Icahn School of Medicine at Mount Sinai

Microbiome Analytical Validation

Version 1.1
Date: Nov 4, 2019

Microbiome Translational Center (MTC)
Facility Director: Gerold Bongers, PhD, Assistant Professor
Icahn School of Medicine at Mount Sinai
1470 Madison Avenue, New York, NY 10028

Table of Contents

Signature page

Microbiome Translation Center
Facility director: Gerold Bongers, PhD Assistant Professor, Oncological Sciences

Name	Signature	Date

List of tables

List of figures

Bioassay tables

1.1 Bioassay Description

Bioassay	Brief Assay Description	Application in this study (fit-for-purpose)
16S microbiome sequencing	16S rRNA amplicon sequencing is performed to determine the relative or absolute abundance of microbes in biological sample. Microbial DNA isolated by a combination of chemical and mechanical lysis and the V4 hypervariable region of the 16S rRNA gene is amplified by PCR. DUAL-index amplicons are normalized and combined into a library for sequencing on the Illumina MiSeq platform. Computational analysis is performed to determined relative abundances of taxa, alpha- and beta diversity and association taxa with phenotypes.	Application in this study: The microbiome is suggested to be involved in cancer initiation, progression and clinical response to immunotherapy.[1,2] Clinical response to immunotherapy is associated with a specific microbiome.[3–5] Although, detailed mechanistic understanding is currently lacking, characterization of the microbiome might help to stratify responders from non-responder, contribute to the development of microbiome-based combination therapy to improve response rate and/or reduce adverse events such as colitis.[1,6] Platform: MiSeq Analyte: Metagenome Analytical performance parameters Specificity: see Bioassay performance (v)s Sensitivity: 1 in 1,000 Reproducibility: CV <10%

1.2 Bioassay performance

16S microbiome sequencing	
Primary Assay Outputs	Operational Taxonomic Units (OTU) or Amplicon Sequence Variants (ASV).
Pre-processing/Normalization/QC	Fastq quality filtering and trimming as implemented in Qiime1[7], Qiime2[8] or DADA2[9]. Normalization by relative abundance, rarefaction or by the DESeq2[10] algorithm.
Initial Analyses	OTU or ASV count tables with taxonomic classification.
Derived Data Outputs	Alpha and beta diversity, differential abundances and phenotype-taxa associations.
(i) accuracy	Accuracy was determined by analysis of whole cell and DNA mock communities with known amounts of microbes. DNA extraction is known to enrich for DNA from gram-negative microbes over gram-positive microbes. Overall we found a media absolute error of 0.5x (Figure 4a), in line with results obtained by Costea et al.[11]
(ii) precision	Precision was determined by analysis of whole cell and DNA mock communities with known amounts of microbes across batches and fecal aliquots obtained from donors. Intra-batch CVs are generally <10% for the whole cell mock community and <5% for the DNA mock community. Biological samples focusing just on the highest abundant genera across CVs are generally <50%, but these are within 2 SD of the mean (Figure 4a, Figure 5c).
(iii) analytical sensitivity	Analytical sensitivity is dependent on many factors such as the characteristic of the microbe, diversity in the sample, 16S rRNA subregion amplified and sequencing depth. In a typical experiment sensitivity to quantify an AVS is 1 in 1,000 and to detect 1 in 10,000 (Figure 3).
(iv) Coefficient of variation and reproducibility	See precision.
(v) analytical specificity including interfering substances	Amplification and sequencing errors can decrease specificity, although to a certain extent this can be addressed computationally by quality filtering[12] and algorithms such as Deblur and DADA2.[9,13] Low biomass samples, such as skin, false positive results can be obtained due to amplification of environmental contaminant microbial DNA. Environmental contamination

