

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/137377/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Weiser, Rebecca , Rye, Phillip D. and Mahenthiralingam, Eshwar 2021. Implementation of microbiota analysis in clinical trials for cystic fibrosis lung infection: experience from the OligoG phase 2b clinical trials. *Journal of Microbiological Methods* 181 , 106133. 10.1016/j.mimet.2021.106133

Publishers page: <http://dx.doi.org/10.1016/j.mimet.2021.106133>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Supplementary Data File

Implementation of microbiota analysis in clinical trials for cystic fibrosis lung infection: experience from the OligoG Phase 2b clinical trials

Authors: Rebecca Weiser^{1*}, Philip D. Rye², and Eshwar Mahenthiralingam^{1*}

Affiliations:

¹ Microbiomes, Microbes and Informatics Group, Organisms and Environment Division, School of Biosciences, Cardiff University, The Sir Martin Evans Building, Museum Avenue, Cardiff, Wales, CF10 3AX, UK.

²AlgiPharma AS, Industriveien 33, N-1337 Sandvika, Norway

* Corresponding authors:

Rebecca Weiser, School of Biosciences, Cardiff University, The Sir Martin Evans Building, Museum Avenue, Cardiff, Wales, CF10 3AX, UK.

Co-correspondence: Eshwar Mahenthiralingam, School of Biosciences, Cardiff University, The Sir Martin Evans Building, Museum Avenue, Cardiff, Wales, CF10 3AX, UK.

Tel: +44 (0)29 2087 5875, Fax: +44 (0)29 2087 4305

E-mail addresses: WeiserR@cardiff.ac.uk (R. Weiser), phil.rye@algipharma.com (P.D. Rye), MahenthiralingamE@cardiff.ac.uk (E. Mahenthiralingam)

ORCID ID:

0000-0003-3983-3272 (R. Weiser)

0000-0001-9014-3790 (E. Mahenthiralingam).

0000-0001-7762-3300 (P.D. Rye)

43	Contents
44	Supplementary tables
45	Supplementary Table S1. Analyses conducted on samples in the Bcc trial (13 patients, 155
46	samples)
47	Supplementary Table S2. Analyses conducted on samples in the <i>P. aeruginosa</i> trial (45
48	patients, 511 samples)
49	Supplementary Table S3. Statistical analysis of Shannon diversity at three paired start-end
50	points for the Bcc and <i>P. aeruginosa</i> trials
51	Supplementary Table S4. Statistical analysis of the total abundance key pathogens at three
52	paired start-end points for the Bcc and <i>P. aeruginosa</i> trials
53	Supplementary Table S5. Statistical analysis of the relative abundance of key pathogens at
54	three paired start-end points for the Bcc and <i>P. aeruginosa</i> trials
55	Supplementary Table S6. Supplementary Table 3. The relative and total abundance of
56	<i>Burkholderia</i> in Bcc trial samples
57	
58	Supplementary figures
59	Supplementary Figure S1. Paired samples from the Bcc and <i>P. aeruginosa</i> trials were
60	concordant as shown by high Pearson product-moment correlation coefficients (PPMCC).
61	Supplementary Figure S2. The lung microbiota in the <i>P. aeruginosa</i> trial was linked to the
62	individual rather than treatment.
63	Supplementary Figure S3. Analysis of microbiota present between paired start and end
64	time-points collected during the Bcc trial.
65	Supplementary Figure S4. Correlation of loss of bacterial diversity with poor lung function
66	for the Oligo G trial participants.
67	
68	Supplementary methods
69	Supplementary method S1: Standardised protocol 1, Sputum sample processing for DNA
70	extraction
71	Supplementary method S2: Standardised protocol 2, Identification of <i>Burkholderia</i> species
72	in sputum samples
73	Supplementary method S3: Standardised protocol 3, Quantitative PCR (qPCR) using
74	TaqMan probes to determine bacterial load (<i>P. aeruginosa</i> and <i>Burkholderia</i>)
75	Supplementary method S4: Standardised protocol 4, Ribosomal RNA Intergenic Spacer
76	Analysis (RISA)
77	Supplementary method S5: Standardised protocol 5, Bacterial diversity analysis (16S
78	rRNA gene sequencing and analysis)
79	Supplementary method S6: R scripts for statistical analysis of microbiota data

80 **Supplementary Table S1. Analyses conducted on samples in the *Bcc* trial (13 patients,**
81 **155 samples)**
82

Patient	Visit number (paired samples)											
	V1		V2		V4		V5		V7		V8	
	1	2	1	2	1	2	1	2	1	2	1	2
27610-001	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-002	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	-	ABC	A
27610-003	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-004	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-005	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-006	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-007	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-008	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-009	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27610-011	ABCD	A	ABC	A	ABC	A	ABC	A	AB	AC	ABC	A
27611-002	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27611-005	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A
27611-006	ABCD	A	ABC	A	ABC	A	ABC	A	ABC	A	ABC	A

83
84 Footnotes: **A**, sample subjected to 16s rRNA gene sequencing and used in paired analysis to determine
85 microbiota concordance; **B**, sample used in qPCR analysis; **C**, 16S rRNA sequencing results used to
86 examine genus relative abundance, alpha and beta diversity metrics and for identification of trends in
87 the dataset; **D**, sample used to determine the identity of the infecting *Burkholderia* species by *recA* and
88 *gyrB* gene amplification, sequencing and analysis; '-' sample not received and not analysed.

82604-008	ABC	A										
82604-009	ABC	A										
82606-001	ABC	A										
82608-002	ABC	A										
57801-002	ABC	A										

91 Footnotes: **A**, sample subjected to 16s rRNA gene sequencing and used in paired analysis to
92 determine microbiota concordance; **B**, sample used in qPCR analysis; **C**, 16S rRNA gene sequencing
93 results used to examine genus relative abundance, alpha and beta diversity metrics and for
94 identification of trends in the dataset; '-' sample not received and not analysed.

