iScience, Volume *24*

Supplemental Information

Impaired host response and the presence of *Acinetobacter baumannii* in the serum microbiome of type-II diabetic patients

Dasith Perera, Sarah E. Kleinstein, Benjamin Hanson, Hatice Hasturk, Ryan Eveloff, Marcelo Freire, and Matthew Ramsey

Supplemental Information

Supplemental Figures and Legends

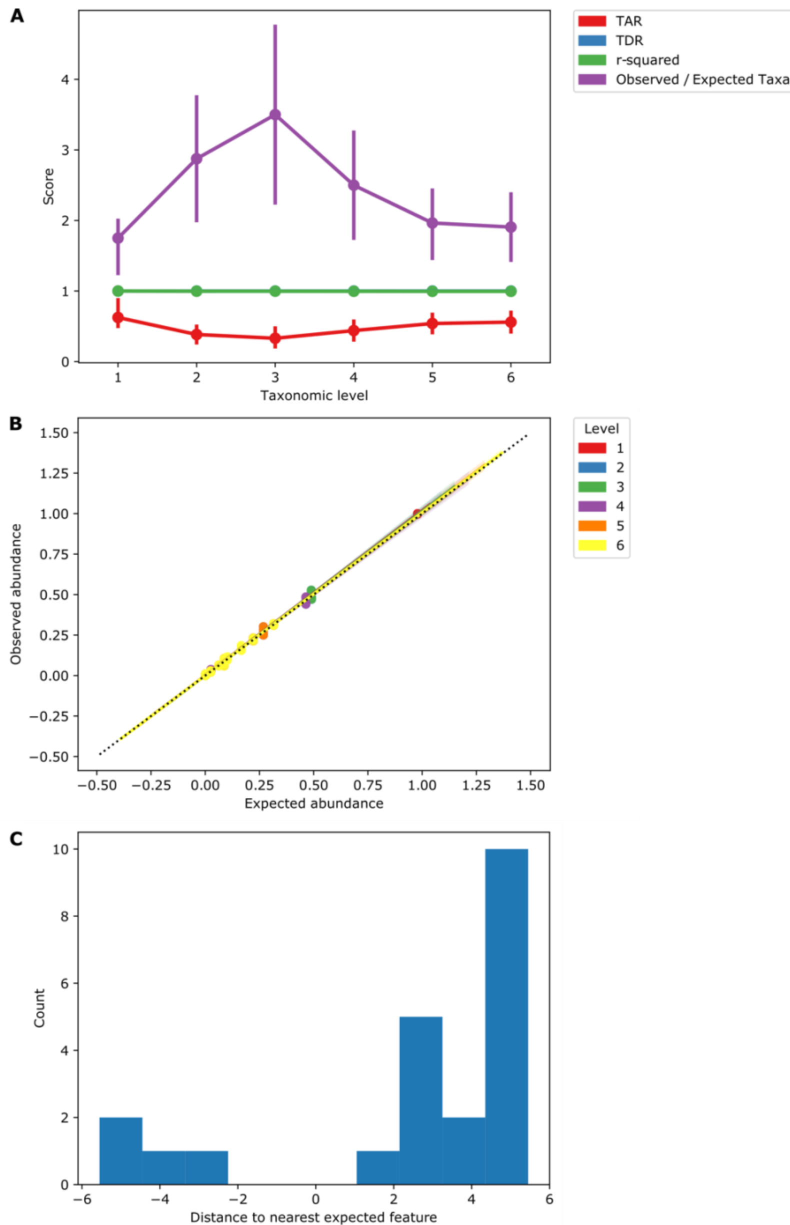
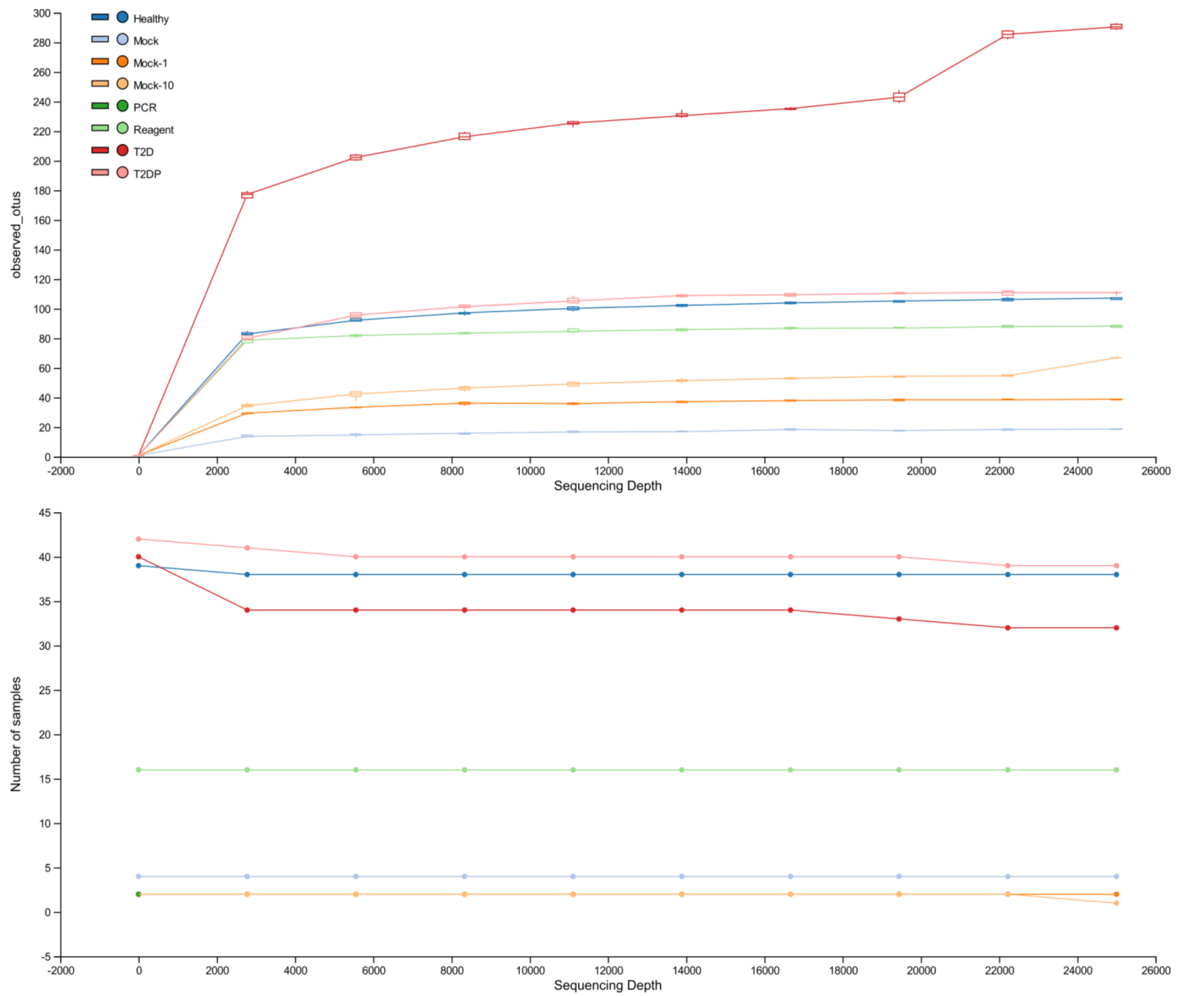
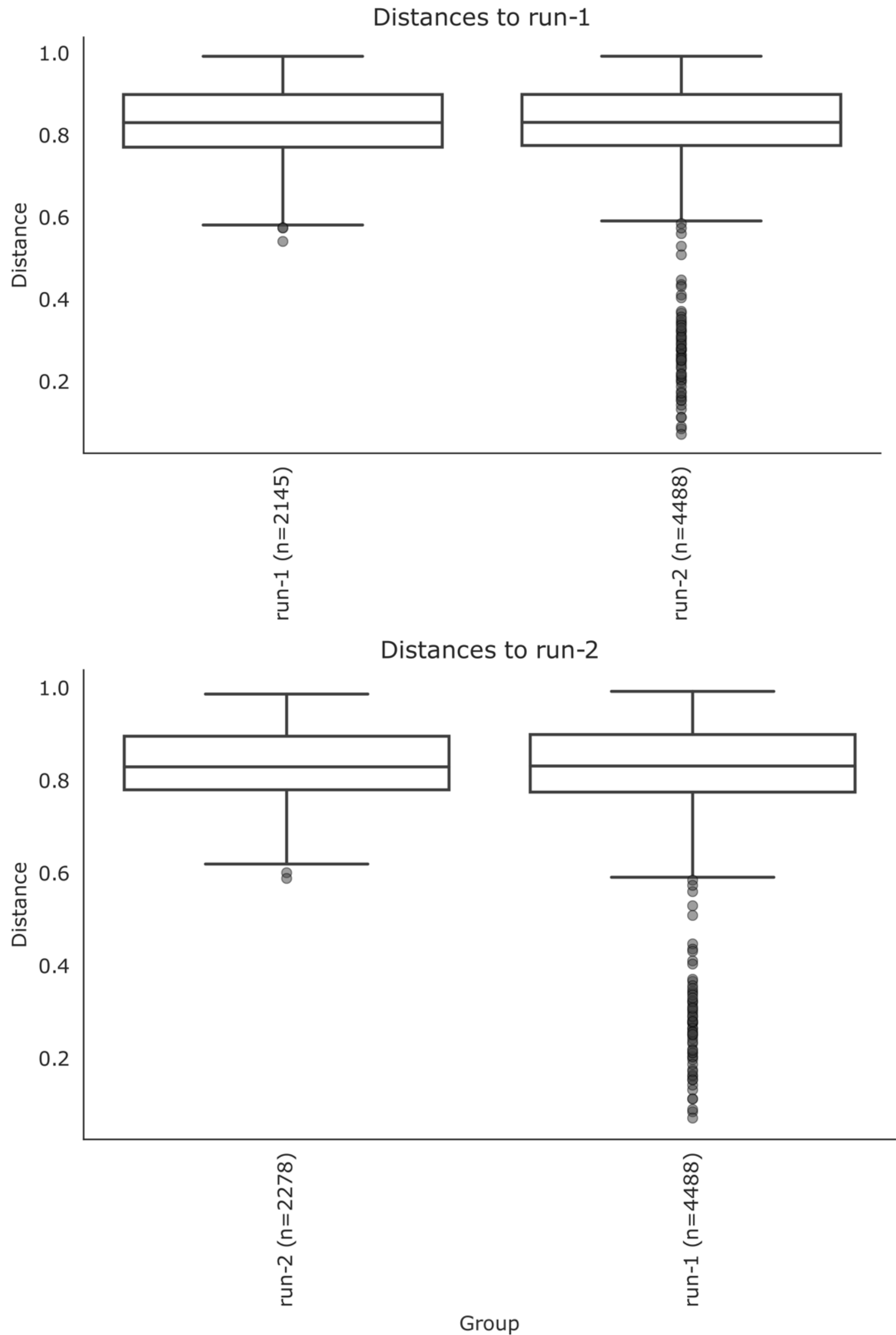


Figure S1. Mock library composition evaluation. Related to Figure 1. A Zymogen mock community library was used for sequence library construction at 10 and 1 ng in duplicate between both sequencing runs. QIIME2 / DADA2 unfiltered data was given taxonomical assignment via SILVA and assessed via the "q2-quality-control" plugin. (A) Taxon accuracy rate (TAR), taxon detection rate (TDR) and linear regression scores (r-squared) are plotted at each taxonomic level. (B) Expected vs observed abundance of each species in the mock community is plotted at each taxonomic level. (C) Distance between false positive observations vs nearest expected feature. Misclassifications were unique to environmental contaminant taxa and no false negative features were observed (not shown).