16S microbiome sequencing	
	is monitored by inclusion of water, no-template PCR controls and is generally not an issue in fecal samples that have a high microbial biomass.
(vi) reference intervals (normal values) with controls and calibrators	Reference controls, whole cell Zymo mock communities and DNA mock communities, are included in each run to ensure batches are within 2xSD and that a genera with an abundance > 1 in 1000 are detected. The Zymo mock community is well-defined, accurately characterized and includes a range of organisms with different properties.
(vii) standardization, harmonization, and ruggedness of analytical performance if the assay is to be performed in multiple laboratories	In addition to the commercially available whole cell Zymo mock community, we have also collected and aliquoted fecal samples from three different donors that is used to evaluated batch variations and can be shared between laboratories.
(viii) establishment of appropriate quality control and improvement procedures	In addition to the overall quality control assessment using the mock communities and fecal aliquots described above, specific steps of the procedure are quality controlled as well. Specifically, extracted DNA is quantified by Quant-IT and normalized. Amplicons are quantified by Quant-IT and a subset is evaluated for yield and size by gel analysis. Sequence libraries are evaluated for size and purity by gel or bioanalyzer and quantified by Qubit and qPCR, the resulting fastq files are quality filtered and the inserted size and number of reads per samples are determined. Samples not passing QC at any step are processed again, if possible, or removed from the analysis altogether.
(ix) any other performance characteristics required for assay performance	Taxonomy in 16S rRNA amplicon sequencing is determined from a specific region of the 16S rRNA gene and often sequences are collapsed into a higher taxonomic rank as different strains share the same or highly similar amplicon sequence. In cases where species or strain level resolution is needed, use whole genome shotgun metagenomics. Metagenome functional content can be predicted from 16S marker data using PICRUSt[14], however if an accurate functional analysis is required genome shotgun metagenomics or transcriptomics should be used.
(x) scoring procedure and criteria for significant change	Group differences with be evaluated using DESeq2[10], LEfSe[15] and non-parametric rank tests (Kruskal-Wallis, Wilcoxon). P-values will be adjusted using Benjamini and Hochberg (FDR) and $p < 0.05$ will be considered significant.

2 Analytical Performance Data

2.1 Introduction

The microbiome is suggested to be involved in cancer initiation, progression and clinical response to immunotherapy.[1,2] Clinical response to immunotherapy is associated with a specific microbiome.[3–5] Although, detailed mechanistic understanding is currently lacking, characterization of the microbiome might help to stratify responders from non-responder, contribute to the development of microbiome-based combination therapy to improve response rate and/or reduce adverse events such as colitis.[1,6]

2.2 Analyte(s)

Metagenome as determined by sequencing of the V4 hypervariable region of the 16S rRNA gene

2.3 Technical platform(s)

Amplicons are sequenced on the Illumina MiSeq platform

2.4 Reagents, controls, and calibrators

2.4.1 DNA Extraction

The method described efficiently DNA from stool samples using a combination of mechanical and chemical lysis. The isolated DNA is suitable for NGS analysis such as 16S rDNA amplicon sequencing and shotgun metagenomics.

Table 1. DNA extraction reagents

Reagent	Vendor	Catalog
0.1mm Zirconia/Silica Beads	Fisher Scientific	11079101Z
Phenol:Chloroform:IAA, 25:24:1, pH 6.6	Life Technologies	AM9732
SDS	Sigma-Aldrich	75746-250G
QIAquick 96 PCR Purification Kit	Qiagen	28183
Axygen 2.0mL Self-Standing Screw Cap Tubes	Fisher Scientific	14-222-626
Quant-iT™ dsDNA Assay Kit, broad range	Life Technologies	Q33130
Corning™ 96-Well Solid Black Plates	Fisher Scientific	07-200-590
Hard-Shell® 96-Well PCR Plates, low profile, thin wall, skirted, white/clear	Bio-Rad	HSP9601
Eppendorf Combitips advanced	Fisher Scientific	13-683-724
ClipTip Filtered Pipette Tips	Fisher Scientific	14-387-978
Tris (1 M), pH 8.0	Life Technologies	AM9855G
EDTA (0.5 M), pH 8.0	Life Technologies	AM9260G
SYRINGE FILTER .2UM CS50	VWR	28145-477
Sterile Luer-Slip Syringes, Air-Tite	VWR	53548-006
96-Well Deep Well Plates	VWR	10755-250
Reagent	Vendor	Catalog

2.4.2 Preparation and sequencing of 16S rRNA gene sequence libraries

The Purpose of this protocol is to define the steps for the preparation and sequencing of 16S rRNA gene sequence libraries using the Illumina MiSeq sequencing platform. Method adapted from Kozich, J. J. et al.[16] see also [https://github.com/SchlossLab/MiSeq_WetLab_SOP/blob/master/MiSeq_WetLab_SOP_v4.md]. Libraries are quantified by qPCR using a previous library and a standard curve generated from PhiX with the following primers qPCR primer 1.1 AATGATACGGCGACCACCGAGAT and qPCR primer 2.1 CAAGCAGAAGACGGCATACGA.