95 **Supplementary Table S3.** Statistical analysis of Shannon diversity at three paired start-end
 96 points for the Bcc and *P. aeruginosa* trials (Wilcoxon signed-rank test)

Treatment	Test statistic	P-value
Bcc trial		
Start-End	V = 31	P = 0.3396
OligoG Start-End	V = 38	P = 0.6355
Placebo Start-End	V = 42	P = 0.8394
<i>P. aeruginosa</i> trial		
Start-End	V = 410.5	P = 0.2293
OligoG Start-End	V = 541	P = 0.7971
Placebo Start-End	V = 498	P = 0.8318

97

98

99

100 **Supplementary Table S4.** Statistical analysis of the total abundance (qPCR) of key
 101 pathogens at three paired start-end points for the Bcc and *P. aeruginosa* trials (Wilcoxon
 102 signed-rank test)

Treatment	Test statistic	P-value
Bcc trial – <i>Burkholderia</i> total abundance		
Start-End	V = 18	P = 0.05737
OligoG Start-End	V = 51	P = 0.7354
Placebo Start-End	V = 78	P = 0.02148
<i>P. aeruginosa</i> trial – <i>P. aeruginosa</i> total abundance		
Start-End	V = 470.5	P = 0.8171
OligoG Start-End	V = 456.5	P = 0.7411
Placebo Start-End	V = 401	P = 0.7071

103

104 Footnotes: Statistically significant differences are highlighted in green.

105

106

107

108 **Supplementary Table S5.** Statistical analysis of the relative abundance of key pathogens at
 109 three paired start-end points for the Bcc and *P. aeruginosa* trials using GAMLSS-BEINF and
 110 (μ) logit links

Treatment	Estimate	Std. Error	t value	Pr (> t)
Bcc trial – <i>Burkholderia</i> relative abundance				
Start-End	0.5107	0.4409	1.158	0.267
OligoG Start-End	0.6461	0.2944	2.195	0.05589
Placebo Start-End	0.5859	0.3010	1.947	0.083261
<i>P. aeruginosa</i> trial – <i>P. aeruginosa</i> relative abundance				
Start-End	-0.5372	0.1656	-3.244	0.00226
OligoG Start-End	-0.1030	0.1729	-0.595	0.555
Placebo Start-End	0.2714	0.1757	1.545	0.129

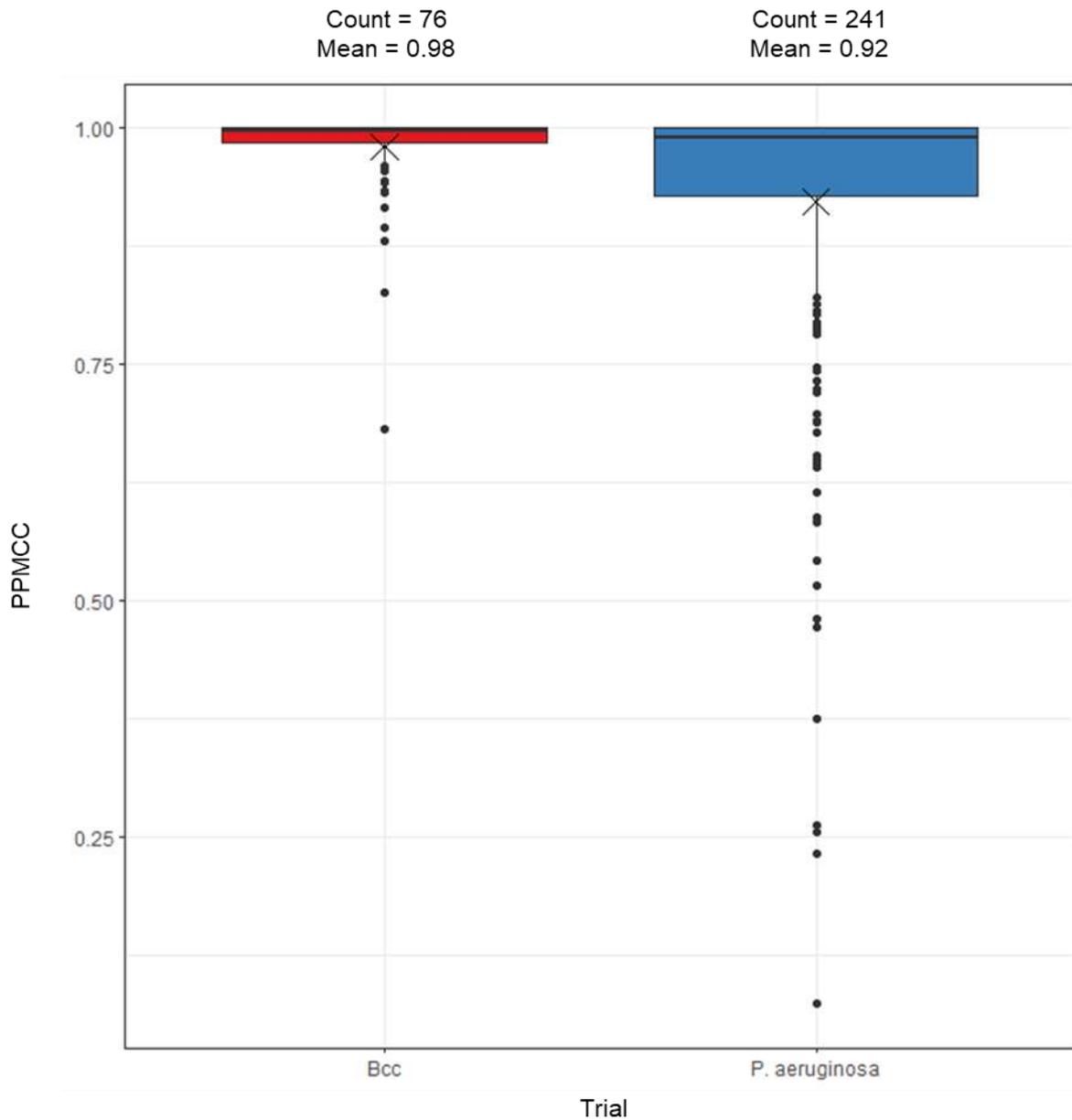
111

112 Footnotes: The estimates from the GAMLSS-BEINF models are the difference in log odds of relative abundances
 113 between groups. In each model, treatment was the response variable and patient ID was the random effect, with
 114 treatment 'start' as the reference class to which treatment 'end' was compared. Statistically significant differences
 115 are highlighted in green.

116 **Supplementary Table S6. The relative and total abundance of *Burkholderia***
 117 **in *Bcc* trial samples**
 118

Patient	<i>Burkholderia</i> species	Relative abundance (%)	Total abundance (Log Bcc/g sputum)
27610-002	<i>B. cenocepacia</i>	90.29	7.74
27610-005		99.62	8.55
27610-006		62.54	8.01
27611-002		66.52	7.36
27610-001		5.2	5.96
27610-004	<i>B. multivorans</i>	98.98	8.33
27610-007		8.41	7.50
27610-008		93.56	8.30
27610-009		79.65	6.96
27611-006		2.10	6.41
27610-003	Unknown	0.07	0.76
27610-011		0.00	5.44
27611-005		0.09	3.69