14

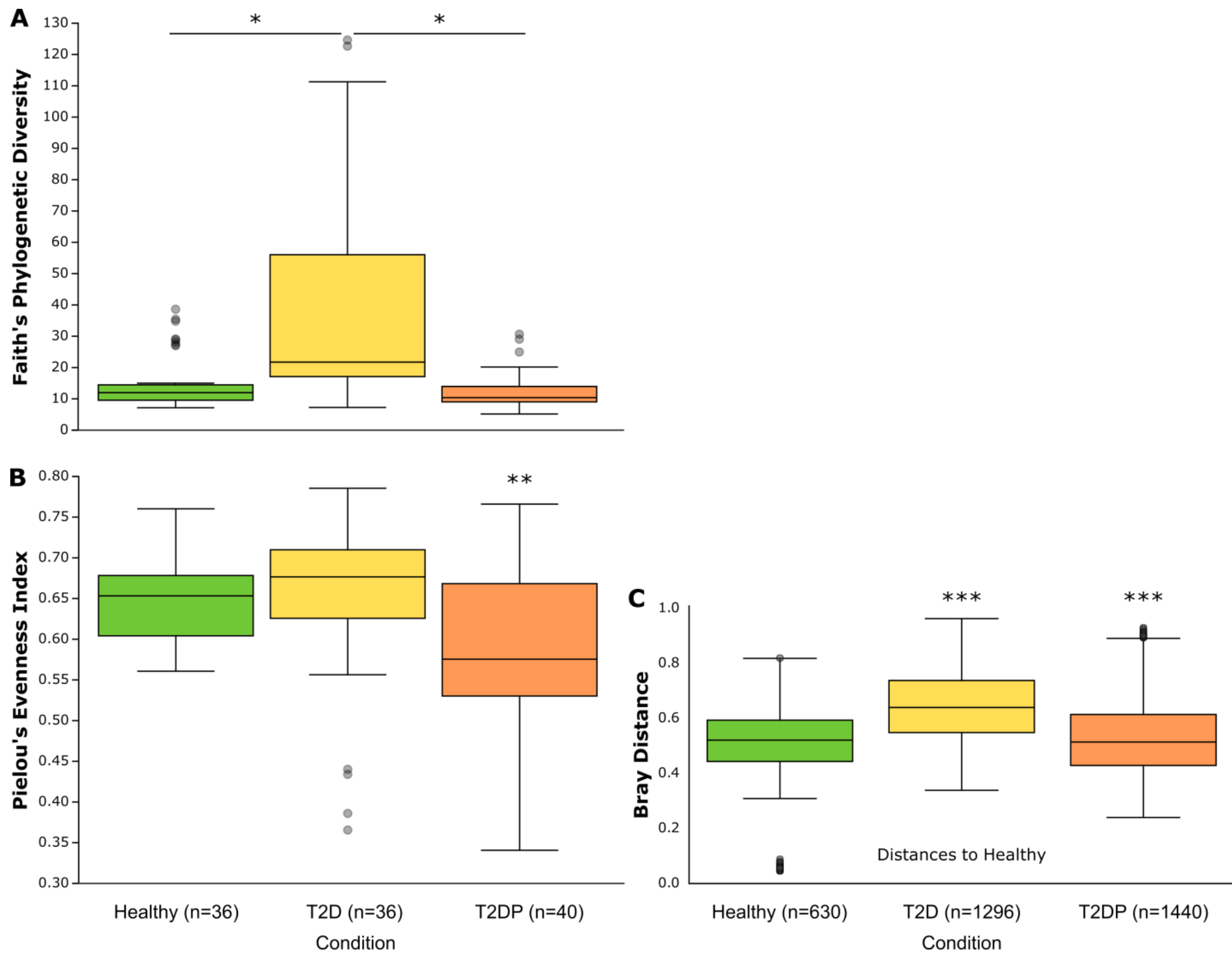


15 **Figure S2. Alpha-diversity at different rarefaction levels.** Related to Figure 1. Total observed
16 OTUs (here ASVs) are plotted (A) vs sequencing depth and each sample that can
17 accommodate that sequencing depth is plotted in (B).
18



19

20 **Figure S3. Run to Run variability quantification.** Related to Figure 1. A Run vs Run Beta
21 diversity comparison was performed for unfiltered data by comparing unweighted Unifrac
22 distances as described in the methods. Pairwise PERMANOVA of Run-1 vs Run-2 (group size
23 of 2, sample size of n=134 total indexes) was performed in 999 permutations resulting in a
24 pseudo-F value of 0.567, p-value of 0.999 and corrected q-value of 0.999. No significant
25 differences were observed between each run.
26



27

28 **Figure S4. 16S Alpha and Beta diversity of serum microbiomes changes with subject**
29 **status.** Related to Figure 1. Species richness is displayed based on Faith's phylogenetic
30 diversity (Faith, 1992) (A) and species evenness based on Pielou's evenness index (Pielou,
31 1966) (B) for healthy, type-II diabetic (T2D) and type-II diabetics with periodontitis (T2DP)
32 samples. Each graph used minimally filtered data for analysis. Significance was determined by
33 Kruskal-Wallis analysis of variance (Kruskal and Wallis, 1952) for each comparison indicated
34 and Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) was applied to generate
35 adjusted q-values. * indicates q-value < 2e-3 vs healthy, ** indicates q-value < 0.024 vs
36 healthy, (C) Beta-diversity Bray Curtis distances. Pairwise PERMANOVA of each category vs
37 each (group size of 3, n=112) was performed in 999 permutations. *** indicates q-value differs
38 from healthy < 0.001.
39

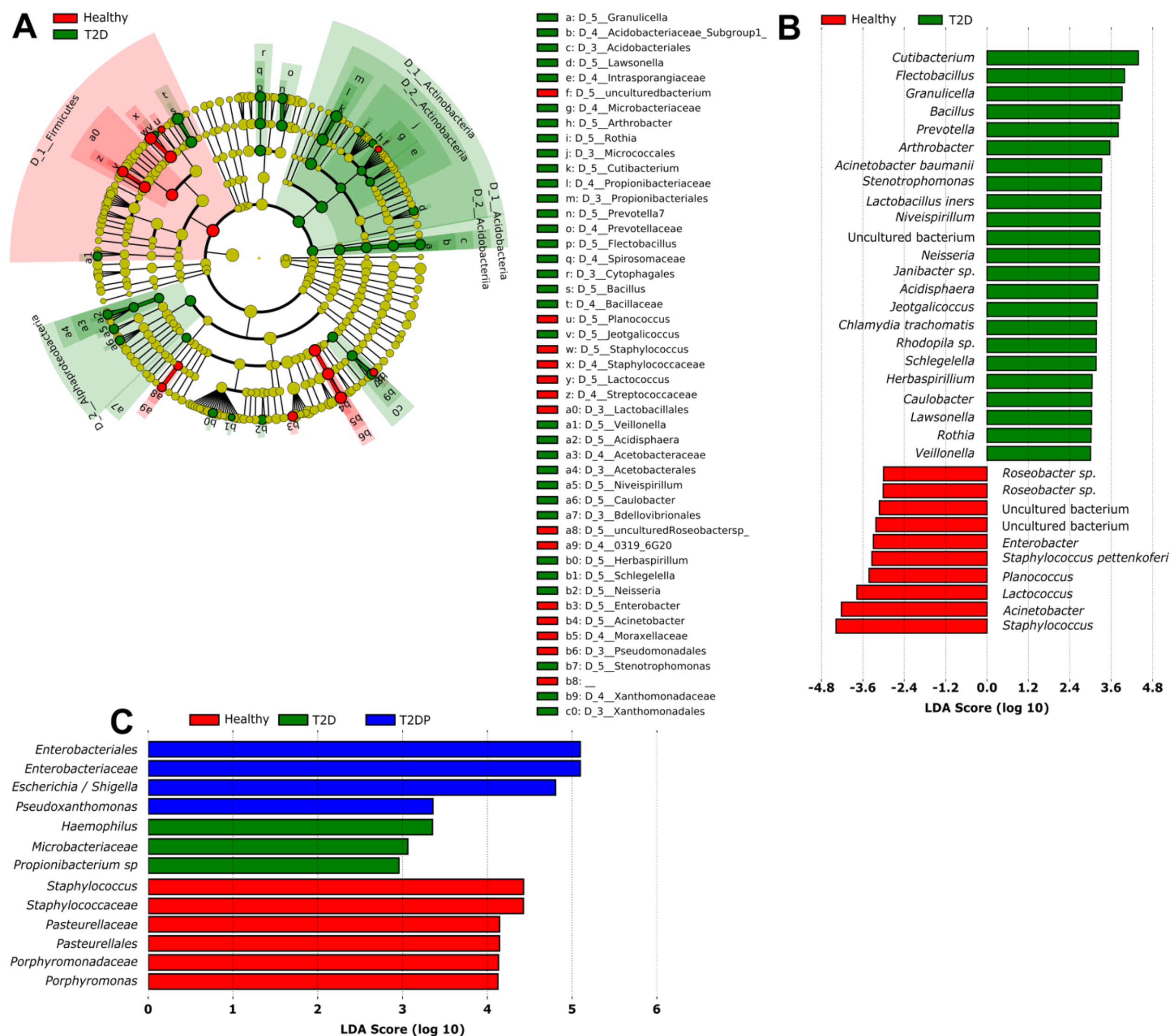
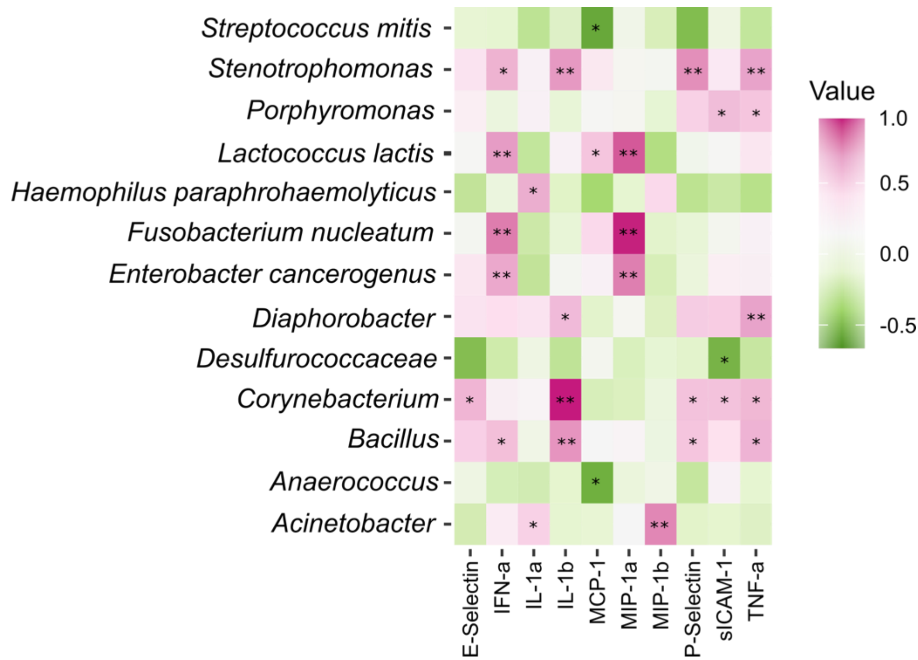