Table 2. Reagent for library preparation

Reagent	Vendor	Catalog
Hard-Shell® 96-Well PCR Plates, low profile, thin wall, skirted, white/clear	Bio-Rad	HSP9601
PCR Plate Heat Seal, foil, peelable	Bio-Rad	1814045
UltraPure™ DNase/RNase-Free Distilled Water	Life Technologies	10977015
Phusion High-Fidelity PCR Master Mix	NEB	M0531L
Quant-iT™ dsDNA Assay Kit, broad range	Life Technologies	Q33130
Corning™ 96-Well Solid Black Plates	Fisher Scientific	07-200-590
Beckman Coulter AMPURE XP	Fisher Scientific	NC9959336
ZymoBIOMICS Microbial Community Standard	Zymo Research	D6300

20 Strain Even Mix Whole Cell Material	ATCC	MSA-2002
TE, pH 8.0	Life Technologies	AM9849
Buffer EB	Qiagen	19086
GEL LOAD DYE PURP (6X) 4 ML	Fisher Scientific	50591186
QUICK-L 2-LOG DNA LADD-125 LAN	Fisher Scientific	50994896
SYBR SAFE DNA GEL STAIN	Fisher Scientific	S33102
AGAROSE LOW EEO 100G	Fisher Scientific	BP160100
NaCl (5 M)	Life Technologies	AM9760G
Quarter Reservoir Sterile Divided by Length	Beckman Coulter	372788

V4 amplicon primers

General amplicon primer design is as follows; forward: AATGATACGGCGACCACCGAGATCTACAC <i5><pad><link><16Sf>, reverse: CAAGCAGAAGACGGCATACGAGAT <i7><pad><link><16Sr>.

The listed sequences in the generic design, above, are the adapter sequences to allow annealing of the amplicons to the flow cell. The i5 and i7 sequences are the 8-nt index sequences. The pad is a 10-nt sequence to boost the sequencing primer melting temperatures. The link is a 2-nt sequence that is anti-complementary to the known sequences. The 16Sf and 16Sr are the gene specific primer sequences. Primers are purchased from IDT with no special purification.

Table 3. Forward primers

Name	Adapter	i5	pad	link	16Sf
SA501	AATGATACGGCGACCACCGAGATCTACAC	ATCGTACG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA502	AATGATACGGCGACCACCGAGATCTACAC	ACTATCTG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA503	AATGATACGGCGACCACCGAGATCTACAC	TAGCGAGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA504	AATGATACGGCGACCACCGAGATCTACAC	CTGCGTGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA505	AATGATACGGCGACCACCGAGATCTACAC	TCATCGAG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA506	AATGATACGGCGACCACCGAGATCTACAC	CGTGAGTG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA507	AATGATACGGCGACCACCGAGATCTACAC	GGATATCT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SA508	AATGATACGGCGACCACCGAGATCTACAC	GACACCGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB501	AATGATACGGCGACCACCGAGATCTACAC	CTACTATA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB502	AATGATACGGCGACCACCGAGATCTACAC	CGTTACTA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB503	AATGATACGGCGACCACCGAGATCTACAC	AGAGTCAC	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB504	AATGATACGGCGACCACCGAGATCTACAC	TACGAGAC	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB505	AATGATACGGCGACCACCGAGATCTACAC	ACGTCTCG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB506	AATGATACGGCGACCACCGAGATCTACAC	TCGACGAG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB507	AATGATACGGCGACCACCGAGATCTACAC	GATCGTGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SB508	AATGATACGGCGACCACCGAGATCTACAC	GTCAGATA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC501	AATGATACGGCGACCACCGAGATCTACAC	ACGACGTG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC502	AATGATACGGCGACCACCGAGATCTACAC	ATATACAC	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC503	AATGATACGGCGACCACCGAGATCTACAC	CGTCGCTA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC504	AATGATACGGCGACCACCGAGATCTACAC	CTAGAGCT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC505	AATGATACGGCGACCACCGAGATCTACAC	GCTCTAGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC506	AATGATACGGCGACCACCGAGATCTACAC	GACACTGA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC507	AATGATACGGCGACCACCGAGATCTACAC	TGCGTACG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SC508	AATGATACGGCGACCACCGAGATCTACAC	TAGTGTAG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD501	AATGATACGGCGACCACCGAGATCTACAC	AAGCAGCA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD502	AATGATACGGCGACCACCGAGATCTACAC	ACGCGTGA	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD503	AATGATACGGCGACCACCGAGATCTACAC	CGATCTAC	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD504	AATGATACGGCGACCACCGAGATCTACAC	TGCGTCAC	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD505	AATGATACGGCGACCACCGAGATCTACAC	GTCTAGTG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD506	AATGATACGGCGACCACCGAGATCTACAC	CTAGTATG	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD507	AATGATACGGCGACCACCGAGATCTACAC	GATAGCGT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
SD508	AATGATACGGCGACCACCGAGATCTACAC	TCTACACT	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA

