515rcbc1.126	Heathy Donor #1	87.9
Batch1	HD.MDA.2.1	Intra	P2020.42.515rcbc1.127	515rcbc1.127	Heathy Donor #2	85.2
Batch1	HD.MDA.2.2	Intra	P2020.42.515rcbc1.128	515rcbc1.128	Heathy Donor #2	77.2
Batch1	HD.MDA.2.3	Intra	P2020.42.515rcbc1.129	515rcbc1.129	Heathy Donor #2	82.5
Batch1	HD.MDA.2.4	Intra	P2020.42.515rcbc1.130	515rcbc1.130	Heathy Donor #2	77.6
Batch1	HD.MDA.2.5	Intra	P2020.42.515rcbc1.131	515rcbc1.131	Heathy Donor #2	95.2
Batch1	HD.MDA.3.1	Intra	P2020.42.515rcbc1.132	515rcbc1.132	Heathy Donor #3	80.1
Batch1	HD.MDA.3.2	Intra	P2020.42.515rcbc1.133	515rcbc1.133	Heathy Donor #3	81.4
Batch1	HD.MDA.3.3	Intra	P2020.42.515rcbc1.134	515rcbc1.134	Heathy Donor #3	86.7
Batch1	HD.MDA.3.4	Intra	P2020.42.515rcbc1.135	515rcbc1.135	Heathy Donor #3	96.4
Batch1	HD.MDA.3.5	Intra	P2020.42.515rcbc1.136	515rcbc1.136	Heathy Donor #3	91.8
Batch1	HD.MDA.4.1	Intra	P2020.42.515rcbc1.137	515rcbc1.137	Heathy Donor #4	86.9
Batch1	HD.MDA.4.2	Intra	P2020.42.515rcbc1.138	515rcbc1.138	Heathy Donor #4	92.3
Batch1	HD.MDA.4.3	Intra	P2020.42.515rcbc1.139	515rcbc1.139	Heathy Donor #4	89.0
Batch1	HD.MDA.4.4	Intra	P2020.42.515rcbc1.140	515rcbc1.140	Heathy Donor #4	89.7
Batch1	HD.MDA.4.5	Intra	P2020.42.515rcbc1.141	515rcbc1.141	Heathy Donor #4	85.1
Batch1	HD.MDA.5.1	Intra	P2020.42.515rcbc1.142	515rcbc1.142	Heathy Donor #5	97.6
Batch1	HD.MDA.5.2	Intra	P2020.42.515rcbc1.143	515rcbc1.143	Heathy Donor #5	82.5
Batch1	HD.MDA.5.3	Intra	P2020.42.515rcbc1.144	515rcbc1.144	Heathy Donor #5	85.1
Batch1	HD.MDA.5.4	Intra	P2020.42.515rcbc1.145	515rcbc1.145	Heathy Donor #5	85.2
Batch1	HD.MDA.5.5	Intra	P2020.42.515rcbc1.146	515rcbc1.146	Heathy Donor #5	89.6
Batch1	HD-7003.1	Intra	P2020.42.515rcbc1.169	515rcbc1.169	Healthy Donor HD7003	16.8
Batch1	HD-7003.2	Intra	P2020.42.515rcbc1.170	515rcbc1.170	Healthy Donor HD7003	40.0
Batch1	HD-7003.3	Intra	P2020.42.515rcbc1.171	515rcbc1.171	Healthy Donor HD7003	61.3
Batch1	HD-7003.4	Intra	P2020.42.515rcbc1.172	515rcbc1.172	Healthy Donor HD7003	19.2
Batch1	HD-7003.5	Intra	P2020.42.515rcbc1.173	515rcbc1.173	Healthy Donor HD7003	46.2
Batch1	HD-7001.1	Intra	P2020.42.515rcbc1.174	515rcbc1.174	Healthy Donor HD7001	37.7
Batch1	HD-7001.2	Intra	P2020.42.515rcbc1.175	515rcbc1.175	Healthy Donor HD7001	74.2
Batch1	HD-7001.3	Intra	P2020.42.515rcbc1.176	515rcbc1.176	Healthy Donor HD7001	37.8
Batch1	HD-7001.4	Intra	P2020.42.515rcbc1.177	515rcbc1.177	Healthy Donor HD7001	68.7
Batch1	HD-7001.5	Intra	P2020.42.515rcbc1.178	515rcbc1.178	Healthy Donor HD7001	76.4
Batch1	HD-7002.1	Intra	P2020.42.515rcbc1.179	515rcbc1.179	Healthy Donor HD7002	10.1
Batch1	HD-7002.2	Intra	P2020.42.515rcbc1.180	515rcbc1.180	Healthy Donor HD7002	4.5
Batch1	HD-7002.3	Intra	P2020.42.515rcbc1.181	515rcbc1.181	Healthy Donor HD7002	2.8
Batch1	HD-7002.4	Intra	P2020.42.515rcbc1.182	515rcbc1.182	Healthy Donor HD7002	17.0
Batch1	HD-7002.5	Intra	P2020.42.515rcbc1.183	515rcbc1.183	Healthy Donor HD7002	8.8

2.1 DNA Isolation

Bacterial DNA was extracted from pre-weighed stools using zirconium beads in InhibitEX lysis buffer following protocol mentioned in the MDACC microbiome SOP (in a separate document)[2].

2.2 DNA quantitation and quality control

DNA concentration was measured using a Nanodrop Spectrophotometer and UV plate quantitation method[3] (Table 2). Direct measurements of DNA samples at OD260 can be converted to concentration using the Beer-Lambert law. The equation for calculating

concentration for nucleic acids: Nucleic Acid Concentration = OD260/path length x standard coefficient 50 for double-stranded DNA x sample dilution.

Table 2: DNA concentration obtained from each sample.

Sample ID	DNA Conc.	Sample ID	DNA Conc.	Sample ID	DNA Conc.
P.MDA.1.1	263.4	P.MDA.5.3	100	HD.MDA.4.5	55.6
P.MDA.1.2	119.2	P.MDA.5.4	65.2	HD.MDA.5.1	157.9
P.MDA.1.3	111.1	P.MDA.5.5	110.3	HD.MDA.5.2	70.2
P.MDA.1.4	201.6	HD.MDA.1.1	784.8	HD.MDA.5.3	151.3
P.MDA.1.5	180.4	HD.MDA.1.2	733.3	HD.MDA.5.4	60.7
P.MDA.2.1	70.7	HD.MDA.1.3	596.1	HD.MDA.5.5	92
P.MDA.2.2	543.2	HD.MDA.1.4	624.1	HD-7003.1	85.7
P.MDA.2.3	233	HD.MDA.1.5	574.2	HD-7003.2	96.6
P.MDA.2.4	173.1	HD.MDA.2.1	241.5	HD-7003.3	147
P.MDA.2.5	196	HD.MDA.2.2	475.9	HD-7003.4	80.8
P.MDA.3.1	72	HD.MDA.2.3	463.4	HD-7003.5	131.3
P.MDA.3.2	58.6	HD.MDA.2.4	316.5	HD-7001.1	116.4
P.MDA.3.3	96.3	HD.MDA.2.5	162.9	HD-7001.2	194.7
P.MDA.3.4	58.1	HD.MDA.3.1	156.9	HD-7001.3	105.4
P.MDA.3.5	86.6	HD.MDA.3.2	47.6	HD-7001.4	146.7
P.MDA.4.1	159.8	HD.MDA.3.3	141.4	HD-7001.5	200.7
P.MDA.4.2	31	HD.MDA.3.4	72.3	HD-7002.1	77.2
P.MDA.4.3	32.8	HD.MDA.3.5	190.2	HD-7002.2	70.7
P.MDA.4.4	33.5	HD.MDA.4.1	247	HD-7002.3	70
P.MDA.4.5	29.5	HD.MDA.4.2	251.6	HD-7002.4	77.9
P.MDA.5.1	89.3	HD.MDA.4.3	114.8	HD-7002.5	75.7
P.MDA.5.2	83	HD.MDA.4.4	78.6	Nuclease free water	0
Zymo mock 1	7	Zymo Mock 2	7.7	Zymo Mock 3	7.5

Concentration in ng/μl

Note: Mock community standards are the mixtures of inactivated microorganisms made by Zymo research. It is not surprising that the standards have a low amount of DNA compared to stool samples. Stool contains a range of DNA, e.g., host DNA from colon epithelial cells, parasite DNA, bacterial DNA, DNA from food, or DNA from gastrointestinal.