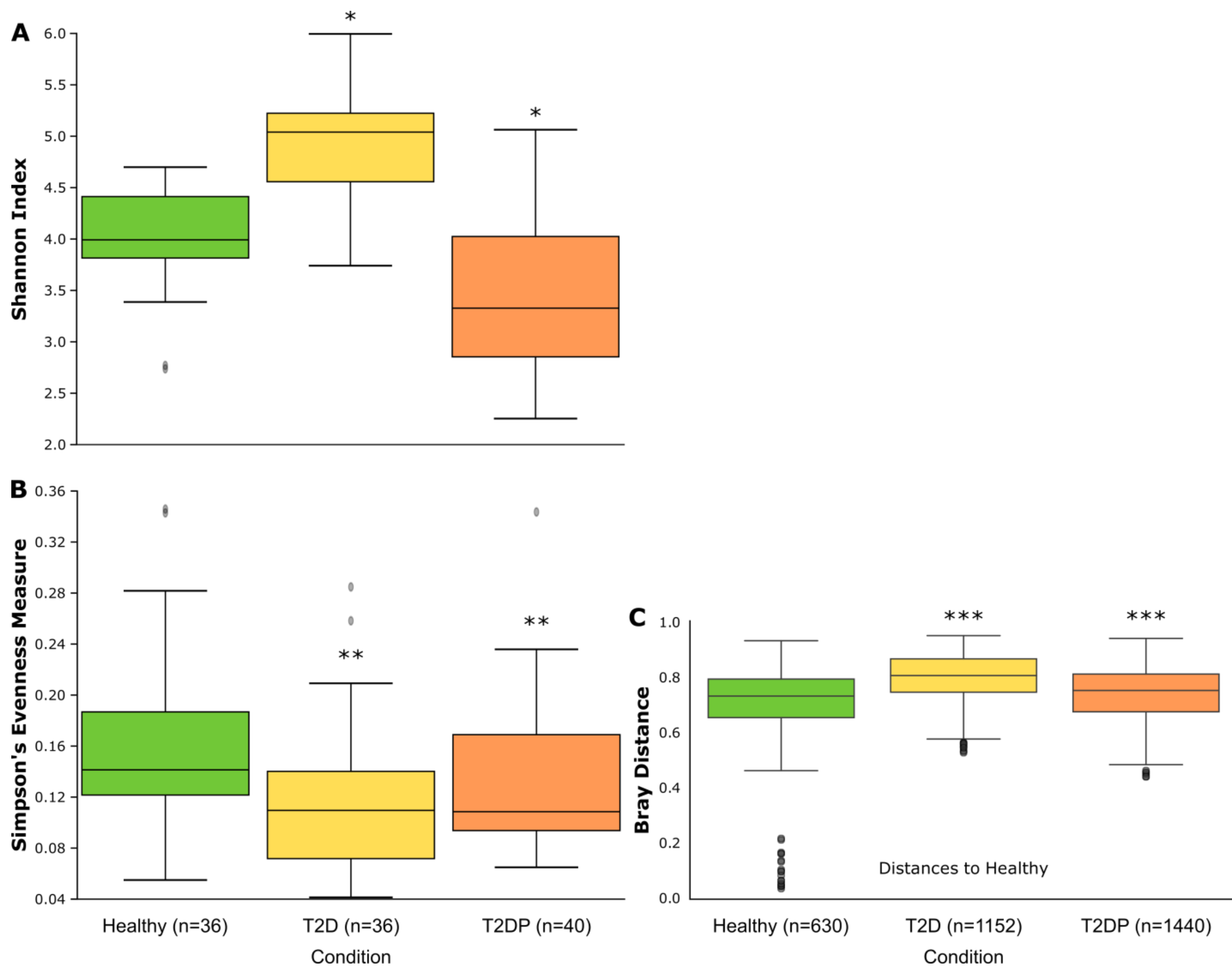
Figure S5. LEfSe Comparisons reveal taxonomical differences between subject groups. Related to Figure 3. (A) LEfSe Cladogram plots (Segata *et al.*, 2011) reveal general taxonomic shifts compared to (B) specifically enriched genera or species in all diabetics vs healthy and (C) compared across all 3 subject groups. Data was generated from DADA2 / QIIME2 analyzed data aligned to the SILVA database that was then filtered for contaminating taxa abundant in no-template control indexes. *Pseudomonas* genus / species level results were also removed for clarity.



49

50 **Figure S6. Cytokine correlation with specific taxa in healthy subjects.** Related to Figure 7.
51 HOMD/RDP assigned taxonomy of MED analyzed 16S data for healthy samples and cytokine
52 concentrations were analyzed via Pearson correlation coefficient in R using the *rcorr* function.
53 Significance was determined using the asymptotic p-values generated by *rcorr* with * = p-value
54 <0.05 ** = p-value <0.01. Data are strictly filtered with taxa present in no-template controls
55 subtracted.
56

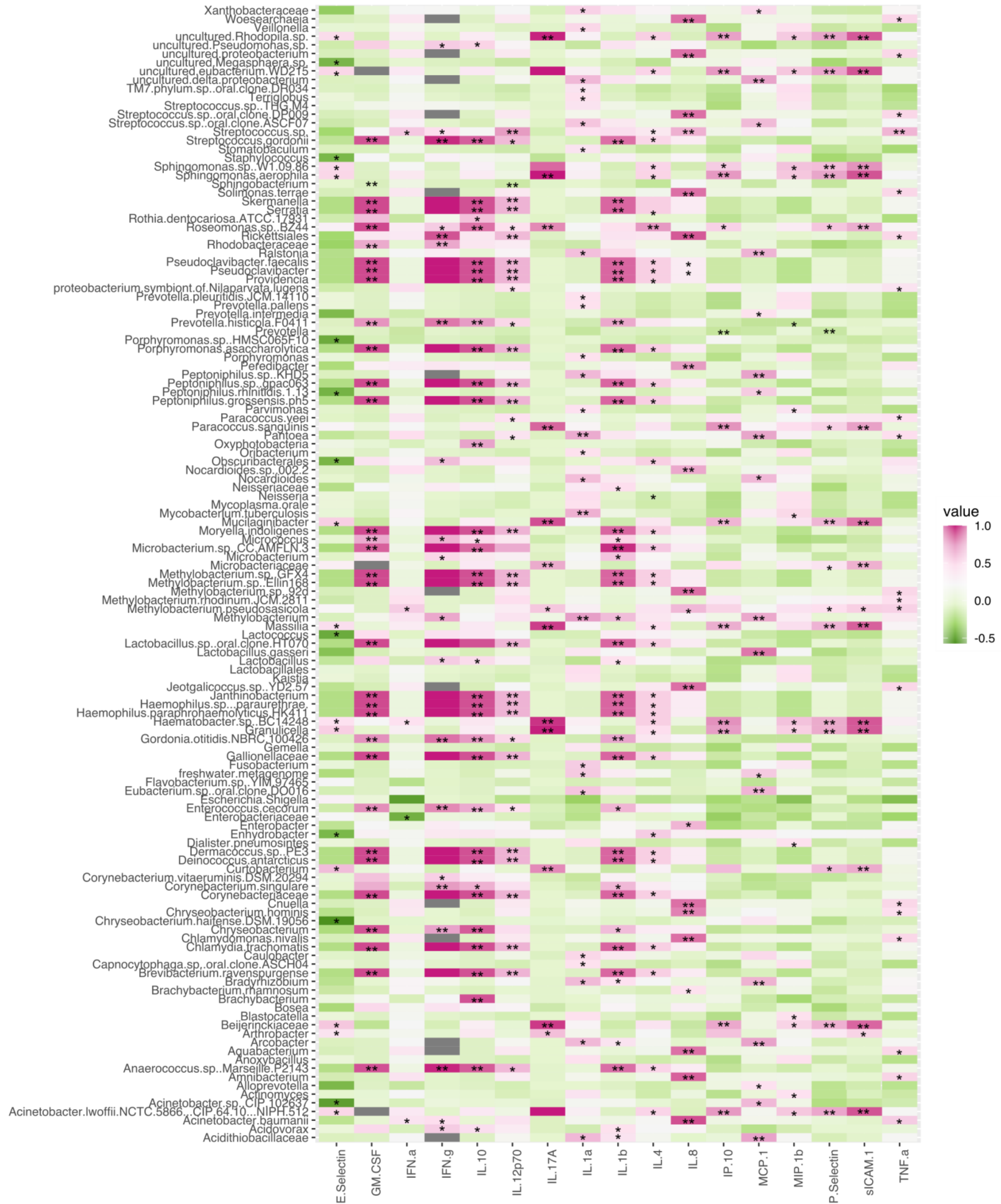
57



58

59 **Figure S7. 16S Alpha and Beta diversity of serum microbiomes changes with subject**
60 **status.** Related to Figure 1. Alpha diversity is displayed based on Shannon's index (Shannon
61 and Weaver, 1975) (A) and species evenness based on Simpson's evenness measure
62 (Simpson, 1949)) (B) for healthy, type-II diabetic (T2D) and type-II diabetics with periodontitis
63 (T2DP) samples. Each graph used minimally filtered data for analysis. Significance was
64 determined by Kruskal-Wallis analysis of variance (Kruskal and Wallis, 1952) for each
65 comparison indicated and Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) was
66 applied to generate adjusted q-values. * indicates q-value < 1e-5, ** indicates q-value < 0.007,
67 (C) Beta-diversity unweighted Unifrac distances. Pairwise PERMANOVA of each category vs
68 each (group size of 3, n=112) was performed in 999 permutations. *** indicates q-value differs
69 from 'healthy' < 0.02.
70

71



72
73 **Figure S8. Host-microbial correlation resolved type-II diabetic associations with**
74 **inflammatory cytokines.** Related to Figure 7. SILVA assigned taxonomy of DADA2 analyzed
75 16S data for both T2D and T2DP samples and cytokine concentrations were analyzed via
76 Pearson correlation coefficient in R using the *rcorr* function. Significance was determined using
77 the asymptotic p-values generated by *rcorr* with * = p-value <0.05 ** = p-value <0.01. Data are
78 strictly filtered with taxa present in no-template controls subtracted.