119
 120



121

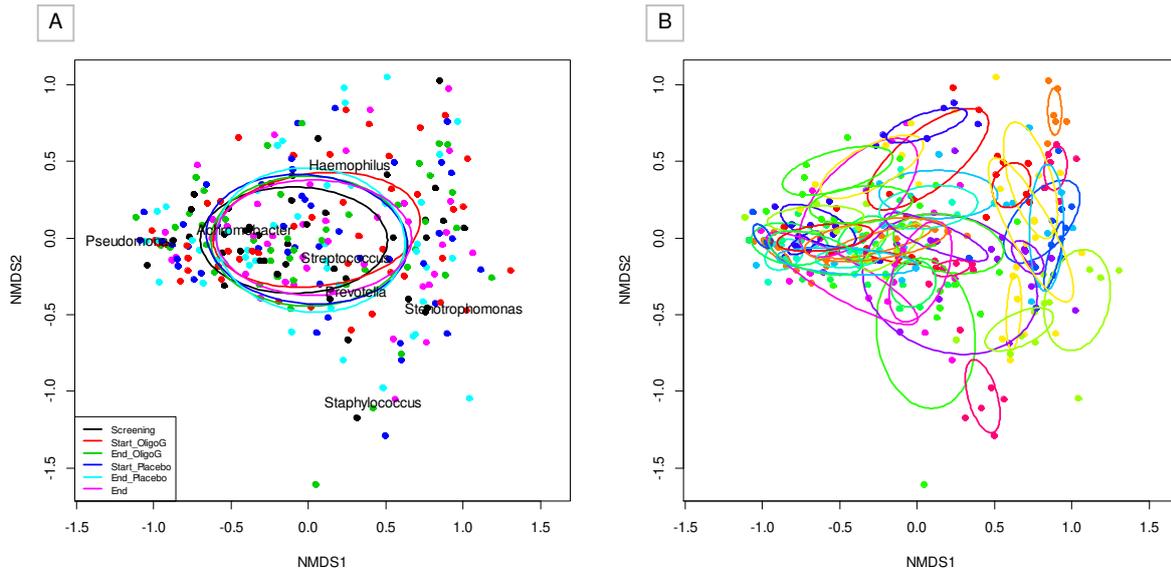
122

123

124

125 **Supplementary Figure S1. Paired samples from the Bcc and *P. aeruginosa* trials were**
 126 **concordant as shown by high Pearson product-moment correlation coefficient**
 127 **(PPMCC) values.** The boxplots show the spread of the PPMCC values for each trial, with the
 128 mean value highlighted by a cross in the centre of the plots. The number of samples in each
 129 trial and the mean values are shown above the plots.

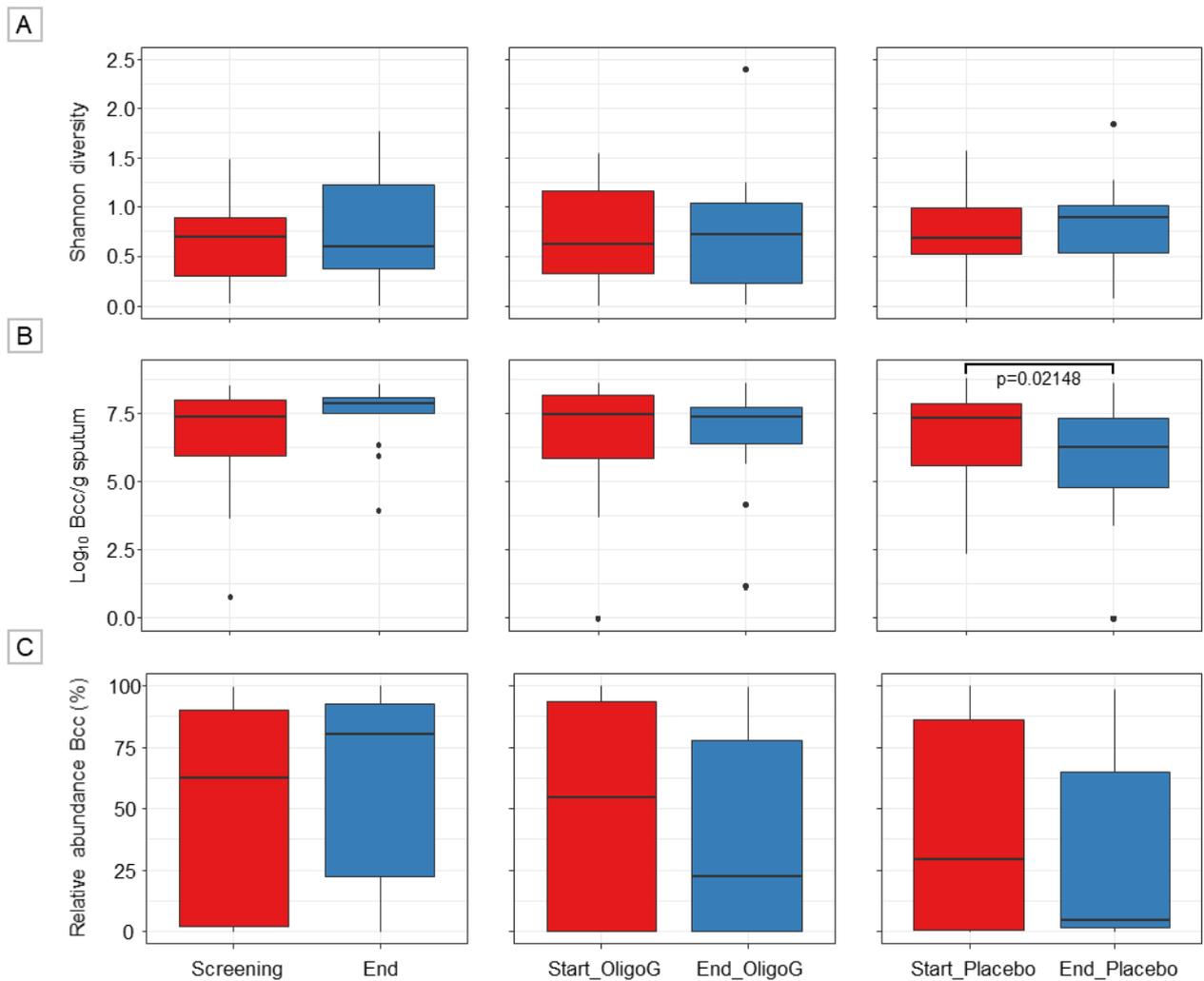
130



131
132

133 **Supplementary Figure S2. The lung microbiota in the *P. aeruginosa* trial was linked to**
 134 **the individual rather than treatment.** NMDS analysis of Bray-Curtis dissimilarity values for
 135 S1 samples from all 6 time points for the 45 patients on the *P. aeruginosa* trial (except
 136 82601003 that only had S2 for V1) are shown and grouped by treatment (**A**) and patient (**B**).
 137 Points represent individual samples, ellipses are standard deviations of points scores for each
 138 grouping. The top 10% genera based on abundance across the whole dataset are shown in
 139 (**A**). A significant difference was observed between patient groups (PERMANOVA, $R^2=0.01$
 140 $p=0.034$), but not between treatment groups (PERMANOVA, $R^2=0.006$, $p=0.998$). For the
 141 treatment sample grouping (**A**) the group variances were homogeneous, satisfying the
 142 conditions of the PERMANOVA model. This was not the case for the patient sample grouping
 143 (**B**), indicating that the significant result should be interpreted with caution.