2.3 Amplicon amplification

The hypervariable V4 region of the 16S rRNA gene was amplified by PCR of an approximate 400 bp, using extracted, and purified genomic DNA from stool samples.

2.4 Library Quality Control (QC) and High-Throughput Library Pooling for NGS

Amplicon quality and quantity were assessed with the Agilent D1000 DNA ScreenTape assay on an Agilent 4200 TapeStation system. The barcoded amplicons were normalized in equal concentrations and pooled into one single tube using the epMotion system. Final library quality and quantity were assessed with the Qubit dsDNA BR Assay Kit and qPCR on QuantStudio 6 Flex Real-Time PCR Systems. The molarity was calculated based on the size of the amplicon. The sequencing run was performed using a 2 x 250 bp paired-end protocol on the Illumina Miseq platform.

3. Data Analyses and Bioinformatic pipeline

Sequencing data from 2 x 250 bp paired-end reads were de-multiplexed and split using QIIME[4]. Merging of paired-end reads creates consensus sequences using VSEARCH v7, allowing up to a maximum of 10 mismatched. The cluster_otus command, an implementation of the UPARSE algorithm, was used to perform 97% related operational taxonomic units (OTU) clustering. Denoising was done by unoise3 command, allowing to identify all correct biological sequences in the reads. A denoised sequence is called a zero-radius OTU (ZOTU). Mothur with Silva database[5] was then used for taxonomic assignment. The alpha_diversity.py script in QIIME was then used to estimate alpha diversity. A phylogenetic tree was generated using a FastTree method in the

QIIME package and considered a cluster of related OTU (crOTUs). The abundance of each crOTU was calculated as the sum of abundances of its member OTUs.

3.1. Bioassay performance

Intra- Assay reproducibility: The intra-assay reproducibility was performed using three healthy donors received from the Icahn School of Medicine at Mount Sinai, five healthy donors at MD Anderson Cancer Center, and five patients at MD Anderson Cancer Center (Table 1). Each individual sample has five aliquots and sequenced within the same sequencing run. The highest average target coverage was found in HD.MDA.2.3 (MD Anderson donor 2, aliquot 2) compared to other samples. The mean reads are 152,899 (Figure 1, Figure 2, and Table 3). Principal coordinate analyses (PCoA), also known as metric multidimensional scaling, showed high intra-assay reproducibility between aliquots despite inter-sample variability (Figure 3, Figure 5, and Figure 7). The number of OTU observed in the sample showed high reproducibility across aliquots (Figure 5 and Figure 7). Pairwise analysis at the genus level also showed a high correlation value, indicating high reproducibility and high consistency between each aliquot within the same sample (Figure 6 and Figure 9). The stacked bar plot of bacterial composition at the genus level showed high intra-assay reproducibility (Figure 5 and Figure 8). (Note: The counts of HD-7003 is much smaller than HD-7001 and HD-7002, which could be due to the condition of the donors during sample collection. It also noted that normal controls have higher OTU than patients. This possibly could be due to diet habits and treatments in the patients, which can shift microbiome diversity in the host. The 16S sequencing is sensitive enough to capture more than 1000 of OTUs in human samples. All five patients underwent a stem cell transplant. Patient #1, #4, and #5 have received antibiotics (Levofloxacin) during stool collection. Patient #5 has received two additional, Vancomycin and Cefepime, during stool collection and showed much lower oberseved OTU.)

Inter- Assay reproducibility and Precision (run-to-run variation): Accuracy was determined by the analysis of mock communities with known amounts of microbes (Table 4). To test the precision and reproducibility of the assay, DNA was extracted from three ZymoBIOMICS standard aliquots and analyzed in three independent sequencing runs. High reproducibility and consistency and low standard deviation in the relative abundance of taxonomy were observed in specimens across three separate sequencing runs (Figure 4). The microbiome composition profile was determined based on sequence counts after mapping amplicon sequences to the Silva ribosomal database. The bar plot of taxonomy composition indicates each detected genus in the ZymoBIOMICS Microbial Community Standard, which is consistent with the datasheet provided by Zymo Research (Table 4).

Patient – Normal comparisons: Five healthy donors and five patient samples were tested within the same sequencing run. The analysis was performed on the output data to compare the microbiome composition between healthy donors and patient samples (Figure 7 and Figure 8).

3.2. Evaluation of procedure

Sensitivity

Analytical sensitivity depends on sequencing depth, DNA extraction method, the amount of bacterial DNA, and bacterial diversity. The Illumina Miseq instrument can generate approximately 24 million reads per run passing filter for 250bp paired-end sequencing, which is around 30K read counts per sample if we pool around 350 amplicons in a run. Increasing the sample size would lower the read counts and sequencing depth. The core does not pool more than 350 samples within the same sequencing run to ensure enough read depth for bacterial identification (Table 3).

Specificity

Sequence reads are clustered into operational taxonomic units (OTUs) at a defined threshold to reduce errors generated during PCR amplification and sequencing. Three denoising packages, UNOISE3, Deblur, and DADA2, improved quality-filtering and correct sequencing errors.

Unoise3[6] has been implemented into our current pipeline, which runs faster than DADA2 and Deblur, respectively[7,8].

Accuracy

Accuracy was determined by a mixed microbial community of well-defined composition. ZymoBIOMICS standard contains a composition of eight validated microbes and two yeast strains. To test the accuracy of the assay, DNA was extracted from the ZymoBIOMICS standard with known amounts of microbes to determine the accuracy (Table 4). Overall, we found a consistent mock community composition across three mock community aliquots in three independent sequencing runs (Figure 4).

Rarefaction analysis

Rarefaction allows the calculation of species richness and estimates numbers of species for individual samples to meaningfully estimate and compare alpha diversity. Rarefaction was done with vegan package in R (R version 3.6.0, vegan_2.5-6). In order to provide a more robust estimate of species richness, we removed OTUs with only 1 read in a sample. To demonstrate adequate read depth to detect most species present, we calculated the percentage of detected by dividing the rarefied number of species by the estimated asymptotic richness (Figure 10). The percentage of detected is above 80% in all samples, indicating most taxa are recovered.