Table 4. Reverse primers

Name	Adapter	i7	pad	link	16Sf
SA701	CAAGCAGAAGACGGCATACGAGAT	AACTCTCG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA702	CAAGCAGAAGACGGCATACGAGAT	ACTATGTC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA703	CAAGCAGAAGACGGCATACGAGAT	AGTAGCGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA704	CAAGCAGAAGACGGCATACGAGAT	CAGTGAGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA705	CAAGCAGAAGACGGCATACGAGAT	CGTACTCA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA706	CAAGCAGAAGACGGCATACGAGAT	CTACGCAG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT

SA707	CAAGCAGAAGACGGCATACGAGAT	GGAGACTA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA708	CAAGCAGAAGACGGCATACGAGAT	GTCGCTCG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA709	CAAGCAGAAGACGGCATACGAGAT	GTCGTAGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA710	CAAGCAGAAGACGGCATACGAGAT	TAGCAGAC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA711	CAAGCAGAAGACGGCATACGAGAT	TCATAGAC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SA712	CAAGCAGAAGACGGCATACGAGAT	TCGCTATA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB701	CAAGCAGAAGACGGCATACGAGAT	AAGTCGAG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB702	CAAGCAGAAGACGGCATACGAGAT	ATACTTCG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB703	CAAGCAGAAGACGGCATACGAGAT	AGCTGCTA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB704	CAAGCAGAAGACGGCATACGAGAT	CATAGAGA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB705	CAAGCAGAAGACGGCATACGAGAT	CGTAGATC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB706	CAAGCAGAAGACGGCATACGAGAT	CTCGTTAC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB707	CAAGCAGAAGACGGCATACGAGAT	GCGCACGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB708	CAAGCAGAAGACGGCATACGAGAT	GGTACTAT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB709	CAAGCAGAAGACGGCATACGAGAT	GTATACGC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB710	CAAGCAGAAGACGGCATACGAGAT	TACGAGCA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB711	CAAGCAGAAGACGGCATACGAGAT	TCAGCGTT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SB712	CAAGCAGAAGACGGCATACGAGAT	TCGCTACG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC701	CAAGCAGAAGACGGCATACGAGAT	ACCTACTG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC702	CAAGCAGAAGACGGCATACGAGAT	AGCGCTAT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC703	CAAGCAGAAGACGGCATACGAGAT	AGTCTAGA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC704	CAAGCAGAAGACGGCATACGAGAT	CATGAGGA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC705	CAAGCAGAAGACGGCATACGAGAT	CTAGCTCG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC706	CAAGCAGAAGACGGCATACGAGAT	CTCTAGAG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC707	CAAGCAGAAGACGGCATACGAGAT	GAGCTCAT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC708	CAAGCAGAAGACGGCATACGAGAT	GGTATGCT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC709	CAAGCAGAAGACGGCATACGAGAT	GTATGACG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC710	CAAGCAGAAGACGGCATACGAGAT	TAGACTGA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC711	CAAGCAGAAGACGGCATACGAGAT	TCACGATG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SC712	CAAGCAGAAGACGGCATACGAGAT	TCGAGCTC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD701	CAAGCAGAAGACGGCATACGAGAT	ACCTAGTA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD702	CAAGCAGAAGACGGCATACGAGAT	ACGTACGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD703	CAAGCAGAAGACGGCATACGAGAT	ATATCGCG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD704	CAAGCAGAAGACGGCATACGAGAT	CACGATAG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD705	CAAGCAGAAGACGGCATACGAGAT	CGTATCGC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD706	CAAGCAGAAGACGGCATACGAGAT	CTGCGACT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD707	CAAGCAGAAGACGGCATACGAGAT	GCTGTAAC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD708	CAAGCAGAAGACGGCATACGAGAT	GGACGTTA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD709	CAAGCAGAAGACGGCATACGAGAT	GGTCGTAG	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD710	CAAGCAGAAGACGGCATACGAGAT	TAAGTCTC	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD711	CAAGCAGAAGACGGCATACGAGAT	TACACAGT	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
SD712	CAAGCAGAAGACGGCATACGAGAT	TTGACGCA	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT

Table 5. Read primers

Name	pad	link	16Sf
Read1	TATGGTAATT	GT	GTGCCAGCMGCCGCGGTAA
Read2	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT
ReadI	AGTCAGTCAG	CC	GGACTACHVGGGTWTCTAAT

Table 6. Sequencing reagents

Reagent	Vendor	Catalog
PhiX CONTROL V3 KIT	Illumina	FC-110-3001
MiSeq Reagent Kit v2 (300 cycle)	Illumina	MS-102-2002
Tween® 20, Molecular Biology Grade	Promega	H5152
KAPA SYBR® FAST qPCR Kits	Kappa	KK4603
Qubit™ 1X dsDNA HS Assay Kit	Thermo Fisher	Q32851
MicroAmp™ Fast Optical 96-Well Reaction Plate	Thermo Fisher	4346907
MicroAmp™ Optical Adhesive Film	Thermo Fisher	4360954
Sodium hydroxide solution	Sigma-Aldrich	72068-100ML
PhiX CONTROL V3 KIT	Illumina	FC-110-3001
MiSeq Reagent Kit v2 (300 cycle)	Illumina	MS-102-2002
Tween® 20, Molecular Biology Grade	Promega	H5152
KAPA SYBR® FAST qPCR Kits	Kappa	KK4603

Qubit™ 1X dsDNA HS Assay Kit	Thermo Fisher	Q32851
MicroAmp™ Fast Optical 96-Well Reaction Plate	Thermo Fisher	4346907
MicroAmp™ Optical Adhesive Film	Thermo Fisher	4360954
Sodium hydroxide solution	Sigma-Aldrich	72068-100ML

2.5 Quality control parameters for specimens

2.5.1 Stool sample

Upon receipt, the sample barcodes is scanned and stool samples are inspected to ensure that the package is intact and the sample/ice packs feel cold to the touch before storage at -80°C; deviations are recorded. Stool samples are aliquoted (3x50-100mg, 2x500mg and 8g) under liquid nitrogen using a custom mortar and pestle for efficient storage and downstream processing. Therefore, aliquot barcodes are scanned and tubes are pre- and post-weighed using a analytical scale that directly writes to a computer. For efficient extraction only aliquots between 50 and 100mg are used for metagenomics that includes absolute quantification.

2.5.2 Extracted DNA

To enable absolute quantification it is important that all samples are extracted using the volume during the phenol extraction step. Any deviation of the volume described in the SOP will be recorded. Extracted DNA is quantified using the Quant-iT dsDNA Assay kit and normalized to 2 ng/ul using the Biomek liquid handler. Samples below 2 ng/ul are, if possible, re-extracted from another aliquot; if the aqueous phase during the phenol extraction was <200 ul due to voluminous fecal sample an aliquot with less fecal material otherwise are aliquot with more fecal material will be used for the re-extraction. Any samples that remain below 2 ng/ul are annotated and recorded as such and continue on to the amplification step (samples with less DNA usually still amplify without any issues). Extraction water controls should measure ~0 ng/ul.

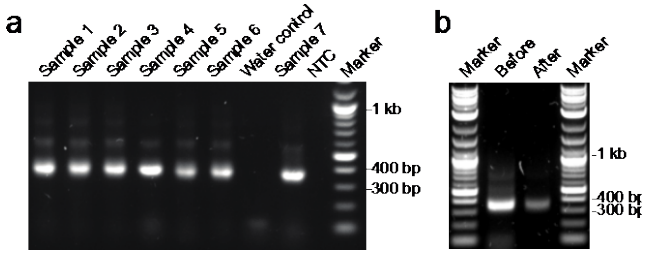

Figure 1. DNA and amplicon QC.
a) Example gel showing amplicons with the expected 352 bp size and no amplification with the water sample control or the no-template PCR control (NTC). b) Gel showing library before and after AMPure cleanup.

2.5.3 Amplicons

Amplicons are quantified by for yield by Quant-IT and size by gel inspection. The amplicon size should be approximately 352 bp and the DNA extraction water controls and the no template PCR controls should not show any amplification (Figure 1a). In case these controls show a band, the PCR should be repeated with new reagents.