144



145

146

147

148 **Supplementary Figure S3. Analysis of microbiota present between paired start and end**

149 **time-points collected during the Bcc trial.** Boxplots show the spread of data for Screening

150 versus End samples, Start OligoG versus End OligoG samples, and Start Placebo (S1 samples only, n=78; except 27610-011 that only had S2 for V7).

151 Placebo (S1 samples only, n=78; except 27610-011 that only had S2 for V7). **(A)** shows the

152 microbiota diversity measured using the Shannon index; **(B)** provides the total abundance of

153 *Burkholderia* per gram of sputum measured using qPCR, and; **(C)** shows the relative

154 abundance of *Burkholderia* from 16S rRNA gene sequencing analysis. For Shannon Diversity

155 and total *Burkholderia* abundance, Wilcoxon signed-rank tests were used to assess the

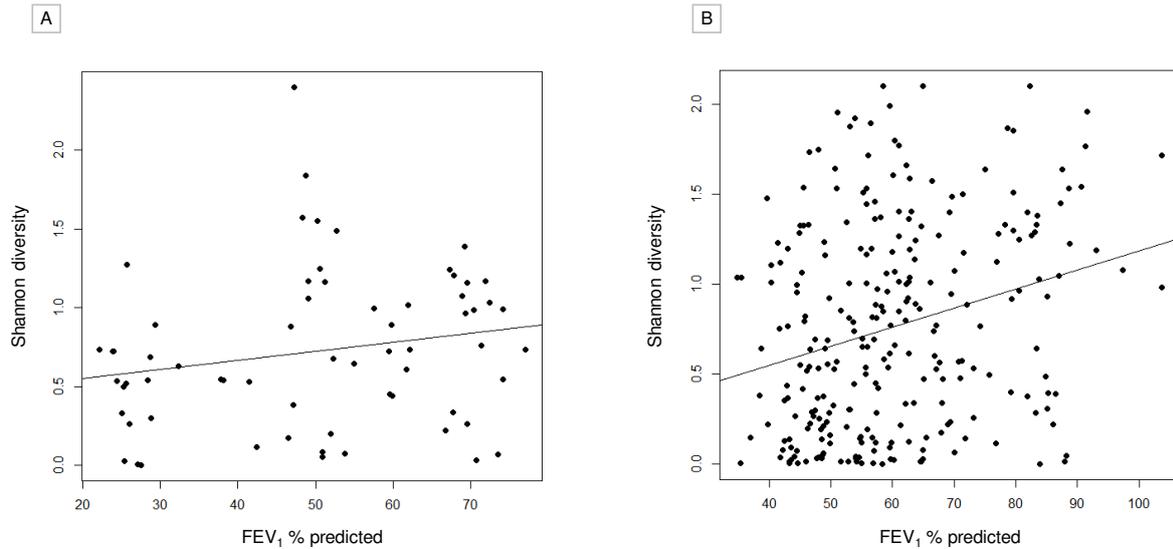
156 differences between paired time-points. Differences in relative abundance were determined

157 using GAMLESS-BEINF models with patient as the random effect, reporting changes in

158 log(odds ratio) between paired time-points. Statistical significance is shown as a bracket

159 above boxplots with the p-value under the bracket.

160



161

162

163 **Supplementary Figure S4. Correlation of loss of bacterial diversity with poor lung**
 164 **function for the Oligo G trial participants.** Linear regression was used to examine the
 165 relationship between Shannon diversity and lung function (FEV₁ % predicted) for: **(A)** S1
 166 samples (n=78) from all 6 time points for the 13 patients on the Bcc trial (except 27610-011
 167 that only had S2 for V7), and **(B)** S1 samples (n=270) from all 6 time points for the 45 patients
 168 on the *P. aeruginosa* trial (except 82601-003 that only had S2 for V1). The regression lines
 169 show a trend for decreased lung function with decreased diversity. These trends were not
 170 significant when analysed with linear mixed models with patient as the random effect.

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187 **Supplementary method S1**

188 Standardised protocol 1: Sputum sample processing for DNA extraction

189

190 **A) Sample pre-processing**

191

192 1. Sputum samples must be processed within 4 weeks of sample collection (stored at
193 frozen at -80°C in the meantime)

194 2. Thaw samples at room temperature for 30-45 minutes

195

196 *Label tubes with a Cardiff number and record all sample information*

197

198 3. Weigh samples using an empty collection tube to tare balance. Record the weight to
199 nearest mg and write on side of tube

200

201 *Record sample weights*

202

203 4. Add 4M guanidine isothiocyanate to sample (1:1 ratio, add equivalent ml to weight
204 e.g. if 1 g sample, add 1 ml of guanidine isothiocyanate)

205 5. Centrifuge tubes to bring sputum to the bottom (2 minutes; 1409 g)

206 6. Vortex mix tubes for 30 seconds to 1 minute

207 *If the sputum is still very viscous incubation at 37°C for 30 minutes to 1 hour can be
208 performed to thin the consistency*

209

210 **B) Bead beating**

211 1. Add 1ml of the sputum mix to 1 g of beads in a 2 ml tube with cap and O-ring

212

213 *if the sputum is viscous it will be difficult to measure 1 ml accurately, in this case just
214 pipette enough sputum to reach the fill line indicated below*

215

216

217  Fill line

218

219

220

221 2. Bead beat for 2 minutes on lowest setting of Beadbug (280 x 10 rpm)

222

223 3. Pulse centrifuge to settle beads

224 **C) DNA extraction**

225 1. Add 400 µl of sputum mix to a Maxwell16® tissue kit cartridge and run DNA
226 extraction programme on Maxwell16® instrument. In one run, 16 samples can be
227 processed.

228

229 *Store remaining sputum mix at -80°C in 1.5 ml non-stick eppendorfs*

230

231 2. Collect DNA from Maxwell16® instrument into 1.5 ml non-stick eppendorfs

- 232 3. Store 20 µl DNA at 4°C for PCR analyses (short term storage)
233 4. Store remaining DNA at -20°C (long term storage)

234

235 **D) Reagents/consumables/equipment**

Reagent/Consumable/Equipment	Supplier	Product Code
Beadbug microtube homogenizer	Benchmark Scientific	D1030
Triple-pure high impact 100 µm zirconium beads	Benchmark Scientific	D1132-01TP
2 ml bead tubes with caps and seals	Benchmark Scientific	D1031-T20
4 M UltraPure™ Guanidine Isothiocyanate Solution	ThermoFisher Scientific	15577018
Maxwell® 16 System for DNA extraction	Promega	AS1250
Maxwell® Tissue DNA Purification kit (48 preps)	Promega	AS1030

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257 **Supplementary method S2**

258 Standardised protocol 2: Identification of *Burkholderia* species in sputum samples

259

260 **A) PCR amplification of *recA* and *gyrB* gene sequences**

261 Perform separate PCRs for the amplification of *Burkholderia recA* and *gyrB* genes using DNA
262 extracted from sputum samples (Standardised protocol 1).