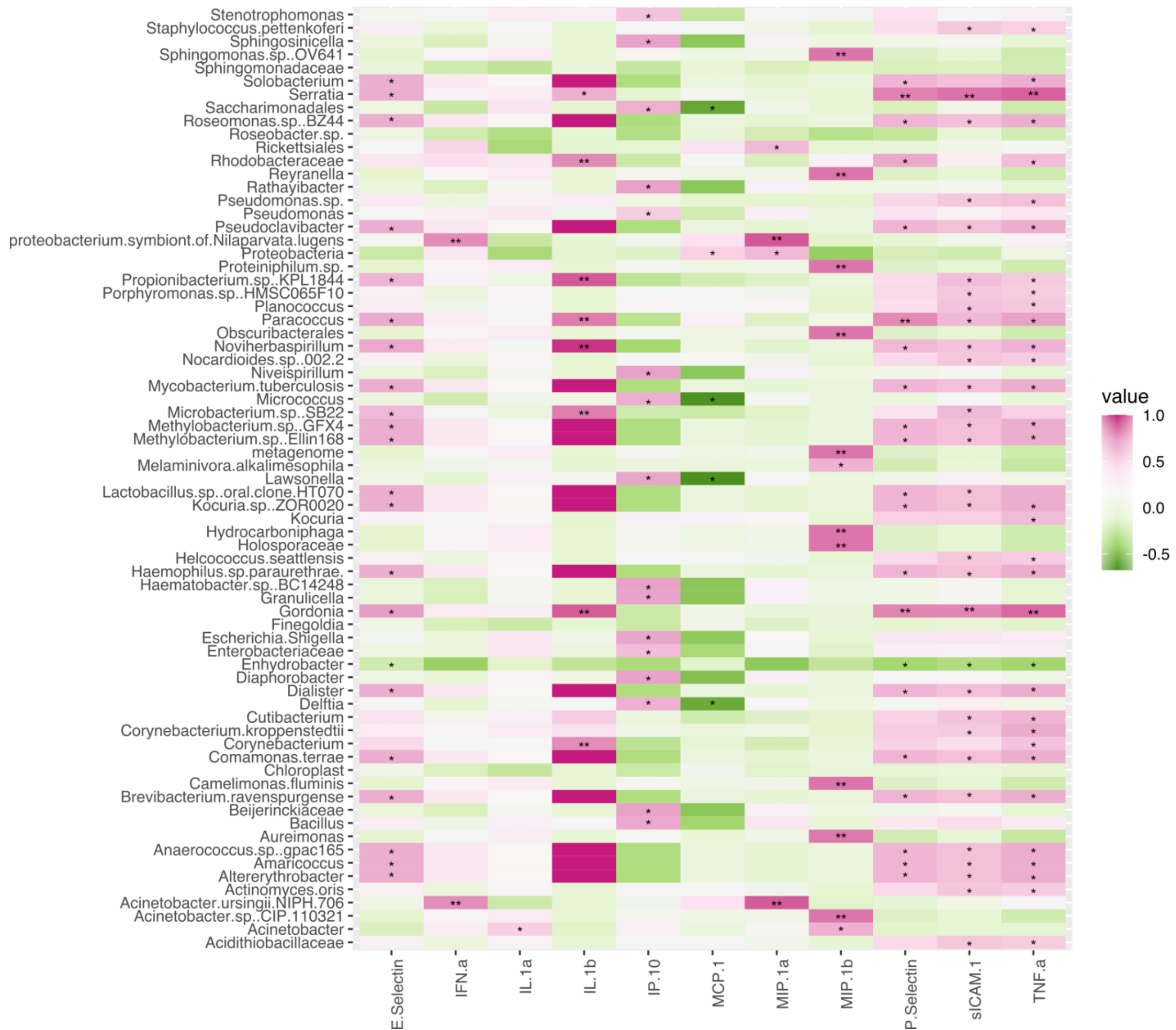


Figure S9. Cytokine correlation with specific taxa in healthy subjects. Related to Figure 7. SILVA assigned taxonomy of DADA2 analyzed 16S data for healthy samples and cytokine concentrations were analyzed via Pearson correlation coefficient in R using the *rcorr* function. Significance was determined using the asymptotic p-values generated by *rcorr* with * = p-value <0.05 ** = p-value <0.01. Data are strictly filtered with taxa present in no-template controls subtracted.

Transparent Methods

Subject recruitment, sampling and storage

Subject recruitment has been described previously (Freire *et al.*, 2017). Peripheral venous blood (~60 ml) was collected from patients diagnosed with T2D and from healthy nondiabetic controls. Blood samples were collected and centrifuged at 2300 rpm, and serum was isolated and frozen at -80°C until analysis under IRB protocol #13-07. All subjects gave signed informed consent prior to study evaluations. Clinical periodontal data and peripheral venous blood were collected. The diagnosis of T2D was made by the subject's primary care physician following American Association of Diabetes guidelines (American Diabetes Association, 2015). Information was collected on subject demographics (age, gender, self-reported ethnicity, and self-reported smoking status), body-mass index (BMI; kg/m2), blood total cholesterol, blood glucose (point-of-care), percent hemoglobin A1C (HbA1c), and periodontal condition (Armitage, 1999). HbA1c was used to determine the level of glycemic control for diabetic subjects. One T2D individual lacked HbA1c measurements but fit based on all other diagnostic criteria (blood glucose >200 mg/dl) as well as cytokine profiles in accord with other T2D individuals. Neutrophil and monocyte cell counts were determined by lab assay (described below). Individuals were excluded if they were taking insulin sensitizers, nonsteroidal anti-inflammatory drugs, or antimicrobials within 3 months of study initiation. Smoking status was defined by CDC NHIS terms (https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm#:~:text=Former%20smoker%3A%20An%20adult%20who,in%20his%20or%20her%20lifetime). Site collection before blood draw: Samples were drawn in clinical phlebotomy settings. From all serum sampled for this microbiome study, a total of 81 subjects (N=24 healthy, N=57 T2D) were included for analysis, all of whom were unrelated and over 18 years of age (range: 28-79 years of age).

Sample preparation for 16S analysis

DNA Extraction

Frozen serum from N=81 samples (N=24 healthy, N=32 T2D and N=25 T2DP) were thawed on ice and aliquots were separated for DNA extraction. DNA was extracted using the Epicentre MasterPure Complete DNA and RNA Purification kit (Lucigen, WI, USA). Extraction was carried out using the manufacturer's instructions, with modifications that enabled bead beating. Briefly, 200µL of Tissue and Cell Lysis Solution (2x), 100µL of nuclease free water and 2µL of Proteinase K were added to Lysing matrix B (LMB) tubes (MP Biomedicals, Santa Ana, CA). Thereafter 100µL of serum was added to the tubes and placed in a Beadbeater (Biospec) for 30 seconds, then placed on ice for 3 minutes and then repeated. Each round of DNA extractions included a no template control, which consisted of 100µL nuclease free water instead of serum. The samples were then incubated and extracted as described in the manufacturer's instructions. Total DNA was precipitated using the manufacturer's instructions, however 300 µL of MPC Protein Precipitation Reagent was used to accommodate the increased volume. DNA was eluted in 50µL of nuclease free water.

16s rDNA primers 27F (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGTTTGATYMTGGCTCAG) and 519R (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGWATTACCGCGGCKGCTG) (Stackebrandt and Goodfellow, 1991; Turner *et al.*, 1999) were used to amplify the V1-V3 regions of 16s rDNA in a 50 µL reaction using 2x Q5 HiFi mastermix (New England Biolabs, Ipswitch, MA) and 23 µL of extracted DNA (35 cycles).

137
138 *Illumina MiSeq Library preparation*
139
140 After the first round of PCR was carried out it was cleaned with Ampure XP beads (Beckman
141 Coulter, Pasadena, CA) and then visualized by agarose gel electrophoresis. Full indices and
142 adapters were added using the Illumina Nextera Index Kit (Illumina, San Diego, CA) by running
143 the second round of PCR (50 ng of template DNA, 5 cycles) in 2x Phusion HF Master Mix (New
144 England Biolabs, Ipswitch, MA). PCR were then cleaned with Ampure XP beads and analyzed
145 by agarose gel electrophoresis and using the Agilent BioAnalyzer DNA1000 chip (Agilent
146 Technologies, Santa Clara, CA). Samples were then quantified and normalized prior to pooling
147 using a Qubit fluorimeter (Invitrogen, Carlsbad, CA). The final pooled library was quantified with
148 the KAPA Biosystems Illumina Kit (KAPA Biosystems, Woburn, MA) via qPCR in a Roche
149 LightCycler480. Samples were sequenced (2x250 bp paired-end) on an Illumina MiSeq
150 (Illumina, San Diego, CA) at the Rhode Island Genomics and Sequencing Center (Kingston, RI).
151
152 *Mock Community Standards Preparation*
153 In order to quantify potential bias due to low template concentrations, PCR amplification, and
154 contamination, we performed mock library assemblies with commercially available bacterial
155 genomic DNA templates (Zymo Research #D6305). We amplified 2 libraries using 10 and 1 ng
156 total starting templates in both sequencing runs for a total of 4 mock libraries synthesized.Total
157 DNA species composition is provided in the manufacturer's instructions.
158 16S clone library resequencing
159
160 Primers were made to amplify regions of 16s rDNA that enable differentiation of the
161 Acinetobacter species. Primers used were
162 oMR328 (TAGCGGCGGACGGGTGAGTAATGCTTA) and
163 oMR329 (TTCCGACTTCATGGAGTCGAGTTGCAGAC). 50 µL reactions in 2x Q5 mastermix
164 (New England Biolabs, Ipswitch, MA) were run for 35 cycles using 23 µL of DNA extracted from
165 serum as the template. The products were then purified using the Qiagen PCR purification kit
166 according to the manufacturer's instructions. Gotaq 2x mastermix (Promega, Madison, WI) was
167 used to generate 'A' overhangs via PCR for 5 cycles using 10 µL of purified DNA as the
168 template. 4µL of the resulting PCR products were then used to generate clones using the TOPO
169 TA Cloning Kit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions which
170 was modified to use 0.5µL of TOPO vector. Clones were then transformed into NEB5α cells
171 using the manufacturer's instructions and plated on LB plates supplemented with 0.1mM IPTG
172 and 40µg/mL Kanamycin. Additionally, 40µl of 40mg/mL of X-gal was spread onto each plate
173 prior to plating to enable blue-white screening. The plates were incubated for 24 hours at 37°C.
174 White colonies were inoculated into liquid cultures of LB supplemented with 40µg/mL
175 kanamycin. Plasmids were extracted using QIAprep Spin Miniprep kit (Qiagen, Venlo,
176 Netherlands). Inserts were sequenced using the M13 Forward (GTAAAACGACGGCCAGTG)
177 and M13 Reverse (CACAGGAAACAGCTATGACC) primers on an Applied Biosystems 3500xl
178 using the "BigDye" Terminator v3.1 Cycle sequencing kit. Sequences were analyzed using
179 NCBI BLAST (NCBI Resource Coordinators, 2018). A maximum of 10 colonies were screened
180 for each sample or until a sequence matched a species belonging to the *Acinetobacter*
181 *baumanii- calcoaceticus* complex.
182
183 Cytokine Quantification
184