2.5.4 Sequence library

Sequence library will be purified using double size selection with AMPure beads and validated on gel or bioanalyzer for the absence of primer dimers or higher molecular weight bands (Figure 1b). The purified library will be quantified by Qubit-HS and diluted to 2 nM. Concentration of the diluted library will be validated by qPCR for which the standard curve should have a $R^2 > 0.9$. The qPCR quantified library should be 2 ± 0.2 nM.

2.6 Any critical pre-analytic variables

Fastq files are quality filtered as implement in Qiime or DADA2 using default values when appropriate. Insert size of the stitched reads should be ~252 bp. Total reads for each sample should be >10K and water kit controls and no template PCR controls (NTC) should all be <2K reads (Figure 2).

2.7 Analytical performance characteristics

2.7.1 Validation of procedure

Sensitivity

Analytical sensitivity is dependent on many factors such as the characteristic of the microbe, diversity in the sample, 16S rRNA subregion amplified and sequencing depth. In a typical experiment at 12M paired-end reads pooling 384 samples a sequence depth ~30K counts per sample equates to a sensitivity to quantify an AVS is 1 in 1,000 and to detect 1 in 10,000 (Figure 3). Increasing the number of samples per run leads to less counts per sample and a reduced sensitivity to quantify and detect an ASV and it is advised not to pool more than 384 samples on a sequencing run.

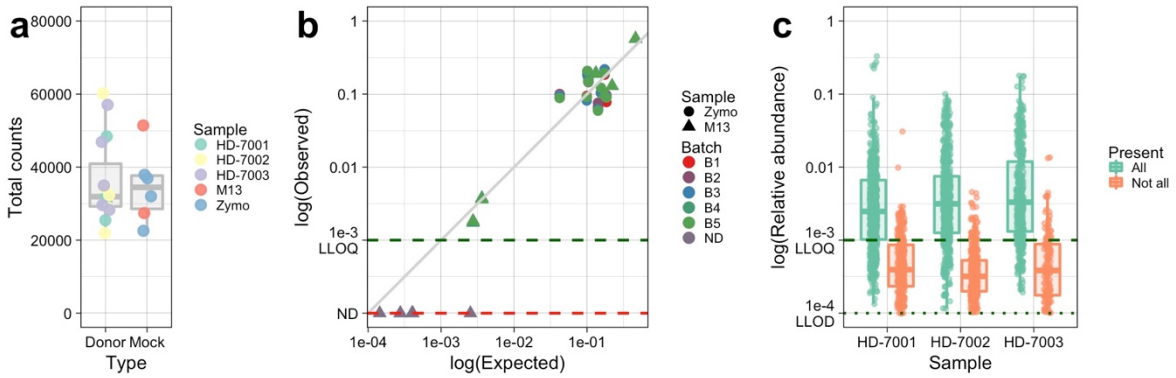

Figure 2. Sensitivity.
a) Read depth to determine sensitivity on the Mock (Zymo, M13) and Donor samples (HD-7001-5). b) Sensitivity in the Mock communities with LLOQ reflecting a reliability cutoff. c) Repeat analysis of the healthy donor samples showing ASVs present in all batches (B1-5) or absent in >1 batch (Not all). LLOQ line reflects the first quantile of ASVs detected in all batches. LLOD reflects detection limit of an ASV.

Specificity

Amplification and sequencing errors can decrease specificity, although to a certain extent this can be addressed computationally by quality filtering[12] and algorithms such as Deblur and DADA2.[9,13] Low biomass samples, such as skin, false positive results can be obtained due to amplification of environmental contaminant microbial DNA. Environmental contamination is monitored by inclusion of water, no-template PCR controls and is generally not an issue in fecal samples that have a high microbial biomass.

Accuracy

Accuracy was determined by analysis of whole cell and DNA mock communities with known amounts of microbes. DNA extraction is known to enrich for DNA from gram-negative microbes over gram-positive microbes**.** Overall we found a median absolute error of 0.5x (Figure 4a), in line with results obtained by Costea et al.[11]