263

264 1. **Order PCR primers** (Spilker *et al.* 2009) from Eurofins Genomics. Primer sequences are
265 shown in Table 1. Prepare stock solutions of individual primers at 100 pmol/μl in nuclease
266 free water according to the synthesis report provided by Eurofins Genomics. Prepare a
267 working solution of a combination of F and R primers at 10 pmol/μl each in nuclease free
268 water (for example 30 μl of F primer and 30 μl R primer in a total volume of 300 μl).

269

270 2. **Prepare PCR Mastermix** for the number of reactions required. This will include the
271 number of samples, a positive and negative control, and one additional reaction to account
272 for any pipetting error. Reagent concentrations and volumes for 1 reaction (50 μl) are
273 shown in Table 2.

274

275 3. **Run the PCRs** in a thermal cycler with the reaction conditions shown in Table 3. Note that
276 different annealing temperatures are used for *recA* and *gyrB*.

277

278 4. **Perform gel electrophoresis** to check for successful amplification of gene products.
279 Prepare a 1.5 % (w/v) gel using molecular grade agarose and Tris-acetate-EDTA (TAE)
280 buffer, stained with SafeView (NBS Biologicals Ltd.; 10 μl SafeView per 100 ml of agarose
281 gel). Load 5-10 μl of PCR product and run in TAE buffer at 80V for approximately 1 hour.
282 Visualise PCR products with a gel imaging system.

283

284 **B) PCR product purification and sequencing**

285 1. **Purify PCR products** using the QIAquick PCR purification kit (Qiagen Ltd.) according to
286 the manufacturer's instructions. Load the remaining volume from the PCR for each sample
287 (approximately 40 – 45 μl) into the kit.

288 2. **Quantify PCR products** using the Qubit fluorometer and the Qubit dsDNA BR Assay kit
289 (ThermoFisher Scientific) according to the manufacturer's instructions.

290

291 3. **Send the PCR products for sequencing** with the eurofins genomics sequencing service.
292 Note that PCR products may need to be diluted to a certain concentration in nuclease free
293 water before sending, and primers will need to be sent with the PCR products. For both
294 *recA* and *gyrB* send F and R primers to obtain Forward and Reverse sequences for each
295 sample.

296

297

298

299 **C) Analysis of *recA* and *gyrB* sequences**

300

301 1. **Create consensus *recA* and *gyrB* gene sequences** from the F and R sequence reads.
 302 Eurofins genomics will email the gene sequences after sequencing has been completed.
 303 Copy and paste the F and R sequences into Notepad and save as .fasta files. Open the
 304 .fasta files with the programme BioEdit (Hall 1999).

305

306 • Highlight the R sequences and select reverse complement from the Sequence tab
 307 (Sequence>Nucleic acid>Reverse Complement)

308 • Highlight a pair of F and R sequences and create a pairwise alignment (allow ends to slide)
 309 of the F and reverse complemented R sequences (Sequence>Pairwise alignment>Align
 310 two sequences [allow ends to slide])

311 • Highlight the F and R sequences and create a consensus sequence (Alignment>Create
 312 Consensus Sequence) and save as a .fasta file (File>Export>Split to Individual Fasta
 313 Files).

314

315 2. **Determine the *Burkholderia* species identity** using the BLAST tool on the *Burkholderia*
 316 Genome Database (www.burkholderia.com) (Winsor *et al.* 2008). Search the database
 317 using the BLASTN tool and the *recA* and *gyrB* consensus sequences. Include all of the
 318 available complete genome sequences in the database in the search. The top database
 319 hits will allow the determination of the species identity.

320

Gene	Gene product	Primers	Sequence 5'>3'	Product size (bp)
<i>recA</i>	Recombinase A	F	AGGACGATTCATGGAAGAWAGC	704
		R	GACGCACYGAYGMRTAGAACTT	
<i>gyrB</i>	DNA gyrase B	F	ACCGGTCTGCAYCACCTCGT	738
		R	YTCGTTGWARCTGTCGTTCCACTGC	

321 **Table 1.** PCR primers for the amplification of *recA* and *gyrB* genes

322

323 All primers are synthesised and supplied by Eurofins Genomics

324

Reagent	Final concentration (50 µl)	Volume 1 reaction (µl)
H ₂ O	-	21.6
PCR buffer (10X)	x1	5.0
Q-solution (5X)	x1	10.0
Primers (10 pmol/ul)	1.6 uM	8.0
dNTPs (10 mM each)	240 µM (each)	1.0
Taq	2U	0.4
DNA	-	4

Table 2. PCR reagents for the amplification of *recA* and *gyrB* genes

331

332

333

334

335

336 All reagents are supplied by Qiagen, with the exception of nuclease free H₂O which is
337 supplied by Severn Biotech Ltd.

PCR step	Temperature (°C)	Time (mins)	Cycles
Initial denaturation	95	2	1
Denaturation	94	0.5	
Annealing	60 (<i>gyrB</i>) 58 (<i>recA</i>)	0.5	30
Extension	72	1	
Final Extension	72	5	1
Indefinite hold	10	-	-

Table 3. PCR reaction conditions for the amplification of *recA* and

343 *gyrB* genes

344

345

346

347

348

349

350 **D) PCR reagents and consumables**

351

Reagent/Consumable	Supplier	Product Code
Taq DNA polymerase (Master Mix kit)	Qiagen	201203
Molecular grade water	Severn Biotech Ltd.	20-9000-01
QIAquick PCR purification kit	Qiagen	28104
Qubit dsDNA BR Assay kit	Invitrogen	Q32853
Qubit Assay tubes	Invitrogen	Q32856

352

353

354 **References**

355 Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis
356 program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**:95-98.

357 Spilker, T., Baldwin, A., Bumford, A., Dowson, C. G., Mahenthalingam, E. and LiPuma, J. J.
358 (2009). Expanded multilocus sequence typing for *Burkholderia* species. *J Clin Microbiol*
359 **47**:2607-2610.

360 Winsor, G. L., Khaira, B., Van Rossum, T., Lo, R., Whiteside, M. D. and Brinkman, F. S. L.
361 (2008). The *Burkholderia* Genome Database: facilitating flexible queries and comparative
362 analyses. *Bioinformatics* **24**:2803-2804.

363

364

365

366

367

368 **Supplementary method S3**

369 Standardised protocol 3: Quantitative PCR (qPCR) using TaqMan probes to determine
370 bacterial load (*Pseudomonas aeruginosa* and *Burkholderia*)

371

372 For qPCR background information please refer to the life technologies Real-time PCR
373 handbook, which can be found here: [http://www.gene-quantification.com/real-time-pcr-](http://www.gene-quantification.com/real-time-pcr-handbook-life-technologies-update-flr.pdf)
374 [handbook-life-technologies-update-flr.pdf](http://www.gene-quantification.com/real-time-pcr-handbook-life-technologies-update-flr.pdf). This protocol uses the Chromo4™ system for real-
375 time detection and the Opticon Monitor Software for analysis. The operation manual can be
376 found here: http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_10498.pdf. The
377 *gyrB* gene is used for the quantification of *P. aeruginosa* and the *rpoD* gene is used for
378 *Burkholderia*.