Table 3. Average coverage in healthy donors and patient samples

Run	Sample ID	Counts	Mean	Run	Sample ID	Counts	Mean	Run	Sample ID	Counts	Mean
Batch1	P.MDA.1.1	139076		Batch1	HD.MDA.1.1	148897		Batch1	HD-7003.1	122599	
Batch1	P.MDA.1.2	188167		Batch1	HD.MDA.1.2	127398		Batch1	HD-7003.2	156791	
Batch1	P.MDA.1.3	153248	158604	Batch1	HD.MDA.1.3	186154	162067	Batch1	HD-7003.3	151654	134641
Batch1	P.MDA.1.4	165386		Batch1	HD.MDA.1.4	164794		Batch1	HD-7003.4	108917	
Batch1	P.MDA.1.5	147146		Batch1	HD.MDA.1.5	183092		Batch1	HD-7003.5	133244	
Batch1	P.MDA.2.1	142960		Batch1	HD.MDA.2.1	117698		Batch1	HD-7001.1	128644	
Batch1	P.MDA.2.2	220744		Batch1	HD.MDA.2.2	179893		Batch1	HD-7001.2	147356	
Batch1	P.MDA.2.3	142635	167603	Batch1	HD.MDA.2.3	279703	201215	Batch1	HD-7001.3	151732	161374
Batch1	P.MDA.2.4	176746		Batch1	HD.MDA.2.4	195864		Batch1	HD-7001.4	174097	
Batch1	P.MDA.2.5	154932		Batch1	HD.MDA.2.5	232919		Batch1	HD-7001.5	205041	
Batch1	P.MDA.3.1	201067		Batch1	HD.MDA.3.1	207073		Batch1	HD-7002.1	235793	
Batch1	P.MDA.3.2	185402		Batch1	HD.MDA.3.2	109800		Batch1	HD-7002.2	183400	
Batch1	P.MDA.3.3	118893	141497	Batch1	HD.MDA.3.3	76360	125428	Batch1	HD-7002.3	173622	169629
Batch1	P.MDA.3.4	108046		Batch1	HD.MDA.3.4	102016		Batch1	HD-7002.4	153493	
Batch1	P.MDA.3.5	94081		Batch1	HD.MDA.3.5	131891		Batch1	HD-7002.5	101839	
Batch1	P.MDA.4.1	101156		Batch1	HD.MDA.4.1	173567		**Total**	**Mean**	**151418**	
Batch1	P.MDA.4.2	124607	110732	Batch1	HD.MDA.4.2	199340	161662				
Batch1	P.MDA.4.3	84982		Batch1	HD.MDA.4.3	121030					
Batch1	P.MDA.4.4	126321		Batch1	HD.MDA.4.4	117654					

Batch1	P.MDA.4.5	116596		Batch1	HD.MDA.4.5	196719	
Batch1	P.MDA.5.1	191386		Batch1	HD.MDA.5.1	105907	
Batch1	P.MDA.5.2	110538		Batch1	HD.MDA.5.2	188448	
Batch1	P.MDA.5.3	167633	139365	Batch1	HD.MDA.5.3	229814	153873
Batch1	P.MDA.5.4	132990		Batch1	HD.MDA.5.4	104835	
Batch1	P.MDA.5.5	94281		Batch1	HD.MDA.5.5	140361	

Run	Sample ID	Counts
Batch1	Mock Community	78990
Batch2	Mock Community	73049
Batch3	Mock Community	71528
Total	**Mean**	**74552**

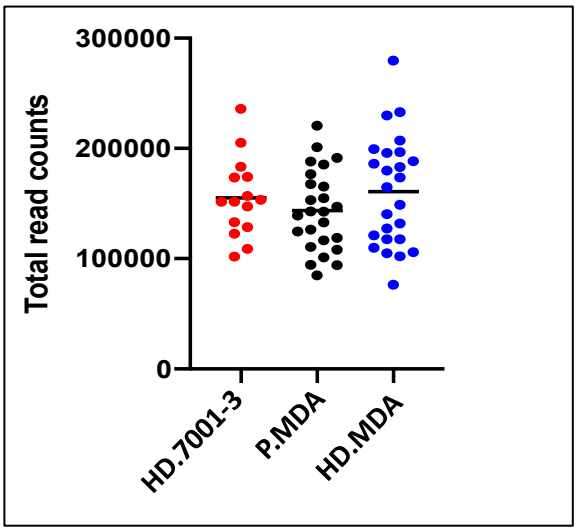


Figure 1. Total read counts on healthy donor samples (HD7001, HD7002, HD7003, and HD.MDA) and patient samples (P.MDA) collected from two institutions, the Icahn School of Medicine at Mount Sinai and MD Anderson Cancer Center.

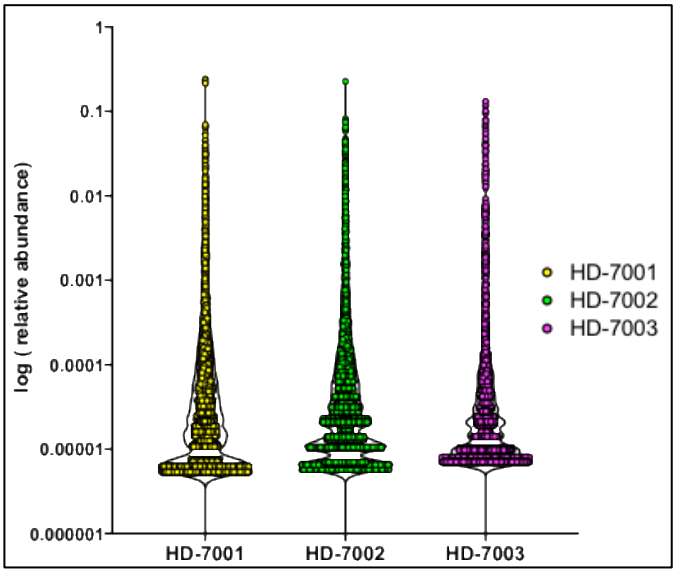


Figure 2. The relative abundance of all OTUs from three healthy donors received from the Icahn School of Medicine at Mount Sinai.

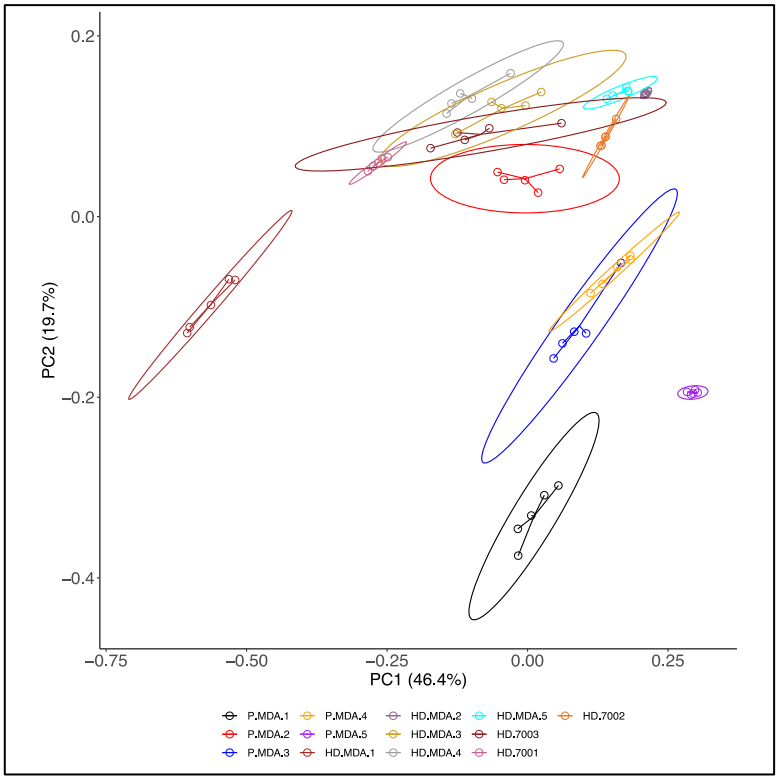


Figure 3. Beta-diversity, with weighted-UniFrac distances, was used to generate a principal coordinate analysis (PCoA) for all samples collected from the Icahn School of Medicine at Mount Sinai and MD Anderson Cancer Center.