185 Frozen (-80°C) subject serum was brought up to room temperature and assayed using the
186 Invitrogen human inflammation 20-plex ProcartaPlex cytokine panel (Thermo Fisher Scientific,
187 Waltham, MA) on a Luminex 200 instrument (Luminex, Austin, TX) in universal assay buffer.
188 Assayed cytokines included: MIP-1α, IL-1β, IL-4, IP-10, IL-6, IL-8, IL-10, IL-12p70, IL-13, IL-
189 17A, IFN-γ, GM-CSF, TNF-α, MIP-1β, IFN-α, MCP-1, P-Selectin, IL-1α, sICAM-1, and E-
190 Selectin. Following manufacturer protocols, all samples were run on a plate with 7 standards
191 (diluted 1:4) and a control (universal assay buffer only), with all samples, standards, and
192 controls run in duplicate similar to our methods previously (Pessoa *et al.*, 2019; Kleinstein *et al.*,
193 2020) .
194
195 Quality control (QC) steps were conducted according to manufacturer recommendations by
196 xPONENT 4.2 software (Affymetrix eBioscience, San Diego, USA). Any standards with <70 or
197 >130 % recovery of beads were invalidated. Samples were required to have a bead count of
198 >30 beads recovered (all samples had >100 beads recovered and none were excluded at this
199 step). Following QC, results were reported as average pg/mL for all measured cytokines. The
200 lower limit of quantification (LLOQ) was determined based on the standard curve (after QC) as
201 the average value of the lowest validated standards. Values at or below the LLOQ for each
202 cytokine were reported at the LLOQ. As samples were run in two batches (with similar LLOQs
203 for each batch), LLOQ for the cytokines (in pg/mL) were averaged across both runs: MIP-
204 1α=1.79, IL-1β=5.58, IL-4=23.99, IP-10= 1.17, IL-6=18.89, IL-8=2.48, IL-10=2.01, IL-
205 12p70=11.72, IL-13=5.63, IL-17A=4.63, IFN-γ=11.93, GM-CSF=14.57, TNF-α=9.66, MIP-
206 1β=7.66, IFN-α=1.32, MCP-1=3.73, P-Selectin=1077.15, IL-1α=0.74, sICAM-1=442.80, and E-
207 Selectin=441.00. To compare cytokine levels between study groups, unpaired t-tests were used
208 with a significance threshold of $p<0.05$ and no assumption of consistent standard deviation.
209
210 Minimum Entropy Decomposition analysis of sequence libraries
211
212 Poor quality reads were discarded using flexbar (Dodt *et al.*, 2012). Reads were then treated as
213 single end reads and trimmed for quality and to a length of 160bp using Trimmomatic (Bolger,
214 Lohse and Usadel, 2014). Minimum Entropy Decomposition (MED) was performed as described
215 in the oligotyping pipeline (Eren *et al.*, 2015), generating a list of node representative sequences
216 and their relative abundances. Taxonomic assignment of representative sequences were
217 generated based on aligning 16s rRNA sequences at 98.5% identity against the eHOMD
218 database (Escapa *et al.*, 2018) and further evaluated using the Ribosomal Database Project
219 (RDP) database (Cole *et al.*, 2014). MED nodes were assigned species level taxonomy based
220 on >98.5% identity matches to eHOMD and/or RDP with eHOMD designations used as 1st
221 priority as our intial aim was to identify oral taxa. Node frequency tables now assigned
222 taxonomy were then used in association with sample metadata for Linear Discriminate Effect
223 Size analysis (LEfSe) (Segata *et al.*, 2011).

224 QIIME2 analysis of sequence libraries
225

226 *All Scripts at end of this document. All work was performed on the University of Rhode Island
227 High Performance Computing Bluewaves cluster:

228 https://web.uri.edu/hpc-research-computing/cluster-specifications/

229 *Data Import*

230 Fastq sequence data were imported using the **import-fastq.sh** script. Generated demux.qzv
231 files were viewed using https://view.qiime2.org/ via the “quality viewer” function to determine
232 location for sequence quality trimming used in the following step.

233 *Fastq Trimming, DADA2 Analysis, Run-run Merging*

234 A second sequencing run of the prepared Illumina libraries was performed to ensure maximum
235 sequencing depth without further PCR amplification of starting template. At the end of the “D2-
236 merge.sh” script we then used the **Rarefaction-alpha.sh** script output (Fig. S2) to determine
237 read depth (22,000) to use in downstream commands.

238 *Initial Alpha and Beta Diversity Measurements and Run-run Comparison Testing*

239 Merged libraries were used in the **Alpha-Beta-An.sh** script to determine initial phylogenetics on
240 unfiltered data using the QIIME2 *core-metrics-phylogenetic* function, Alpha and Beta diversity
241 calculations via QIIME2 *alpha-* or *beta-group-significance* functions as well as Beta diversity
242 difference calculation (*beta-group-significance*) between sequencing Run 1 vs. Run 2 (Figs. S3,
243 S4).

244 *Mock Community Standards Quality Control Testing*

245 This section uses the **mocklibQC1.sh** and **mocklibQC2.sh** scripts. Taxonomy files used are
246 described below. The *-bar-plots.qzv file generated in the 1st script was viewed at
247 https://view.qiime2.org/ and the level 6 data from this plot was exported as a .csv file which
248 provides the reads per each taxa. This data was converted to proportion of total reads per each
249 taxa and converted into the Zymo_actual.tsv file. The Zymo_expected.tsv file was generated by
250 editing the Zymo_actual.tsv file with the proportion for each taxa present as provided in the
251 manufacturers reference:
252 https://files.zymoresearch.com/protocols/_d6305_d6306_zymobiomics_microbial_community_d
253 na_standard.pdf

254 These .tsv inputs were then used in the **mocklibQC2.sh** script to output the visualizations seen
255 in Figure S1.

256 *Classifier Setup and Editing*

257 Initial import of the SILVA 132 classifier was performed by running the following command:

258 *wget https://data.qiime2.org/2019.7/common/silva-132-99-nb-classifier.qza*

259 Classifier setup was then performed using the **classifier.sh** script. Upon searching the silva-
260 taxonomy.qzv file generated, we noted that there were no reads that matched *A. baumanii*
261 sequences at all. We used the rep-seqs.merge.qzv file output from the **D2merge.sh** script
262 saved as a FASTA file from within the QIIME2 viewer function and performed a pairwise blast vs
263 the full length *A. baumanii* 16S sequence from NCBI:

264 >A.baumanii 16s

265 AACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGGGGGAAGGTAGCTTGCTACCGG
266 ACCTAGCGGCGGACGGGTGAGTAATGCTTAGGAATCTGCCTATTAGTGGGGGACAACATC
267 TCGAAAGGGATGCTAATACCGCATACGTCCTACGGGAGAAAGCAGGGGATCTTCGGACCT
268 TGCGCTAATAGATGAGCCTAAGTCGGATTAGCTAGTTGGTGGGGTAAAGGCCTACCAAGG
269 CGACGATCTGTAGCGGGTCTGAGAGGATGATCCGCCACACTGGGACTGAGACACGGCCC

270 AGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGGGGAACCCTGATCCAG
271 CCATGCCGCGTGTGTGAAGAAGGCCTTATGGTTGTAAAGCACTTTAAGCGAGGAGGAGGC
272 TACTTTAGTTAATACCTAGAGATAGTGGACGTTACTCGCAGAATAAGCACCGGCTAACTCT
273 GTGCCAGCAGCCGCGGTAATACAGAGGGTGCGAGCGTTAATCGGATTTACTGGGCGTAAA
274 GCGTGCGTAGGCGGCTTATTAAGTCGGATGTGAAATCCCCGAGCTTAACTTGGGAATTGC
275 ATTCGATACTGGTGAGCTAGAGTATGGGAGAGGATGGTAGAATTCCAGGTGTAGCGGTGA
276 AATGCGTAGAGATCTGGAGGAATACCGATGGCGAAGGCAGCCATCTGGCCTAATACTGAC
277 GCTGAGGTACGAAAGCATGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCATGCCGTA
278 AACGATGTCTACTAGCCGTTGGGGCCTTTGAGGCTTTAGTGGCGCAGCTAACGCGATAAG
279 TAGACCGCCTGGGGAGTACGGTCGCAAGACTAAAACTCAAATGAATTGACGGGGGCCCGC
280 ACAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTGGCCTTGA
281 CATACTAGAAACTTTCCAGAGATGGATTGGTGCCTTCGGGAATCTAGATACAGGTGCTGCA
282 TGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCT
283 TTTCCTTACTTGCCAGCATTTCGGATGGGAACTTTAAGGATACTGCCAGTGACAAACTGGA
284 GGAAGGCGGGGACGACGTCAAGTCATCATGGCCCTTACGGCCAGGGCTACACACGTGCT
285 ACAATGGTCGGTACAAAGGGTTGCTACACAGCGATGTGATGCTAATCTCAAAAAGCCGATC
286 GTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCGGAATCGCTAGTAATCGCGG
287 ATCAGAATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGG
288 GAGTTTGTTGCACCAGAAGTAGCTAGCCTAACTGCAAAGAGGGCGGTTACCACGGTGTGG
289 CCGATGACTGGGGTGAAGT

290 One ASV from the rep-seqs.merge.qzv file (e875355f3179838110485d8f5013d4a6) returned a
291 100% match homology to the above *A. baumanii* sequence. This ASV was just annotated as
292 *Acinetobacter* in the SILVA-132-99 taxonomy. We then edited the silva classifier taxonomy file
293 by first exporting it using the following commands:

294 *module load QIIME2/2019.7*

295 *qiime tools export *

296 *--input-path silva-taxonomy.qza *

297 *--output-path silva-taxonomy*

298 This output the file in *.tsv format which was then opened in a text editor and searched for the
299 ASV node identity and then modified as indicated here:

300 **before**

301 e875355f3179838110485d8f5013d4a6
302 D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pseudomonadales;D
303 _4__Moraxellaceae;D_5__Acinetobacter 0.9999662516375396

304 **after**

305 e875355f3179838110485d8f5013d4a6
306 D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Pseudomonadales;D
307 _4__Moraxellaceae;D_5__Acinetobacter;D_6__Acinetobacter baumanii
308 0.9999662516375396

309 The edited file was then saved as taxonomy-1.tsv and imported via the following command:

310 *qiime tools import *

311 *--input-path /data/mramseylab/classifiers/silva-taxonomy/taxonomy-1.tsv *

312 *--type 'FeatureData[Taxonomy]' *

313 *--input-format TSVTaxonomyFormat *

314 *--output-path silva-mod-taxonomy.qza*

315 *qiime metadata tabulate *

316 *--m-input-file silva-mod-taxonomy.qza *

317 *--o-visualization silva-mod-taxonomy.qzv*

318 The silva-mod-taxonomy.qza file was used for the rest of our analyses.

319 *Data Filtering*

320 Data presented throughout the manuscript is generally described as "Minimally Filtered" or
321 "Strict Filtered". We 1st began with all ASV assigned data from the above scripts and used
322 metadata based filtering to separate out all data from control samples which included no-
323 template and PCR only control indexes. This was done using the **control-filter.sh** script. Data
324 output from the script was exported so it could be viewed and then used to subtract taxa from
325 human-derived datasets in downstream filtering steps. Next we applied the **meta-filter.sh** script
326 to our data to extract only human derived samples for further analysis and filtering. After this we
327 applied the **minimal-filter.sh** script. Outputs from this script were used in all "Minimally Filtered"
328 described data in the manuscript. Minimal filtering included removal of instances of taxa that
329 appeared in only 2 samples or less, features (ASV sequences) that appeared less than 10 times
330 across all samples and any taxa with less than 20 reads per sample.