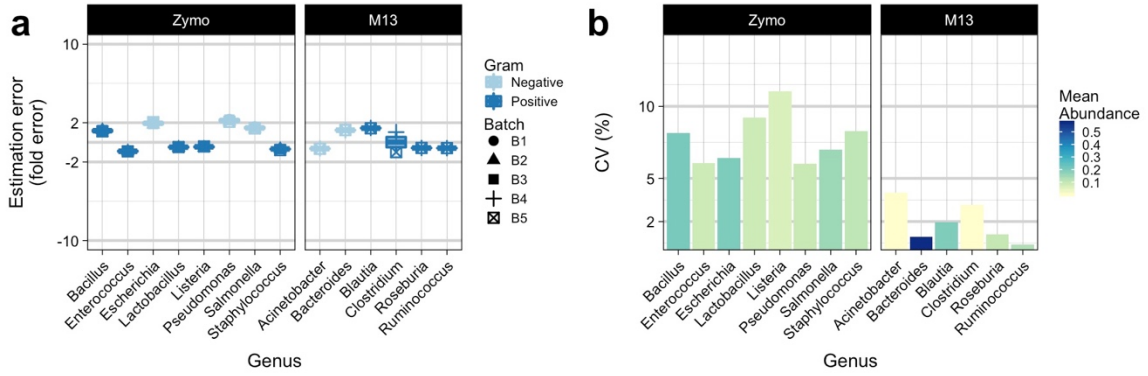

Figure 3. Accuracy and precision of mock communities.
a) Boxplot overlaid with a dot plot for the estimation error for each of the detected genera in the Zymo and M13 mock communities. Batches are denoted by the different shapes and colors indicate gram positive or negative microbes. b) Coefficient of variation (CV) across the 5 batches for each for the detected genera in the two mock communities. Fill color denotes the mean abundance of the microbes in the mock community.

Precision (run-to-run variation) and reproducibility

Precision was determined by analysis of whole cell and DNA mock communities with known amounts of microbes across batches and fecal aliquots obtained from donors. Intra-batch CVs are generally <10% for the whole cell mock community and <5% for the DNA mock community. Biological samples focusing just on the highest abundant genera across CVs are generally <50%, but these are within 2 SD of the mean (Figure 4a, Figure 5c).

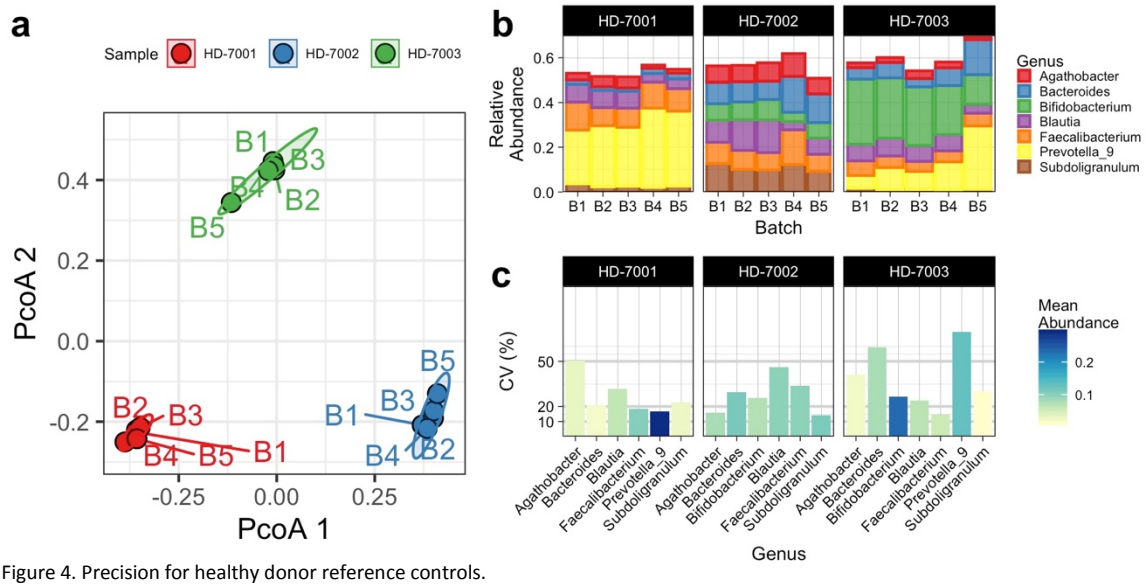

Figure 4. Precision for healthy donor reference controls.
a) Principal coordinate analysis (PCoA) plot with Bray-Curtis dissimilarity for the 3 healthy donor samples (HD-7001, HD-7002 and HD-7003) across the 5 batches (B1 – B5); ellipse indicates the 95% confidence interval. b) Barplots showing the top genera in each of the healthy donor samples for each of the batches. c) Coefficient of variation (CV) for the top genera in the healthy donor samples. Fill color denotes the mean abundance of the microbes.