379

380 **A) qPCR standards**

381

382 • Reaction efficiency and replicate reproducibility are best assessed through the
383 generation of a standard curve. This is based on a dilution series of the sample nucleic
384 acid which is included in the qPCR run.

385 • This protocol describes the use of PCR products as standards for qPCR. For this, a
386 DNA fragment larger (>500 bp) than the target qPCR product (generally 100-200 bp)
387 is amplified by PCR, purified and quantified to determine the concentration (and copy
388 number) of the target qPCR product per µl of DNA.

389

390 **1. Prepare a high concentration 'stock' of the qPCR standard for each gene target**

391 ○ PCR amplify the fragment of DNA containing the target qPCR gene. The PCR
392 reagents, primers and reaction conditions for the amplification of the *rpoD* and *gyrB*
393 genes are shown in Tables 1, 2 and 3. To obtain a large volume of the DNA fragment,
394 run 3 x 50 µl reactions and include a negative control. The species used as positive
395 controls are shown in Table 1.

396 ○ Perform standard gel electrophoresis as described in Standardised protocol 2. Load
397 5 µl of each PCR product.

398 ○ Pool the remaining PCR products (45 µl left in each tube, giving a total of
399 approximately 135 µl) and purify using the QIAquick PCR purification kit (Qiagen Ltd.)
400 according to the manufacturer's instructions.

401 ○ Quantify PCR products using the Qubit fluorometer (Invitrogen) and Qubit dsDNA BR
402 Assay kit (Invitrogen) according to the manufacturer's instructions to obtain a
403 concentration in ng/µl

404 ○ See section E for full details of suppliers of PCR and PCR clean-up and QC reagents

405 ○ Calculate the copy number of the target gene using the following equations:

- I. $(\text{qPCR target bp}/\text{DNA fragment bp}) \times 100 = \% \text{ of DNA that is qPCR target}$
 II. $\text{ng}/\mu\text{l DNA fragment} \times \% \text{ of DNA that qPCR target} = \text{ng}/\mu\text{l qPCR target}$
 III. $[6.023 \times 10^{23} (\text{copies/mol}) \times \text{g}/\mu\text{l qPCR target}] /$
 $\text{qPCR target bp} \times \text{DNA molecular weight} = \text{qPCR target copies}/\mu\text{l}$

Equivalent to:

$$[6.023 \times 10^{14} (\text{Da/ng}) \times \text{ng}/\mu\text{l qPCR target}] /$$

$$\text{qPCR target bp} \times 660 (\text{Da/bp}) = \text{qPCR target copies}/\mu\text{l}$$

406

- 407 ○ A concentration of approximately 10^{10} - 10^{11} copies/ μl should be obtained from a
 408 successful PCR and purification. Store the stock at -20°C .

409

2. Prepare a dilution series

- 411 ○ Use nuclease free water to prepare a dilution series from the stock to cover 10^2 – 10^8
 412 copies/ μl
 413 ○ The dilution series should be made up on the day of the qPCR and stored at -20°C
 414 for no longer than 24 hours before being discarded. Ideally a fresh dilution series
 415 should be made for each qPCR run. Using the Qubit fluorometer, quantify the 10^{10}
 416 copies/ μl dilution of each dilution series to accurately calculate the copy number for
 417 the lower dilutions.

418

419 **Table 1.** PCR primers for the amplification of qPCR standards

Gene	Primers	Sequence 5'>3'	Product size (bp)	Positive control species
<i>rpoD</i>	F	GATCTTGACATCGTCGTC	1011	<i>Burkholderia cenocepacia</i> (J2315)
	R	GTTCGTAACGGAGACGCTG		
<i>gyrB</i>	F	GAGTCGATCACTGTCCGC	1186	<i>Pseudomonas aeruginosa</i> (PAO1)
	R	GCATCTTGTCGAAGCGCG		

420 All primers are synthesised and supplied by Eurofins Genomics; refer to Standardised
 421 protocol 2 for preparation of primer stock solutions.

422

423 **Table 2.** PCR reagents for the amplification of qPCR standards

Reagent	Final concentration	Volume 1 reaction (μl)	424
H ₂ O	-	21.6	425
PCR buffer (10X)	x1	5.0	
Q-solution (5X)	x1	10.0	426
Primers (10 pmol/ μl)	1.6 μM	8.0	
dNTPs (10 mM each)	240 μM (each)	1.0	427
Taq	2U	0.4	
DNA	-	4	428

429 All reagents are supplied by Qiagen, with the exception of nuclease free H₂O which is
 430 supplied by Severn Biotech Ltd. Refer to Standardised protocol 2 for preparation of PCR
 431 Mastermix.

432

433

434 **Table 3.** PCR reaction conditions for the amplification of qPCR standards

PCR step	Temperature (°C)	Time (mins)	Cycles
Initial denaturation	95	2	1
Denaturation	94	0.5	30
Annealing	59 (<i>rpoD</i>) 58 (<i>gyrB</i>)	0.5	
Extension	72	1	
Final Extension	72	5	1
Indefinite hold	10	-	-