331 Next we took our output files from the above **control-filter.sh** script and used them as part of
332 the input for the **auto-filter.sh** script to exclude these taxa from the remaining minimally filtered
333 data. While this removed many spurious taxa from our samples we observed high abundance of
334 known aquatic microbial contaminant sequences not typically human associated (ex:
335 *Sphingomonas, Ralstonia*). These and other taxa were manually excluded using the **strict-**
336 **filter.sh** script and further removal of highly abundant *Pseudomonas* sequences were also
337 removed via the **nopa-filter.sh** script. Data filtered to this extent are referred to as "Strictly"
338 filtered in the manuscript.

339 *Re-analysis of Filtered Data*

340 A reanalysis of filtered data was performed 1st for Alpha and Beta diversity measurements via
341 the **alpha-beta2.sh** script (Fig. 1). Further description and statistical analysis of Beta diversity
342 differences were performed in R using the *vegan* package via the metaMDS and Adonis
343 functions primarily as demonstrated in the **MMR20_ellipses_NMDS.R** script below.

344 *LEfSe Analysis Comparison*

345 Initial LEfSe analysis was performed using ASV output abundance data directly from the initial
346 MED pipeline (Fig. 3). Further analysis on QIIME2 / DADA2 assigned strictly filtered data (Fig.
347 S5) was performed first using the **lefse-noTax.sh** script to export and format data for analysis
348 on the LEfSe Galaxy server (https://huttenhower.sph.harvard.edu/galaxy/). This script only
349 outputs ASV node information without taxonomic assignment. Manual taxonomic assignment
350 was performed comparing ASV node names to the rep-seqs-merge.qza file from the

351 **D2merge.sh** script. LEfSe analysis was performed using default settings, *.svg output files were
352 further edited in InkScape software for clarity.

353

354 **Supplemental References**

355 American Diabetes Association (2015) '(2) Classification and diagnosis of diabetes', *Diabetes*
356 *Care*, 38 Suppl, pp. S8–S16.

357 Armitage, G. C. (1999) 'Development of a classification system for periodontal diseases and
358 conditions', *Annals of Periodontology*, 4(1), pp. 1–6.

359 Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and
360 Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society: Series B*
361 *(Methodological)*, 57(1), pp. 289–300.

362 Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina
363 sequence data', *Bioinformatics (Oxford, England)*, 30(15), pp. 2114–2120.

364 Cole, J. R. *et al.* (2014) 'Ribosomal Database Project: data and tools for high throughput rRNA
365 analysis', *Nucleic Acids Research*, 42(Database issue), pp. D633-642.

366 Dodt, M. *et al.* (2012) 'FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation
367 Sequencing Platforms', *Biology*, 1(3), pp. 895–905.

368 Eren, A. M. *et al.* (2015) 'Minimum entropy decomposition: unsupervised oligotyping for
369 sensitive partitioning of high-throughput marker gene sequences', *The ISME journal*, 9(4), pp.
370 968–979.

371 Escapa, I. F. *et al.* (2018) 'New Insights into Human Nostril Microbiome from the Expanded
372 Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human
373 Aerodigestive Tract', *mSystems*. Edited by J. Xu, 3(6), pp. e00187-18,
374 /msystems/3/6/msys.00187-18.atom.

375 Faith, D. P. (1992) 'Conservation evaluation and phylogenetic diversity', *Biological*
376 *Conservation*, 61(1), pp. 1–10.

377 Freire, M. O. *et al.* (2017) 'Neutrophil Resolvin E1 Receptor Expression and Function in Type 2
378 Diabetes', *Journal of Immunology (Baltimore, Md.: 1950)*, 198(2), pp. 718–728.

379 Kleinstein, S. E. *et al.* (2020) *Transcriptomics of Type 2 Diabetic and Healthy Human*
380 *Neutrophils*. preprint. In Review.

381 Kruskal, W. H. and Wallis, W. A. (1952) 'Use of Ranks in One-Criterion Variance Analysis',
382 *Journal of the American Statistical Association*, 47(260), pp. 583–621.

383 NCBI Resource Coordinators (2018) 'Database resources of the National Center for
384 Biotechnology Information', *Nucleic Acids Research*, 46(D1), pp. D8–D13.

385 Pessoa, L. *et al.* (2019) 'Host-Microbial Interactions in Systemic Lupus Erythematosus and
386 Periodontitis', *Frontiers in Immunology*, 10, p. 2602.

387 Pielou, E. C. (1966) 'The measurement of diversity in different types of biological collections',
388 *Journal of Theoretical Biology*, 13, pp. 131–144.

389 Segata, N. *et al.* (2011) 'Metagenomic biomarker discovery and explanation', *Genome Biology*,
390 12(6), p. R60.

391 Shannon, C. E. and Weaver, W. (1975) *The mathematical theory of communication*. Urbana:
392 University of Illinois Press.

393 Simpson, E. H. (1949) 'Measurement of Diversity', *Nature*, 163(4148), pp. 688–688.

394 Stackebrandt, E. and Goodfellow, M. (eds) (1991) *Nucleic acid techniques in bacterial*
395 *systematics*. Chichester ; New York: Wiley (Modern microbiological methods).

396 Turner, S. *et al.* (1999) 'Investigating deep phylogenetic relationships among cyanobacteria and
397 plastids by small subunit rRNA sequence analysis', *The Journal of Eukaryotic Microbiology*,
398 46(4), pp. 327–338

399

400 **Supplemental Data (Data S1.)** Related to Transparent Methods

401 Scripts used for analyses:

402

403 **import-fastq.sh**

```
404 module load QIIME2/2019.7
405
406 for d in /data/mramseylab/raw_reads/2018_Serum/Run* ;
407 do
408 SUFX=${d#*Run}
409
410 qiime tools import \
411   --type 'SampleData[PairedEndSequencesWithQuality]' \
412   --input-path /data/mramseylab/raw_reads/2018_Serum/Run$SUFX \
413   --input-format CasavaOneEightSingleLanePerSampleDirFmt \
414   --output-path /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/demux-
415 $SUFX.qza
416
417 qiime demux summarize \
418   --i-data /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/demux-
419 $SUFX.qza \
420   --o-visualization /data/mramseylab/visualizations/demux-$SUFX.qzv
421
422 done
423
```

424 **D2merge.sh**

```
425
426 module load QIIME2/2019.7
427
428 for d in /data/mramseylab/raw_reads/2018_Serum/Run* ;
429 do
430 SUFX=${d#*Run}
431
432 qiime dada2 denoise-single \
433   --i-demultiplexed-seqs
434 /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/demux-$SUFX.qza \
435   --p-trim-left 13 \
436   --p-trunc-len 250 \
437   --o-table /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/denoise-
438 table-$SUFX.qza \
439   --o-representative-sequences
440 /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/rep-seqs-$SUFX.qza \
441   --o-denoising-stats
442 /data/mramseylab/raw_reads/2018_Serum/Run$SUFX/denoising-stats-
443 $SUFX.qza
444
445 done
446 #Have to MERGE Run1 and Run2 to get pairwise comparison
447 #Must make directory before running below else it will error out.
448
449 mkdir /data/mramseylab/raw_reads/2018_Serum/Merge_runs/
450
451 qiime feature-table merge \
452   --i-tables /data/mramseylab/raw_reads/2018_Serum/Run1/denoise-table-
453 1.qza \
454   --i-tables /data/mramseylab/raw_reads/2018_Serum/Run2/denoise-table-
455 2.qza \
```

```
456     --o-merged-table
457   /data/mramseylab/raw_reads/2018_Serum/Merge_runs/denoise-table-
458   merge.qza
459
460   qiime feature-table merge-seqs \
461     --i-data /data/mramseylab/raw_reads/2018_Serum/Run1/rep-seqs-1.qza \
462     --i-data /data/mramseylab/raw_reads/2018_Serum/Run2/rep-seqs-2.qza \
463     --o-merged-data
464   /data/mramseylab/raw_reads/2018_Serum/Merge_runs/rep-seqs-merge.qza
465
466   qiime feature-table summarize \
467     --i-table /data/mramseylab/raw_reads/2018_Serum/Merge_runs/denoise-
468   table-merge.qza \
469     --o-visualization /data/mramseylab/visualizations/denoise-table-
470   merge.qzv \
471     --m-sample-metadata-file /data/mramseylab/metadata/Serum5.tsv
472
473   qiime feature-table tabulate-seqs \
474     --i-data /data/mramseylab/raw_reads/2018_Serum/Merge_runs/rep-seqs-
475   merge.qza \
476     --o-visualization /data/mramseylab/visualizations/rep-seqs-merge.qzv
477
478   qiime phylogeny align-to-tree-mafft-fasttree \
479     --i-sequences /data/mramseylab/raw_reads/2018_Serum/Merge_runs/rep-
480   seqs-merge.qza \
481     --o-alignment
482   /data/mramseylab/raw_reads/2018_Serum/Merge_runs/aligned-rep-seqs-
483   merge.qza \
484     --o-masked-alignment
485   /data/mramseylab/raw_reads/2018_Serum/Merge_runs/masked-aligned-rep-
486   seqs-merge.qza \
487     --o-tree /data/mramseylab/raw_reads/2018_Serum/Merge_runs/unrooted-
488   tree-merge.qza \
489     --o-rooted-tree
490   /data/mramseylab/raw_reads/2018_Serum/Merge_runs/rooted-tree-merge.qza
```

491 **Rarefaction-alpha.sh**

```
492 module load QIIME2/2019.7
493
494 rawdir=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/
495 procdir=/data/mramseylab/proc_reads/
496 metadir=/data/mramseylab/metadata/
497 visdir=/data/mramseylab/visualizations/
498
499 qiime diversity alpha-rarefaction \
500   --i-table $rawdir\denoise-table-merge.qza \
501   --i-phylogeny $rawdir\rooted-tree-merge.qza \
502   --p-max-depth 25000 \
503   --m-metadata-file $metadir\Serum4.tsv \
504   --o-visualization $visdir\alpha-rarefaction.qzv
```