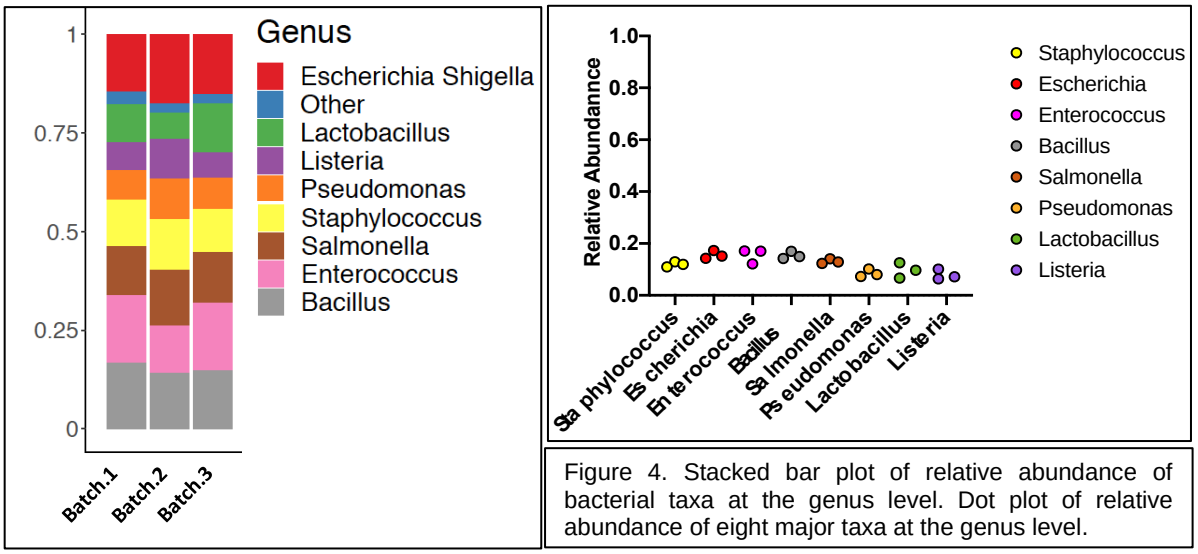
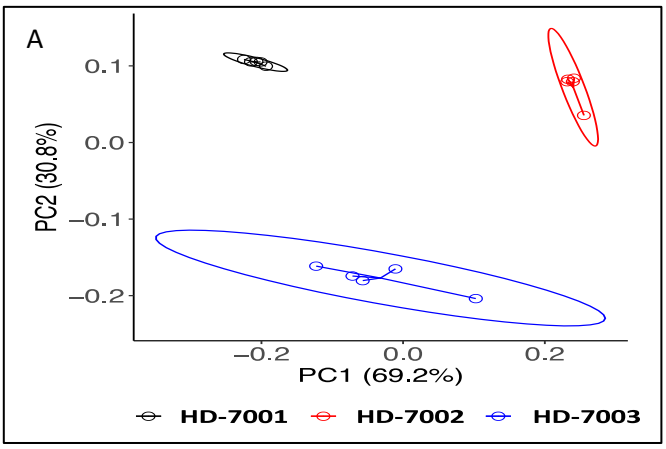
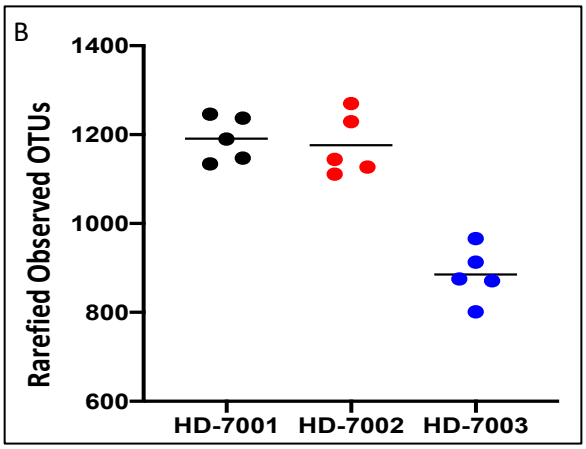


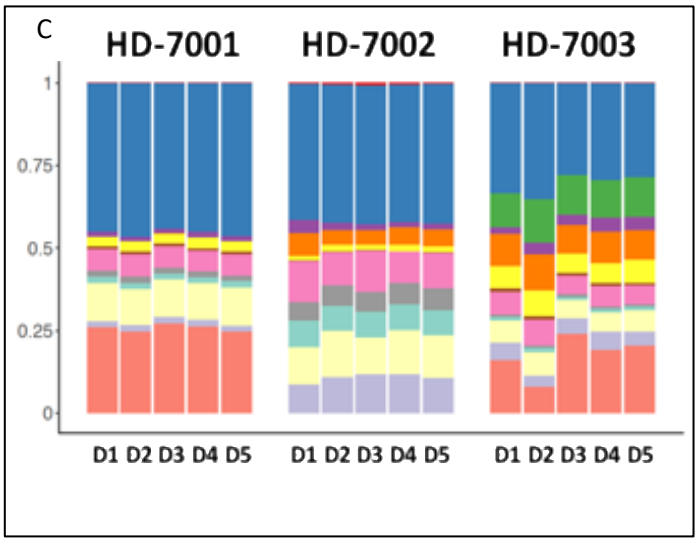
Figure 4. Stacked bar plot of relative abundance of bacterial taxa at the genus level. Dot plot of relative abundance of eight major taxa at the genus level.

Table 4. Microbiome composition and strain information

Species	Genomic DNA	16S only	Genome Copy	GC content (%)	Gram Stain
Listeria monocytogenes	12	14.1	13.9	38.0	+
Pseudomonas aeruginosa	12	4.2	6.1	66.2	-
Bacillus subtilis	12	17.4	10.3	43.9	+
Escherichia coli	12	10.1	8.5	46.7	-
Salmonella enterica	12	10.4	8.7	52.2	-
Lactobacillus fermentum	12	18.4	21.6	52.4	+
Enterococcus faecalis	12	9.9	14.6	37.5	+
Staphylococcus aureus	12	15.5	15.2	32.9	+
Saccharomyces cerevisiae	2	NA	0.57	38.3	Yeast
Cryptococcus neoformans	2	NA	0.37	48.3	Yeast







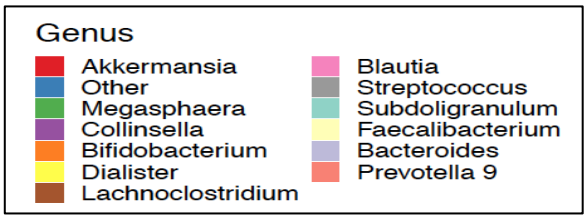


Figure 5. Beta-diversity, alpha-diversity, and bacterial composition analysis on healthy donor samples, HD-7001, HD-7002, HD-7003, collected from the Icahn School of Medicine at Mount Sinai. A. PCoA with weighted-UniFrac distances was used to analyze the beta-diversity. B. Alpha diversity was analyzed in dot plots, indicating the number of OTUs in the sample. The rarefied OTUs were calculated without removing any reads. C. The stacked bar plots of relative abundance of bacterial taxa at the genus level.