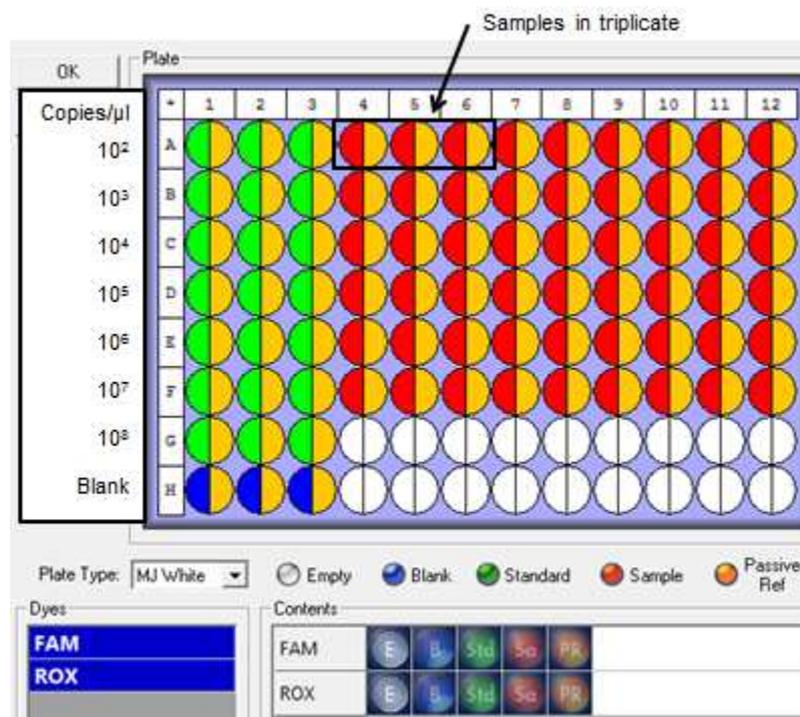
440 **B) qPCR**

- 441
- 442 1. Set up the Chromo4™ real-time detector. Using the Opticon Monitor Software, edit the
- 443 plate layout and reaction conditions for the qPCR run. An example 96 well-plate layout
- 444 is shown below in Figure 1. Under Master>Plate Setup>Edit>Specify Quant Standards,
- 445 the copies/μl of each of the standards can be entered. Under Master>Plate Setup
- 446 Setup>Edit>Sets, replicates can be grouped together into sets and given an ID. The
- 447 reaction conditions for the 16S rRNA, *rpoD* and *gyrB* gene qPCRs are shown in Table
- 448 4.

449

450

451



452

453

454 **Figure 1.** Example qPCR plate layout on the Opticon Monitor Software. Standards,

455 standards, samples and blanks are run in triplicate. The passive reference (PR) dye ROX is a

456 component of the qPCR Mastermix and this is highlighted for each well.

457

458

459

460

461

462

Table 4. qPCR reaction conditions for the amplification of *rpoD* and *gyrB* genes

Step	Time	Temp (°C)	N° cycles
UNG treatment	3 min	50	1
Taq activation	10 min	95	1
Denaturation	15 secs	95	
Annealing and Extension	30 secs (<i>rpoD</i>)	67	40
	30 secs (<i>gyrB</i>)	60	
Plate Read			
Hold	10 mins	25	1

463

464

465

466

467

468

469

470

471

472

473

474

475

476

2. Prepare the qPCR Mastermix for the number of reactions required, bearing in mind that everything is performed in triplicate. This will include the 10² – 10⁸ dilution series, the samples and the blanks. Add an extra reaction to account for any pipetting error. The qPCR primers and TaqMan probes, and reagents are shown in Tables 5 and 6, respectively.
3. Load the Mastermix into the qPCR plate, 9 µl into each well.
4. Load the standards, blanks and samples into the qPCR plate, 1 µl into each well.
5. Seal the plate with a plate seal and pulse centrifuge the plate to draw the liquid to the bottom of the wells
6. Transfer the plate into the qPCR machine and run the qPCR. The qPCR plate can be discarded once the run is complete.

477

Table 5. qPCR primers and TaqMan probes for the amplification of *rpoD* and *gyrB* genes

Gene	Primers/Probe	Sequence 5'>3'	Product size (bp)	Reference
<i>rpoD</i>	F	GAGATGAGCACCGATCACAC	143	(Sass <i>et al.</i> 2013)
	R	CCTTCGAGGAACGACTTCAG		This study
	PROBE	5'FAM-CTGCGCAAGCTGCGTCACC-3'MGBNFQ		
<i>gyrB</i>	F	CCTGACCATCCGTCGCCACAAC	220	(Anuj <i>et al.</i> 2009)
	R	CGCAGCAGGATGCCGACGCC		
	PROBE	5'FAM-GGTCTGGGAACAGGTCTACCACCACGG-3'MGBNFQ		

478

479

480

Table 6. qPCR Mastermix: reagents for the amplification of 16S rRNA, *rpoD* and *gyrB* genes

Reagent	Final concentration	Volume 1 reaction (µl)	481
qPCR Supermix with ROX (2X)*	1X	5	482
F&R primers (10 pmol/µl)	1.8 µM	1.8	
TaqMan Probe (100 pmol/µl) [FAM]*	225 nM	0.0225	483
H ₂ O	-	2.1775	484

485

*See Section E for the suppliers of the qPCR Supermix and TaqMan Probes

486

487

488

489

490 **C. Analysis of qPCR results in Opticon Monitor Software**

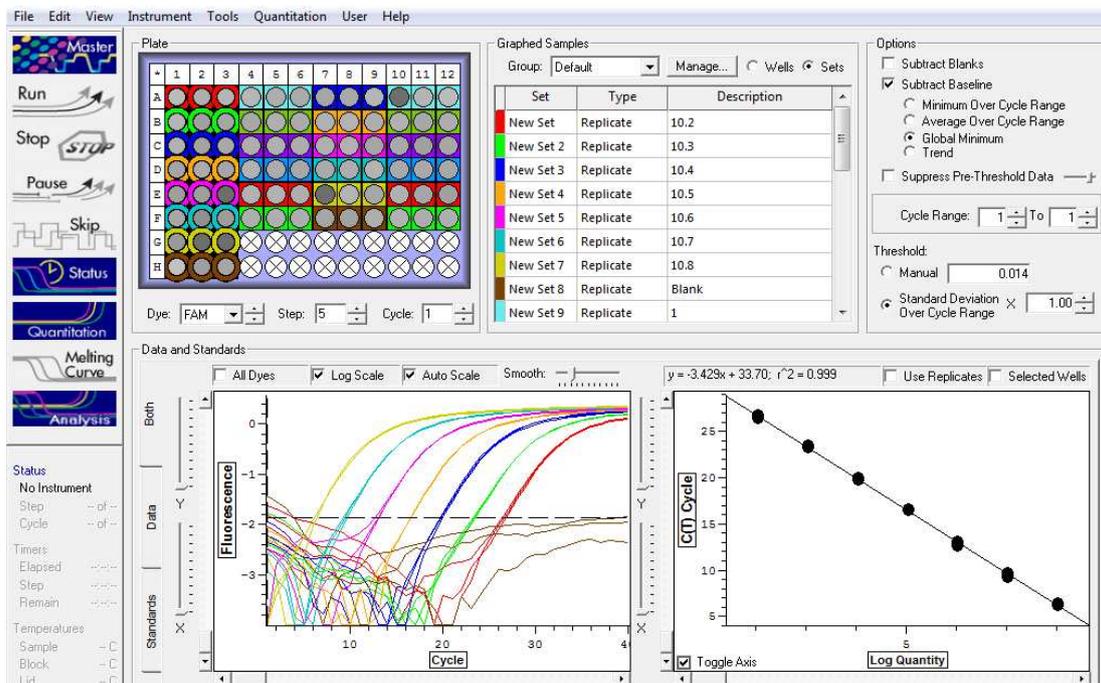
- 491 • In the Opticon Monitor Software click the 'Quantitation' view to see the results. This
- 492 view will show you the plate layout, wells and sets, and the data and standards graphs.
- 493 An example is shown in Figure 2.
- 494 • To calculate the efficiency of the qPCR amplification, use the slope of the standard
- 495 curve and the below equation. The 'Toggle Axis' box needs to be checked to get the
- 496 correct x value to use in the equation. The acceptable range is 90 -110%.