505

506

507

508

509

510

511 **Alpha-Beta-An.sh**

```
512 module load QIIME2/2019.7
513
514 rawdir=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/
515 procdir=/data/mramseylab/proc_reads/
516 metadir=/data/mramseylab/metadata/
517 cmr="core-metrics-results"
518
519 qiime diversity core-metrics-phylogenetic \
520   --i-phylogeny $rawdir\rooted-tree-merge.qza \
521   --i-table  $rawdir\denoise-table-merge.qza \
522   --p-sampling-depth 22000 \
523   --m-metadata-file $metadir\Serum5.tsv \
524   --output-dir $procdir$cmr\-initial
525
526 qiime diversity alpha-group-significance \
527   --i-alpha-diversity $procdir$cmr\-initial/faith_pd_vector.qza \
528   --m-metadata-file $metadir\Serum5.tsv \
529   --o-visualization $procdir$cmr\-initial/faith-pd-group-
530 significance.qzv
531
532 qiime diversity alpha-group-significance \
533   --i-alpha-diversity $procdir$cmr\-initial/evenness_vector.qza \
534   --m-metadata-file $metadir\Serum5.tsv \
535   --o-visualization $procdir$cmr\-initial/evenness-group-
536 significance.qzv
537
538 array=( unweighted_unifrac_distance_matrix
539 weighted_unifrac_distance_matrix bray_curtis_distance_matrix )
540
```

541 for i in "${array[@]}"

542 do

543

544 qiime diversity beta-group-significance \

545 --i-distance-matrix $procdir$cmr\-initial/$i.qza \

546 --m-metadata-file $metadir\Serum5.tsv \

547 --m-metadata-column Condition \

548 --o-visualization $procdir$cmr\-initial/$i.qzv \

549 --p-pairwise

550

551 done

552

553 #Generates data for run1 vs run2 variability to measure batch effect

554 qiime diversity beta-group-significance \

555 --i-distance-matrix $procdir$cmr\-
556 initial/unweighted_unifrac_distance_matrix.qza \

557 --m-metadata-file $metadir\Serum4.tsv \

558 --m-metadata-column Run \

559 --o-visualization $procdir$cmr\-initial/unweighted-unifrac-Run-
560 significance.qzv \

561 --p-pairwise

562

563 **Classifier.sh**

```
564 module load QIIME2/2019.7
565
566 rawdir=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/
567 clsdir=/data/mramseylab/classifiers/
568
569 #for the silva release 132 99 .fna file
570
571 qiime feature-classifier classify-sklearn \
572   --i-classifier $clsdir\silva-132-99-nb-classifier.qza \
573   --i-reads $rawdir\rep-seqs-merge.qza \
574   --o-classification $clsdir\silva-taxonomy.qza
575
576 #output the taxonomy table to check for A. baumanii strains in it
577 afterwards.
578
579 qiime metadata tabulate \
580   --m-input-file $clsdir\silva-taxonomy.qza \
581   --o-visualization $clsdir\silva-taxonomy.qzv
582
```

583 **Control-filter.sh**

```
584 module load QIIME2/2019.7
585
586 tablein=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/denoise-
587 table-merge.qza
588 clsdir=/data/mramseylab/classifiers/
589 metadir=/data/mramseylab/metadata/
590 visdir=/data/mramseylab/visualizations/
591 filtdir=/data/mramseylab/proc_reads/
592 # filter status of input files, "ctrl-filter" is just for taxa
593 belonging to no template controls
594 fil=ctrl-filter
595
596
597 #must make the directory you are filtering to 1st or else it will
598 error
599 mkdir $filtdir$fil
600
601
602 qiime feature-table filter-samples \
603   --i-table $tablein \
604   --m-metadata-file $metadir\Serum5.tsv \
605   --p-where "[Source]='Control'" \
606   --o-filtered-table $filtdir$fil/$fil-table.qza
607
608 qiime taxa collapse \
609   --i-table $filtdir$fil/$fil-table.qza \
610   --i-taxonomy $clsdir\silva-mod-taxonomy.qza \
611   --p-level 6 \
612   --o-collapsed-table $filtdir$fil/$fil-collapse-table.qza
```

```
613
614  qiime feature-table relative-frequency \
615    --i-table $filtdir$fil/$fil-collapse-table.qza \
616    --o-relative-frequency-table $filtdir$fil/$fil-relative-collapse-
617  table.qza
618
619  qiime tools export \
620    --input-path $filtdir$fil/$fil-relative-collapse-table.qza \
621    --output-path $filtdir$fil/
622
623  biom convert \
624  -i $filtdir$fil/feature-table.biom \
625  -o $filtdir$fil/$fil-relative-collapse-table.txt \
626  --header-key "taxonomy" \
627  --to-tsv
628
629  #Use above taxa table to filter out based on taxa present in controls
630
631  qiime feature-table filter-features \
632    --i-table $tablein \
633    --m-metadata-file $filtdir$fil/$fil-collapse-table.qza \
634    --o-filtered-table $filtdir$fil/$fil-excluded-table.qza \
635    --p-exclude-ids
636
637  #Use excluded table to generate barplot for checking
638
639  qiime taxa barplot \
640    --i-table $filtdir$fil/$fil-excluded-table.qza \
641    --i-taxonomy $clsdir\silva-mod-taxonomy.qza \
```

```
642    --m-metadata-file $metadir\Serum5.tsv \
643    --o-visualization $filtdir$fil/$fil-excluded-table.qzv
644
```

645 **Meta-filter.sh**

```
646 module load QIIME2/2019.7
647
648 tablein=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/denoise-
649 table-merge.qza
650 clsdir=/data/mramseylab/classifiers/
651 metadir=/data/mramseylab/metadata/
652 visdir=/data/mramseylab/visualizations/
653 filtdir=/data/mramseylab/proc_reads/
654 # filter status of input files, "initial" is the 1st pass no filter at
655 all
656 # "initial-human" is the initial data but for only human samples, no
657 mock or control samples
658 fil=initial-human
659
660 #note must make directores needed BEFORE running the below command
661 mkdir $filtdir$fil
662
663 qiime feature-table filter-samples \
664   --i-table $tablein \
665   --m-metadata-file $metadir\Serum5.tsv \
666   --p-where "[Source]='Human'" \
667   --o-filtered-table $filtdir$fil/$fil-table.qza
668
```

669 **Auto-filter.sh**

```
670 module load QIIME2/2019.7
671
672 tablein=/data/mramseylab/proc_reads/minF-hum/table3.qza
673 clsdir=/data/mramseylab/classifiers/
674 metadir=/data/mramseylab/metadata/
675 visdir=/data/mramseylab/visualizations/
676 filtdir=/data/mramseylab/proc_reads/
677
678 # filter status of input files, "ctrl-filter" is just for taxa
679 belonging to no template controls
680 fil=AF-hum
681 #must make the directory you are filtering to 1st or else it will
682 error
683 mkdir $filtdir$fil
684
685 #Filtering list from mock samples was generated previously with the
686 control-filter.sh script
687
688 qiime feature-table filter-features \
689   --i-table $tablein \
690   --m-metadata-file /data/mramseylab/proc_reads/ctrl-filter/ctrl-
691 filter-collapse-table.qza \
692   --o-filtered-table $filtdir$fil/$fil-excluded-table.qza \
693   --p-exclude-ids
694
695 #Use excluded table to generate barplot for checking
696 qiime taxa barplot \
697   --i-table $filtdir$fil/$fil-excluded-table.qza \
698   --i-taxonomy $clsdir\silva-mod-taxonomy2.qza \
699   --m-metadata-file $metadir\Serum5.tsv \
```

```
700      --o-visualization $filtdir$fil/$fil-excluded-table.qzv
701
702    #Generate viewable feature table to look for ASVs of interest /
703    filtering stats
704    qiime feature-table summarize \
705      --i-table $filtdir$fil/$fil-excluded-table.qza \
706      --o-visualization $filtdir$fil/$fil-excluded-table2.qzv \
707      --m-sample-metadata-file $metadir\Serum5.tsv
708
```

709 **Strict-filter.sh**

```
710 module load QIIME2/2019.7
711
712 tablein=/data/mramseylab/proc_reads/AF-hum/AF-hum-excluded-table
713 clsdir=/data/mramseylab/classifiers/
714 metadir=/data/mramseylab/metadata/
715 visdir=/data/mramseylab/visualizations/
716 filtdir=/data/mramseylab/proc_reads/
717
718 # filter status of input files, "ctrl-filter" is just for taxa
719 belonging to no template controls
720 fil=SF-hum
721 #must make the directory you are filtering to 1st or else it will
722 error
723 mkdir $filtdir$fil
724
725 #Filtering list from mock samples was generated previously with the
726 auto-filter.sh script
727
728 #Tidying up name here so I can repeatedly use the same input below
729 through the loop
730 cp $tablein.qza /$filtdir$fil/$fil-table.qza
731 #Array here contains all taxa I wish to remove
732 array=( Chloroplast Mitochondria Ralstonia Chryseobacterium
733 Sphingomonas Enhydrobacter Bradyrhizobium Sphingomonadales Rhizobiales
734 Rhodobacterales Sphingobacteriales Halomonadaceae Deinococcales )
735
736 for i in "${array[@]}"
737 do
738
739 qiime taxa filter-table \
```

```
740   --i-table /$filtdir$fil/$fil-table.qza \
741   --i-taxonomy $clsdir\silva-mod-taxonomy2.qza \
742   --p-mode contains \
743   --p-exclude "$i" \
744   --o-filtered-table /$filtdir$fil/$fil-table.qza
745
746 done
747 #Use excluded table to generate barplot for checking
748 qiime taxa barplot \
749   --i-table /$filtdir$fil/$fil-table.qza \
750   --i-taxonomy $clsdir\silva-mod-taxonomy2.qza \
751   --m-metadata-file $metadir\Serum5.tsv \
752   --o-visualization /$filtdir$fil/$fil-table-barplot.qzv
753
754 #Generate viewable feature table to look for ASVs of interest /
755 filtering stats
756 qiime feature-table summarize \
757   --i-table /$filtdir$fil/$fil-table.qza \
758   --o-visualization /$filtdir$fil/$fil-table.qzv \
759   --m-sample-metadata-file $metadir\Serum5.tsv
760
```

761 **Nopa-filter.sh**

```
762 module load QIIME2/2019.7
763
764 tablein=/data/mramseylab/proc_reads/SF-hum/SF-hum-table.qza
765 clsdir=/data/mramseylab/classifiers/
766 metadir=/data/mramseylab/metadata/
767 visdir=/data/mramseylab/visualizations/
768 filtdir=/data/mramseylab/proc_reads/
769
770 # filter status of input files, "ctrl-filter" is just for taxa
771 belonging to no template controls
772 fil=NP-hum
773 #must make the directory you are filtering to 1st or else it will
774 error
775 mkdir $filtdir$fil
776
777 #Filtering list from mock samples was generated previously with the
778 strict-filter.sh script
779
780 qiime taxa filter-table \
781   --i-table $tablein \
782   --i-taxonomy $clsdir\silva-mod-taxonomy2.qza \
783   --p-mode contains \
784   --p-exclude "Pseudomonas" \
785   --o-filtered-table $filtdir$fil/$fil-table.qza
786
787 #Use excluded table to generate barplot for checking
788 qiime taxa barplot \
789   --i-table $filtdir$fil/$fil-table.qza \
790   --i-taxonomy $clsdir\silva-mod-taxonomy2.qza \
```

```
791     --m-metadata-file $metadir\Serum5.tsv \
792     --o-visualization $filtdir$fil/$fil-table-barplot.qzv
793
794   #Generate viewable feature table to look for ASVs of interest /
795   filtering stats
796   qiime feature-table summarize \
797     --i-table $filtdir$fil/$fil-table.qza \
798     --o-visualization $filtdir$fil/$fil-table.qzv \
799     --m-sample-metadata-file $metadir\Serum5.tsv
800
```