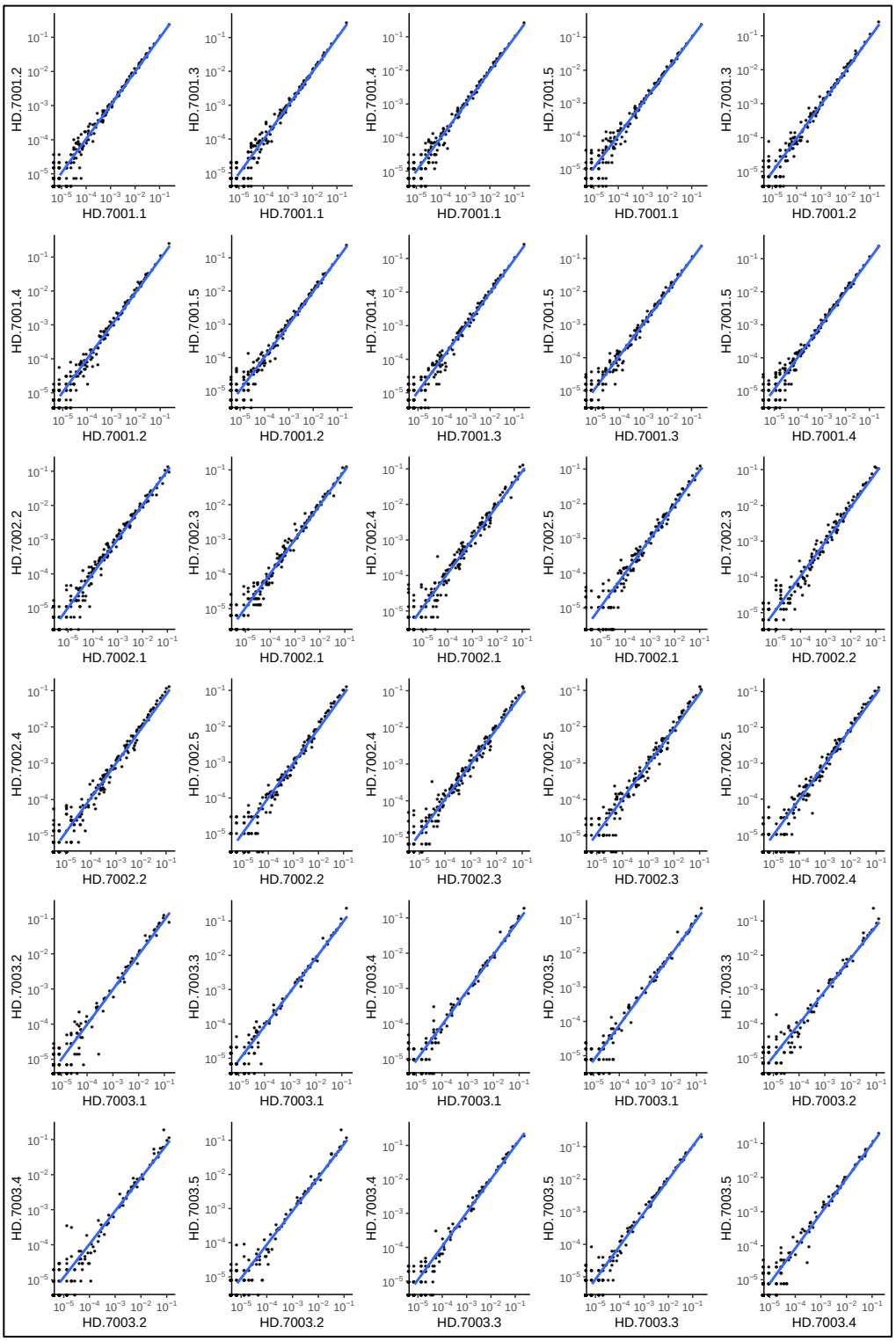


Figure 6. Pairwise analysis of patient samples at the genus level. Sample were collected from the Icahn School of Medicine at Mount Sinai.

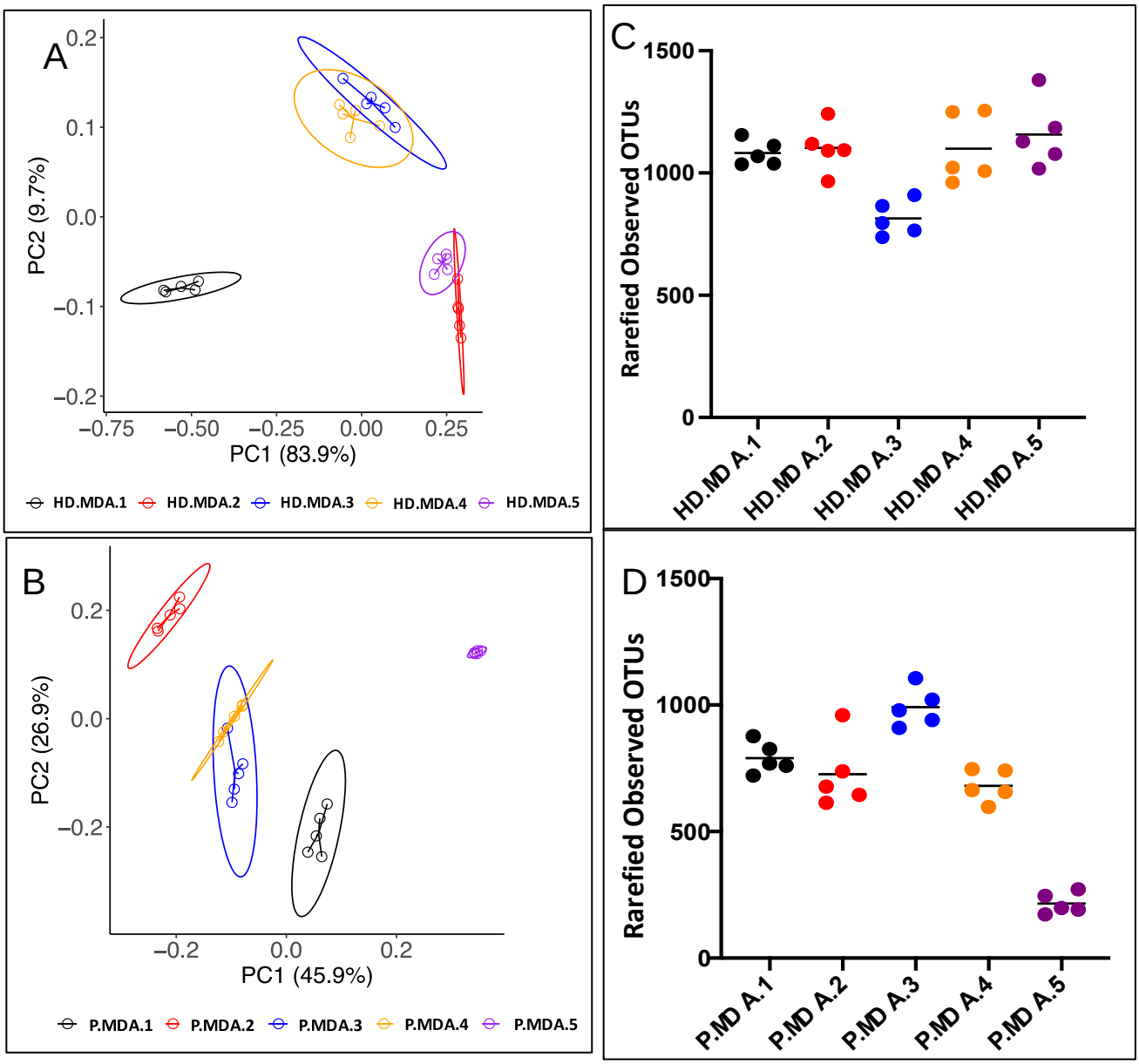


Figure 7. A-B Principal coordinates analysis derived from weighted UniFrac distances among five patient samples and five healthy donors. C-D Dot plots of the number of OTUs in healthy donors and patient samples, which were collected at MD Anderson Cancer Center. The rarefied OTUs were calculated without removing any reads.