Reportable range

Reportable range is dependent on the sequence depth, but reliable results can be expected at relative abundance of $>10^{-4}$ (for high n studies) with higher confidence at $>10^{-3}$ (for low n studies) in at least one of the experimental groups at a sequence depth of ~30K.

Reference ranges/intervals

Reference ranges are provided by Zymo research (zymoresearch.com) and the Clemente lab.

Table 7. Zymo mock community

Species	Genomic DNA	16S	Genome Copy	Cell Number
Bacillus subtilis	12	17.4	10.3	10.2
Enterococcus faecalis	12	9.9	14.6	14.5
Escherichia coli	12	10.1	8.5	8.5
Lactobacillus fermentum	12	18.4	21.6	21.4
Listeria monocytogenes	12	14.1	13.9	13.8
Pseudomonas aeruginosa	12	4.2	6.1	6.1
Salmonella enterica	12	10.4	8.7	8.7
Staphylococcus aureus	12	15.5	15.2	15.1

Table 8. M13 mock community

Genus	Relative abundance
Bacteroides	0.46000
Roseburia	0.17607
Blautia	0.13206
Ruminococcus	0.22000
Clostridium	0.00352

Acinetobacter	0.00264
Staphylococcus	0.00220
Escherichia	0.00031
Streptococcus	0.00030
Staphylococcus	0.00022
Enterococcus	0.00018
Pseudomonas	0.00018
Lactobacillus	0.00004

Turn-around time

Average turnaround time for 384 samples would be 2 – 4 weeks.

Failure rate of the assay

DNA extraction is dependent on the quality of the fecal samples, specifically watery stool samples are problematic and in about 0.5% of the samples PCR amplification fails, though this is highly dependent on the characteristics of the patient cohort. Failure rate for other steps is marginal.

2.7.2 Assay controls

Positive controls

Three characterized donor samples that are pre-aliquoted.

Negative controls

One blank water control that is taken through the complete protocol including extraction, PCR and sequencing for each set of 24 samples and a no template PCR water control for each set of 96 samples.

Reference standards

One whole cell mock community available from Zymo Research and one custom DNA mock community for each set of 96 samples.

2.7.3 Inter-laboratory variability

Not specifically determined, but this is thoroughly reported by Sinha et al.[17]. In this study blinded specimens (human stool, chemostats, mock communities) were sequenced by 15 laboratories and analyzed using 9 bioinformatics protocols. Biospecimen characteristics (type and origin) is the main source of variability. Technical variability is mainly caused by differences in DNA extraction methods followed by the sample handling environment and lastly the computational analysis pipelines.

2.7.4 Scoring procedures

Data types

Primary data are fastq files that containing the amplicon sequence data from which a sequence count table is generated, and taxonomy is assigned for each ASV or OTU. Quantitative count tables are normalized by calculation of relative abundance, rarefaction or DESeq's median or ratios[18]. Continuous metrics such as alpha diversity and beta diversity are calculated from the counts tables using R/phyloseq[19] or Qiime2[8].

Criteria and metrics for defining significant changes

Group differences will be evaluated using DESeq2[10], LEfSe[15] and non-parametric rank tests (Kruskal-Wallis, Wilcoxon). P-values will be adjusted using Benjamini and Hochberg (FDR) and $p < 0.05$ will be considered significant.

3 References

1. Zitvogel, L., Ma, Y., Raoult, D., Kroemer, G. & Gajewski, T. F. The microbiome in cancer immunotherapy: Diagnostic tools and therapeutic strategies. *Science* **359**, 1366–1370 (2018).
2. Tilg, H., Adolph, T. E., Gerner, R. R. & Moschen, A. R. The Intestinal Microbiota in Colorectal Cancer. *Cancer Cell* **33**, 954–964 (2018).
3. Routy, B. *et al.* Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
4. Gopalakrishnan, V. *et al.* Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
5. Matson, V. *et al.* The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients. *Science (80-.).* **359**, 104–108 (2018).
6. Wang, Y. *et al.* Fecal microbiota transplantation for refractory immune checkpoint inhibitor-associated colitis. *Nat. Med.* **24**, 1 (2018).
7. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
8. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
9. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
10. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
11. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069 (2017).
12. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
13. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, e00191-16 (2017).
14. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).
15. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
16. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–20 (2013).
17. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* **35**, (2017).
18. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, (2017).
19. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).