801 **Alpha-beta2.sh**

```
802 module load QIIME2/2019.7
803
804 rawdir=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/
805 procdir=/data/mramseylab/proc_reads/
806 metadir=/data/mramseylab/metadata/
807 visdir=/data/mramseylab/visualizations/
808 cmr="core-metrics-results"
809 # filter status of input files, "initial" is the 1st pass no filter at
810 all
811 fil=-initial
812
813 #note change name of input tables for the 1st command below. Some
814 input tables did not have standardized filename conventions
815 #changed to AF-hum-table.qza and table3.qza to minF-hum-table.qza
816
817 array=( minF AF SF NP )
818
819 for i in "${array[@]}"
820 do
821
822 qiime diversity core-metrics-phylogenetic \
823   --i-phylogeny $rawdir\rooted-tree-merge.qza \
824   --i-table  $procdir$i\-hum/$i\-hum-table.qza \
825   --p-sampling-depth 5000 \
826   --m-metadata-file $metadir\Serum5.tsv \
827   --output-dir $procdir$i\-hum/$cmr
828
829 qiime diversity alpha-group-significance \
830   --i-alpha-diversity $procdir$i\-hum/$cmr/faith_pd_vector.qza \
```

```
831     --m-metadata-file $metadir\Serum5.tsv \
832     --o-visualization $procdir$i\-hum/$cmr/faith-pd-group-
833   significance.qzv
834
835   qiime diversity alpha-group-significance \
836     --i-alpha-diversity $procdir$i\-hum/$cmr/evenness_vector.qza \
837     --m-metadata-file $metadir\Serum5.tsv \
838     --o-visualization $procdir$i\-hum/$cmr/evenness-group-
839   significance.qzv
840
841   done
842
843   #dual array taking the directories above and then running the next
844   command on the 3 filenames in array2 for each directory prefix in
845   array 1
846   #note different array command syntax from above vs below
847
848   array1=( minF AF SF NP )
849   array2=( unweighted_unifrac_distance_matrix
850   weighted_unifrac_distance_matrix bray_curtis_distance_matrix )
851   for indirs in ${array1[@]}
852   do
853       for infils in ${array2[@]}
854       do
855
856   qiime diversity beta-group-significance \
857     --i-distance-matrix $procdir$indirs\-hum/$cmr/$infils\.qza \
858     --m-metadata-file $metadir\Serum5.tsv \
859     --m-metadata-column Condition \
860     --o-visualization $procdir$indirs\-hum/$cmr/$infils\.qzv \
861     --p-pairwise
```

```
862         done
863     done
864
```

865 **MMR20_ellipses_NMDS.R**

```
866 setwd("E:/2019_Diabetic_Serum/Figure Materials/R-Plots/")
867 library(vegan)
868 library(ggplot2)
869 library(dplyr)
870 set.seed(30)
871
872 metadata<-read.csv("Serum5.tsv", header=TRUE, sep="\t",
873 stringsAsFactors = F)
874 featuredf<-read.csv("NP-hum-features.tsv", header=TRUE, sep="\t",
875 stringsAsFactors = F)
876
877 #making a subset of human only metadata and a few other criteria.
878 metadata <- metadata %>% filter(Source == 'Human') %>%
879 select("sample.id","Condition","Run","Source","ABC")
880
881 #Fixing 1st row to make it row names
882 featuredf <- data.frame(featuredf[,-1], row.names = featuredf[,1])
883
884 #transposing trial1 table to match Evelyns data format
885 featuredf <- as.data.frame(t(featuredf))
886
887 #Brings tables into agreement on matching sample.id values
888 featuredf = featuredf %>% mutate(sample.id = rownames(featuredf))
889
890 #Brings tables into agreement on matching sample.id values
891 table_all = left_join(featuredf, metadata)
892
893 #This gets metadata to agree with row numbers of the features table
894 after splitting out from table_all
```

```
895 metadata <-
896 select(table_all,"sample.id","Condition","Run","Source","ABC")
897
898 #filter the "sample.id" column off the very end of the dataframe and
899 turn it into rownames
900 #NOTE THIS VALUE MUST BE CHANGED FOR EACH INPUT
901 featuredf <- data.frame(featuredf[,-1101], row.names =
902 featuredf[,1101])
903
904 # normalize data by sum of ASVs in each sample (this was from ZP's
905 code)
906 featuredf <- sweep(featuredf,2,colSums(featuredf),`/`)
907
908 ##Above this was all data input / manipulation, Below this is data
909 analysis and plotting##
910
911 ### The function metaMDS is used to calculate the dissimilarity matrix
912 using the bray curtis distance metrics and at the same
913 #time generates the values from the dissimilarity matrix for  an
914 ordination plot.
915 MDS <-metaMDS(featuredf ,distance = "bray", k = 3, trymax = 500)
916
917 ## Next, extract the x and y coordinates from the MDS plot into a new
918 data frame and add the metadata factors to the coordinates the data
919 should be plotted based on.
920 #MMR- added ABC=as.factor and changed [,3] to [,5] to agree with my
921 own metadata file.
922 NMDS_t1=data.frame(NMDS1=MDS$point[,1],NMDS2=MDS$point[,2],
923                    Condition=table_all$Condition,ABC=table_all$ABC)
924
925 ## set theme for following plots
926 theme_set(theme_bw())
927
```

928 ## Generate the ordination based on the solution from above and
929 selected grouping factor (Condition)

930 plot.new()

931 ord<-ordiellipse(MDS, table_all$Condition,

932   display = "sites", kind ="sd", conf = 0.95, label
933 = T)

934 dev.off()

935

936 ## Data frame df_ell_t1 contains values to show ellipses. It is
937 calculated with function veganCovEllipse which is hidden in vegan
938 package.

939 #This function is applied to each level of NMDS (group) and it uses
940 also function cov.wt to calculate covariance matrix.

941 veganCovEllipse<-function (cov, center = c(0, 0), scale = 1, npoints =
942 100)

943 {

944   theta <- (0:npoints) * 2 * pi/npoints

945   Circle <- cbind(cos(theta), sin(theta))

946   t(center + scale * t(Circle %*% chol(cov)))

947 }

948

949 ##Generate ellipse points based on 95% confidence (SD) intervals

950 #Reference : http://stackoverflow.com/questions/13794419/plotting-
951 ordiellipse-function-from-vegan-package-onto-nmds-plot-created-in-
952 ggplot

953 #Data frame df_ell contains values to show ellipses. It is calculated
954 with function veganCovEllipse which is hidden in vegan package. This
955 function is applied to each level of NMDS (group) and it uses also
956 function cov.wt to calculate covariance matrix.

957 df_ell_t1 <- data.frame()

958 for(g in levels(NMDS_t1$Condition)){

959   if(g!="" && (g %in% names(ord))){

960   df_ell_t1 <- rbind(df_ell_t1,
961 cbind(as.data.frame(with(NMDS_t1[NMDS_t1$Condition==g,],

962

963 veganCovEllipse(ord[[g]]$cov,ord[[g]]$center,ord[[g]]$scale))),Conditi
964 on=g))}}

965

966 ## Calculate p-value:

967 adon_t1<-adonis2(featuredf ~Condition, data=metadata,
968 by=NULL,method="bray", k=3)

969

970 NMDSplot_t1<-ggplot(data=NMDS_t1,aes(NMDS1,NMDS2,col=Condition))+

971   #update from Evelyn to add metadata text to plot if wanted

972   #This is useful to identify outlier points if using sample ID text
973 here, Remove for final plot

974 ###
975 geom_text(aes(NMDS1,NMDS2,label=table_all$sample.id),size=2,vjust=0)+

976   # add the p-value in the bottom right corner

977   annotate("text",x=min(NMDS_t1$NMDS1),y=min(NMDS_t1$NMDS2-0.5),

978   label=paste("p= ", adon_t1$`Pr(>F)`[1]),size=3)+

979   # draw the ellipses and define color based on the grouping factor

980   geom_path(data=df_ell_t1, aes(x=NMDS1, y=NMDS2, linetype=Condition),
981 size=1)+

982   #scale_linetype_manual(values=c("4-Con"="dotted","3-S4"="solid","2-
983 RI-old"="longdash"))+

984   #scale_colour_manual(values=c("4-Con"="red", "3-S4"="darkgreen", "2-
985 RI-old"="purple"))+

986   # add the points per sample and define shape based on TankLocation

987   geom_point(aes(shape=ABC), size=2) +

988   # Reorder the legend

989   guides(color = guide_legend(order=1),lty= guide_legend(order=1),

990   shape = guide_legend(order=2), legend.position = "bottom")

991

992 # Adding other Aethestics

```
993  NMDS_by_Trial1 <- NMDSplot_t1 + theme(axis.text.y =
994  element_text(size="12", color="black"), axis.title.y =
995  element_text(face="bold",size="12", color="black"))+
996    theme(axis.text.x = element_text(size="12", color="black"),
997  axis.title.x.bottom = element_text(face="bold",size="12",
998  color="black"))+
999    theme(axis.text.x.top= element_text(face="bold",size="12",
1000 color="black")) +
1001   #change name of the legend
1002   theme(legend.title=element_blank())+
1003   theme(legend.position = "right", legend.title =
1004 element_text(colour="black", size=16, face="bold"))
1005
1006
1007 print(NMDS_by_Trial1)
1008 print(NMDSplot_t1)
1009
```

1010 **Lefse-noTax.sh**

```
1011 module load QIIME2/2019.7
1012
1013 clsdir=/data/mramseylab/classifiers/
1014 metadir=/data/mramseylab/metadata/
1015 visdir=/data/mramseylab/visualizations/
1016 filtdir=/data/mramseylab/proc_reads/
1017 rawdir=/data/mramseylab/raw_reads/2018_Serum/Merge_runs/
1018 procdir=/data/mramseylab/proc_reads/
1019
1020 array=( minF-hum AF-hum SF-hum NP-hum )
1021
1022 for i in "${array[@]}"
1023 do
1024
1025 qiime feature-table relative-frequency \
1026  --i-table $filtdir$i/$i\-table.qza \
1027  --o-relative-frequency-table $filtdir$i/$i\-table.notax.qza
1028
1029 qiime tools export \
1030  --input-path $filtdir$i/$i\-table.notax.qza \
1031  --output-path $filtdir$i/
1032
1033 #note, must use single hashes for -i / -o unlike other qiime commands.
1034 biom convert \
1035  -i $filtdir$i/feature-table.biom \
1036  -o $filtdir$i/$i\-notax.table.txt \
1037  --header-key “taxonomy” \
1038  --to-tsv
```

1039

1040 done

1041

1042