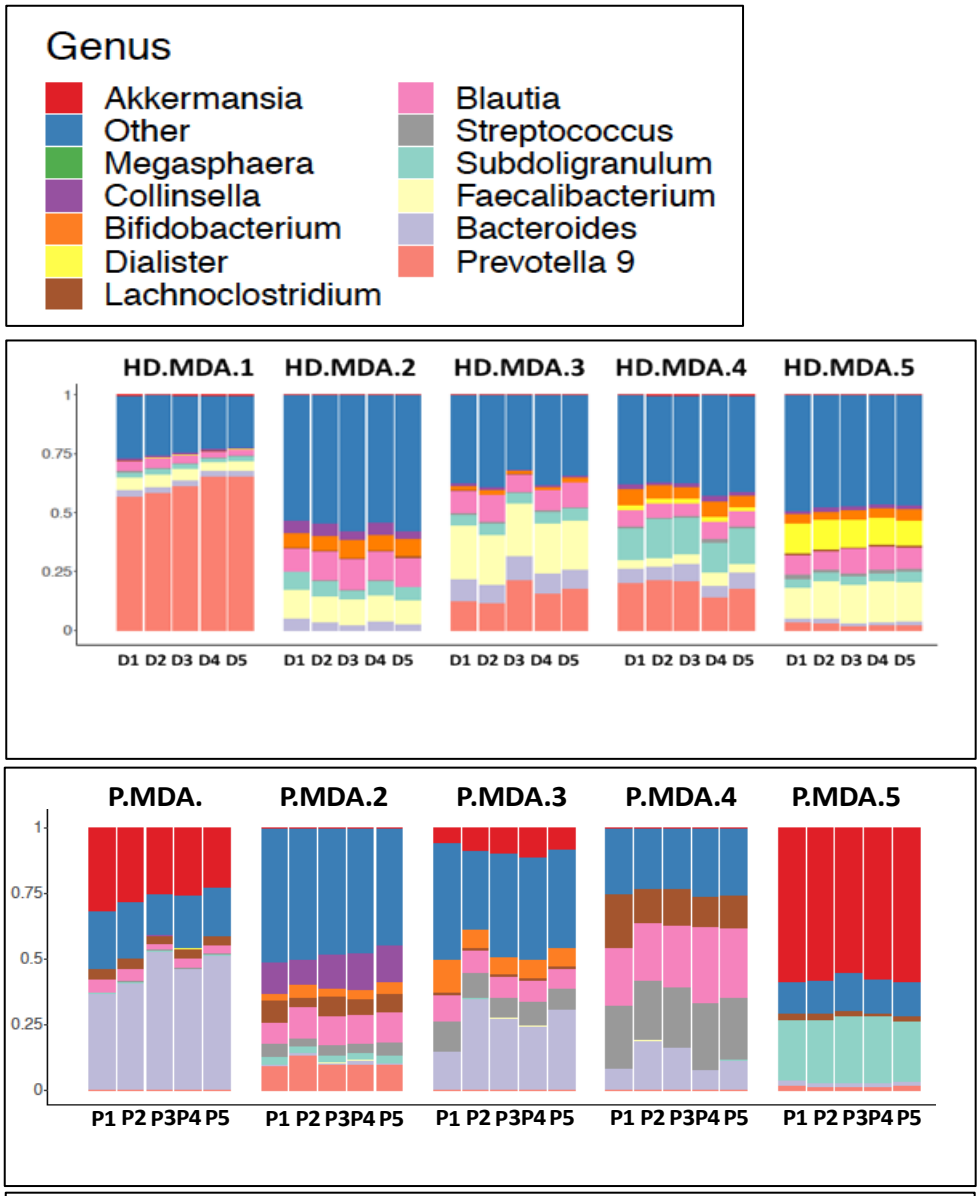


Figure 8. The stacked bar plots of relative abundance of common bacterial taxa at the genus level in healthy donors and patients. Samples were collected at MD Anderson Cancer Center.

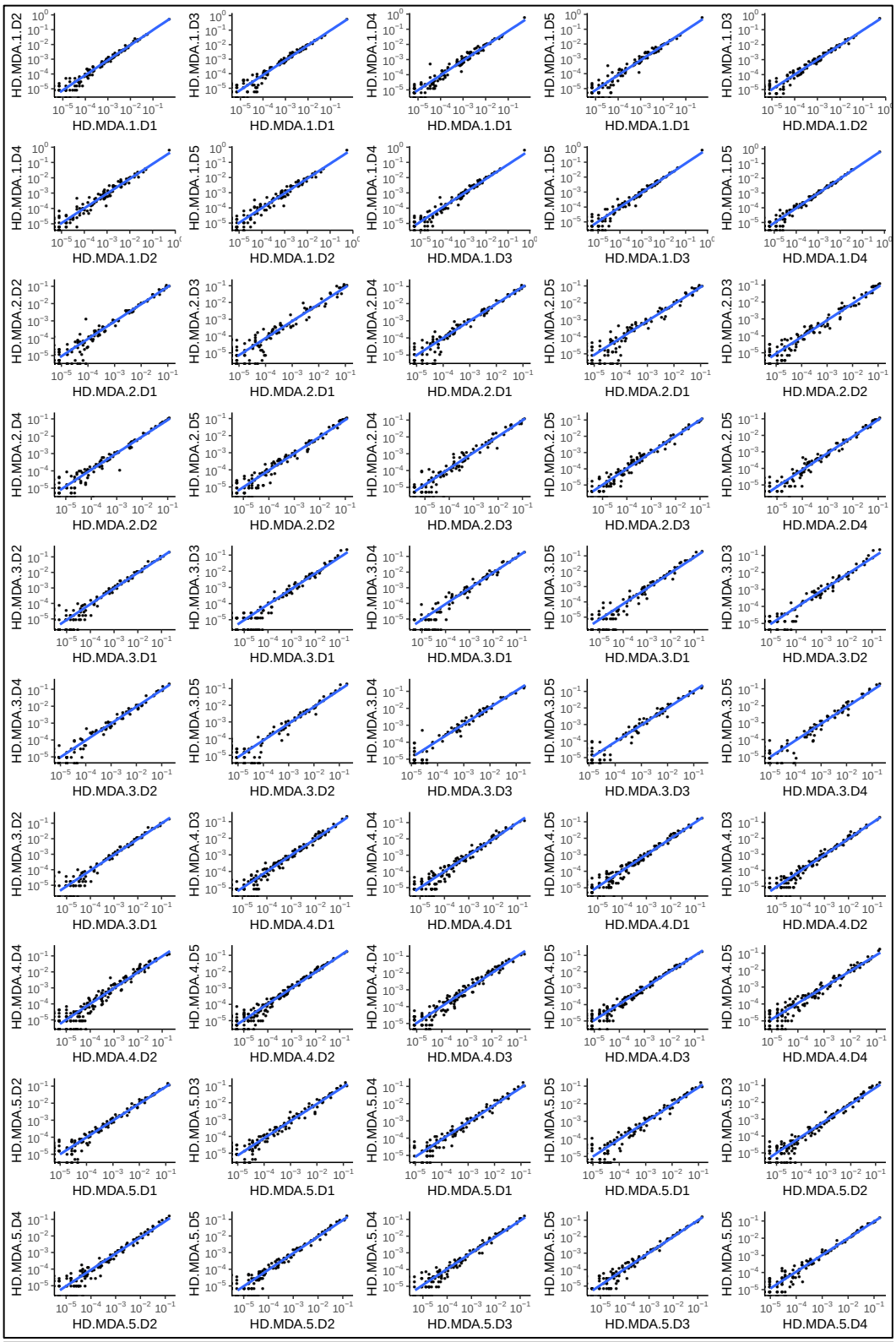


Figure 9. Pairwise analysis of healthy donor samples at the genus level. Samples were collected at MD Anderson Cancer Center.

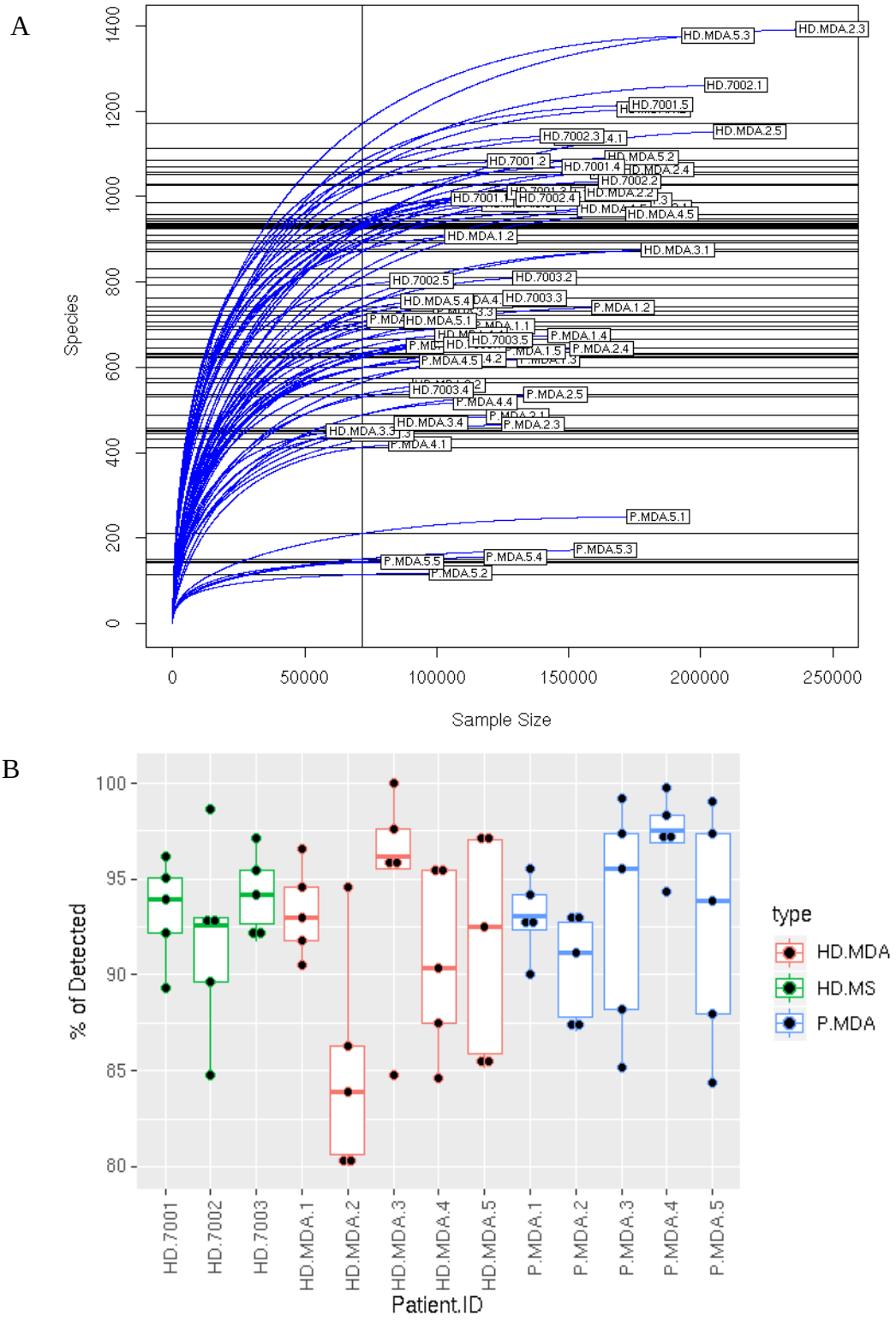


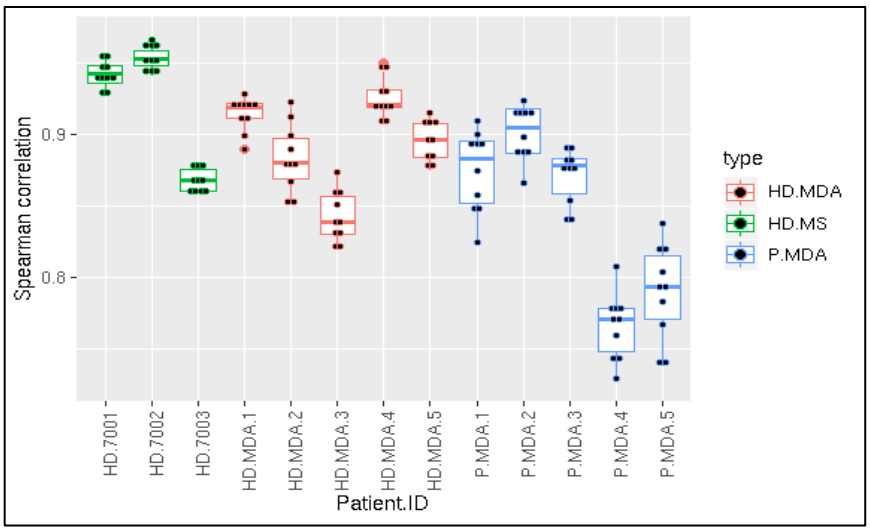
Figure 10. Rarefaction analysis of individual samples collected at MD Anderson Cancer Center and the Icahn School of Medicine at Mount Sinai. OTUs with only 1 read in a sample are removed for more robust estimation of species richness. A: rarefaction curve for each sample. B: percentage of detected calculated by dividing the rarefied number of species by the estimated asymptotic richness.

4. Statistical Analysis

4.1 Sample Reproducibility

The Spearman's correlation was calculated for genus data from each patient. The following boxplot shows the correlation coefficients (rho) grouped by patients and colored by institutions. It was observed that:

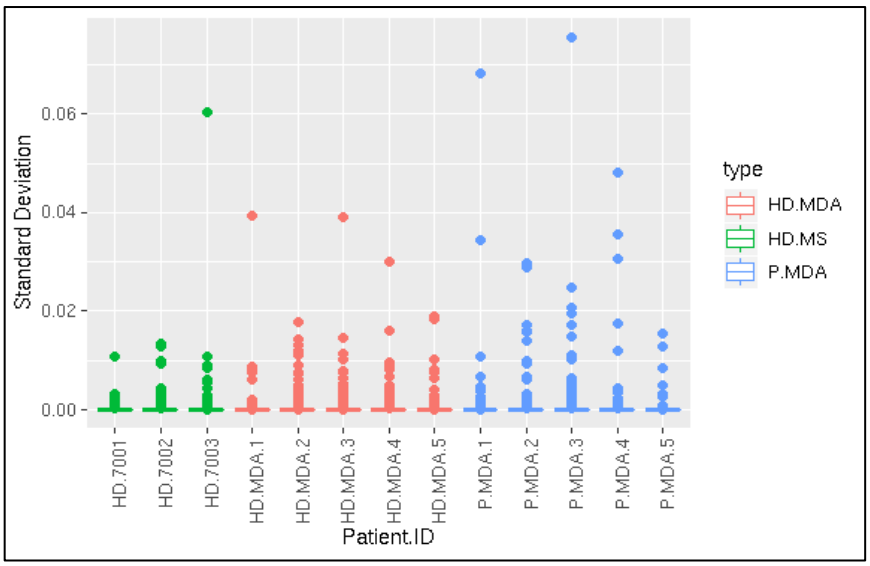
1. All healthy donors and three patients have rho > 0.8
2. Two patients have lower rho, but still > 0.7.



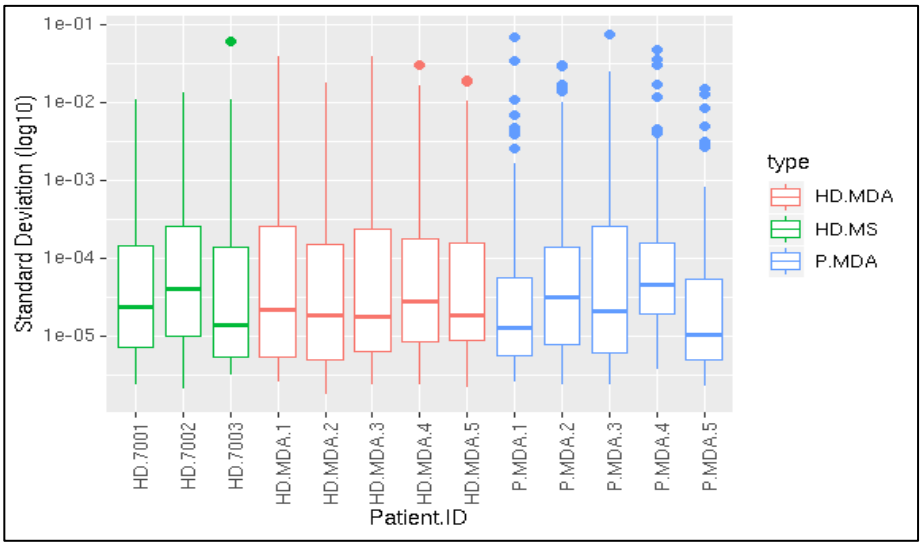
4.2 Sample Precision

Since the genus values are small, precision measurement with a coefficient of variation (CV) is too sensitive to determine small mean values. Thus, the standard deviation (SD) for each genus within each patient was calculated.

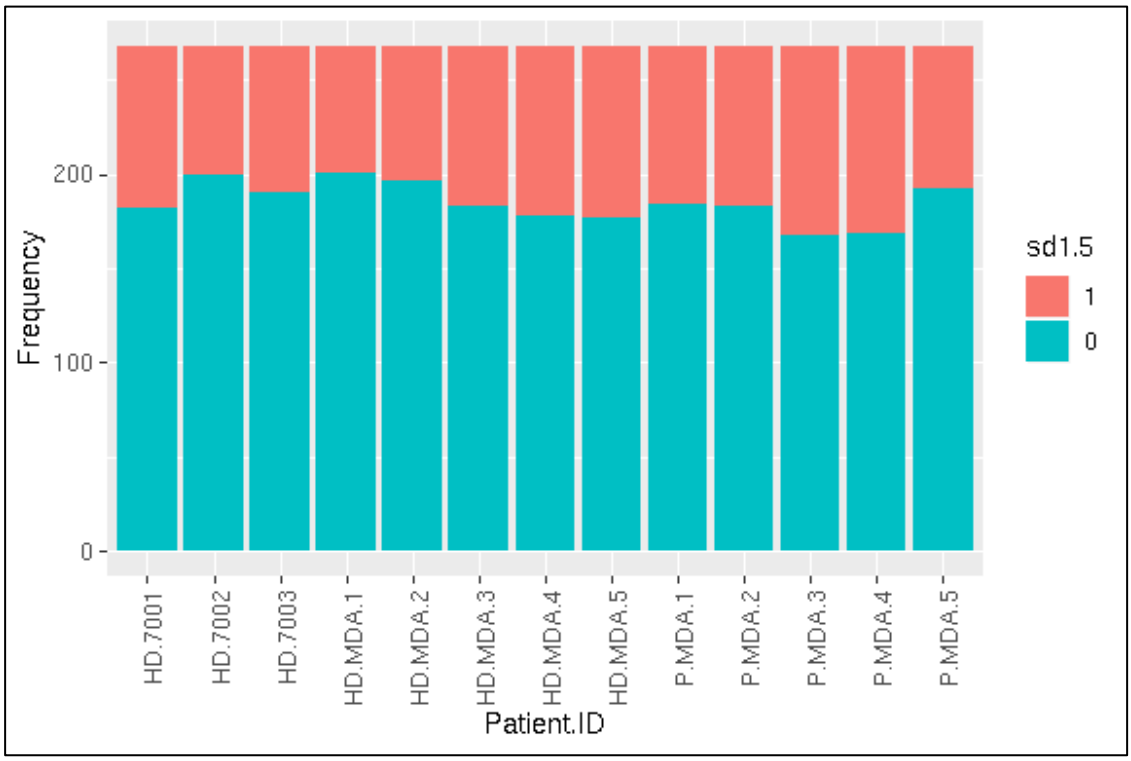
1. The following figure shows the calculated SD for each sample.



2. The standard deviation plotted in log10 transformed scale is shown below. All samples had similar SD distribution, but patient samples tend to have more outliers with higher SD.



3. There are no values outside of 2-SD range, which suggest good precision (95% of values fall within 2 standard deviations of the mean). The following figure shows number of genera with aliquots that are beyond 1.5-SD range in every sample. “0” means that the genus had all 5 aliquots within 1.5-SD range. “1” means that the genus had 1 aliquot beyond 1.5-SD range. Patient samples had slightly more aliquots beyond 1.5-SD range.



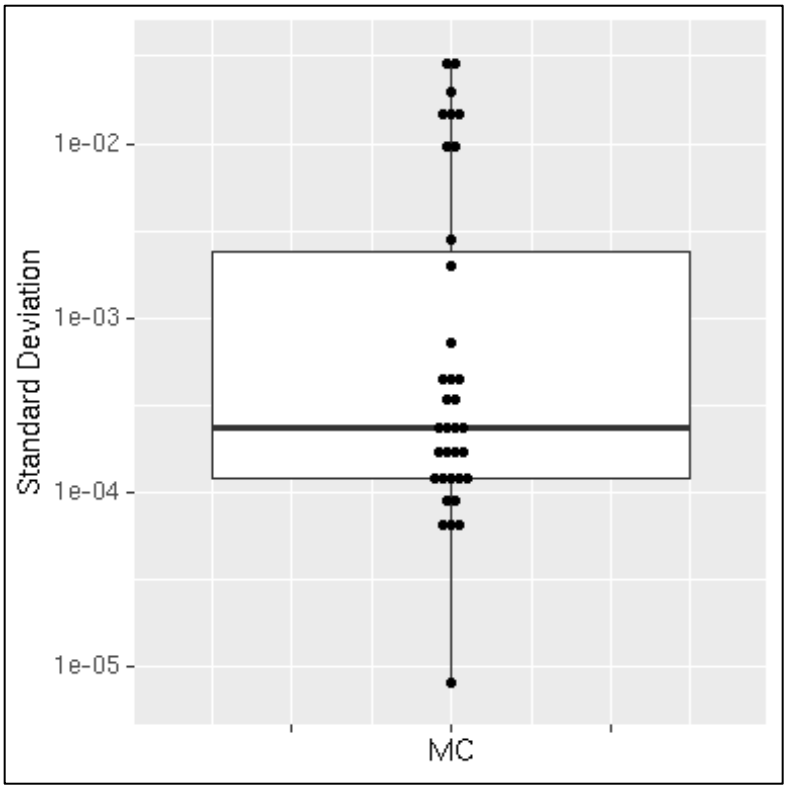
4.3 Mock Communities Reproducibility

Spearman's correlation coefficient for data from three mock communities shows all rho > 0.7.

mc1	mc2	rho
MC.P2019.38	MC.P2020.42	0.786
MC.P2019.38	MC.P2020.43	0.744
MC.P2020.42	MC.P2020.43	0.704

4.4 Mock Communities Precision

SD distribution for all genera calculated for three mock communities showed that SDs are small, and there is no value beyond the 1.5-SD range.



Based on correlation and standard deviation, good reproducibility and precision for all donor samples were observed. While patient samples have lower correlation and larger standard deviation compared to donor samples, they are within an acceptable range (rho > 0.7, no aliquot beyond 2-SD range). Mock communities have correlation > 0.7, and no aliquot is beyond the 1.5-SD range.

5. Reference

1. Watts GS, Youens-Clark K, Slepian MJ, et al. 16S rRNA gene sequencing on a benchtop sequencer: accuracy for identification of clinically important bacteria. *J Appl Microbiol*. 2017;123(6):1584-1596.
2. Caporaso JG, Lauber CL, Walters WA, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. 2012;6(8):1621-1624.
3. Barbas CF, 3rd, Burton DR, Scott JK, Silverman GJ. Quantitation of DNA and RNA. *CSH Protoc*. 2007;2007:pdb ip47.
4. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335-336.
5. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590-596.
6. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*. 2016;081257.
7. Nearing JT, Douglas GM, Comeau AM, Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*. 2018;6:e5364.
8. Schloss PD, Gevers D, Westcott SL. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*. 2011;6(12):e27310.