497 **qPCR reaction efficiency = $[10^{(-1/\text{slope of standard curve})} - 1] \times 100$**

498 **qPCR reaction efficiency (Figure 2) = $[10^{(-1/-3.429)} - 1] \times 100 = 95.7\%$**

- 499 • The R² value is a measure of replicate reproducibility and should ideally be above 0.98
- 500 (98%). The R² value from Figure 2 is 0.999 (99.9%)
- 501 • If you are satisfied with the efficiency and R² values move on to processing the data

502
503



504

505 **Figure 2.** qPCR results under the 'Quantitation' view in Opticon Monitor Software. The data
 506 graph in the bottom left hand window displays the fluorescence curves of the standards, the
 507 bottom right hand window displays the standard curve from which the efficiency and R² values
 508 can be obtained.

509

- 510 • To view the results from all of the wells, highlight all of the wells in the 'Plate' section,
- 511 and all of the sets in the 'Graphed Samples' section. Select Quantitation>Copy to
- 512 Clipboard>Quantity calculations, and paste this into an Excel spreadsheet.
- 513 • Look at the triplicate results for each sample and calculate the coefficient of variation
- 514 (CV = standard deviation/mean). If the CV is equal to or less than 20% then the mean
- 515 of the three results can be taken. If the CV is higher than 20% examine the results

516 further. If two results within a triplicate set are within 0.2 log of each other, the mean of
 517 these two results is taken, and the third 'anomalous' result is excluded from the
 518 dataset. If assays do not meet these criteria, they are considered unacceptable and
 519 the qPCR re-run for the sample. A full explanation of these criteria can be found in
 520 (Zemanick *et al.* 2010).

- 521 • Opticon Monitor Software will perform preliminary analysis of replicate sets for you,
 522 allowing you to observe the average C(t), the average copy number, and some basic
 523 statistics for each set of replicates. Remove wells with anomalous results from the sets
 524 under the Master>Plate Setup>Edit>Sets, return to the Quantitation view, highlight all
 525 of the sets in the Graphed Samples section, and select Quantitation>Copy to
 526 Clipboard>Quantity calculations, and paste this into an Excel spreadsheet.
 527

528 **D. Statistical analysis of qPCR results**

- 529 • There should be at least 3 qPCR runs (3 biological replicates) per sample. These
 530 results can be used for further statistical analyses.
- 531 • The results obtained are gene copy per µl of DNA (equivalent to cell number) extracted
 532 from a sputum sample. Perform the necessary calculations to obtain the number of
 533 gene copies (cell number) per gram of sputum for comparisons.

534

535 **E. Reagents and consumables**

536 **Table 7.** qPCR Reagents and consumables

Reagent/Consumable	Supplier	Product Code
Preparation of qPCR standards		
Taq DNA polymerase (Master Mix kit)	Qiagen	201203
Molecular grade water	Severn Biotech Ltd.	20-9000-01
QIAquick PCR purification kit	Qiagen	28104
Qubit dsDNA BR Assay kit	Invitrogen	Q32853
Qubit Assay tubes	Invitrogen	Q32856
qPCR		
Platinum qPCR supermix-UDG w/Rox	Life Technologies (Thermo Fisher Scientific)	11743-500
Custom TaqMan probe (5'FAM, 3'MGBNFQ)	ThermoFisher Scientific	Order online: https://www.thermofisher.com/order/custom-oligo/custom-taqman-probes
Hard-Shell low-profile Thin-wall 96-well skirted PCR plates (black shell white well)	BioRad	HSP9665
Sealing film for real time PCR (50 µm thick)	ELKAY	SEA-LPTS-RT2

537

538

539

540 **References**

541

542 Anuj, S. N., Whiley, D. M., Kidd, T. J., Bell, S. C., Wainwright, C. E., Nissen, M. D. and Sloots,
543 T. P. (2009). Identification of *Pseudomonas aeruginosa* by a duplex real-time polymerase
544 chain reaction assay targeting the *ecfX* and the *gyrB* genes. *Diagnostic Microbiology and*
545 *Infectious Disease* **63**:127-131.

546

547 Sass, A. M., Schmerk, C., Agnoli, K., Norville, P. J., Eberl, L., Valvano, M. A. and
548 Mahenthiralingam, E. (2013). The unexpected discovery of a novel low-oxygen-activated
549 locus for the anoxic persistence of *Burkholderia cenocepacia*. *Isme j* **7**:1568-1581.

550

551 Zemanick, E. T., Wagner, B. D., Sagel, S. D., Stevens, M. J., Accurso, F. J. and Harris, J. K.
552 (2010). Reliability of quantitative real-time PCR for bacterial detection in cystic fibrosis airway
553 specimens. *PLoS One* **5**.

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576 **Supplementary method S4**

577 Standardised protocol 4: Ribosomal RNA Intergenic Spacer Analysis (RISA)

578
579 The region between the small and large subunit rRNA genes, the 16S rRNA and 23S rRNA
580 gene for bacteria, is known as the Intergenic Transcribed Spacer (ITS). This region varies in
581 length and sequence and is the basis of RISA (Fisher and Triplett 1999). RISA profiles provide
582 the following information about bacterial diversity in polymicrobial samples: (i) the number of
583 RISA bands is a qualitative measure of the taxonomic diversity present; (ii) the size of the ITS
584 amplicon may be correlated to a known species/genera, especially if run alongside a positive
585 control, and provide a presumptive taxonomic identification; and (iii) the RISA profiles are
586 semi-quantitative, with the ITS band intensity correlating to the amount of DNA template and
587 hence the approximate proportion of the original organism present in the sputum sample.

588
589 This protocol describes the use of RISA to check quality of the DNA extracted from sputum
590 (Standardised protocol 1), i.e. to determine if there is sufficient bacterial DNA for a positive
591 PCR result, and to get a preliminary idea of the bacterial diversity in a sputum sample.

592

593

594 **A) RISA-PCR**

595

596 5. RISA-PCR primers are shown in Table 1, please refer to Standardised protocol 2 for more
597 details on primer stock preparation.

598 6. Prepare the RISA-PCR Mastermix for the number of reactions required. This will include
599 the number of samples, a positive and negative control, and one additional reaction to
600 account for any pipetting error. PCR primers, reagent concentrations and volumes for 1
601 reaction (25 µl) are shown in Table 2. Control DNA from *Pseudomonas aeruginosa*,
602 *Burkholderia spp.* and other bacterial species linked with cystic fibrosis can also be run to
603 produce control products for profile analysis.

604 7. Run the RISA-PCR in a thermal cycler with the PCR conditions shown in Table 3.

605 8. Perform standard gel electrophoresis to check for successful amplification of PCR
606 products

607

608

609 **Table 1.** PCR primers for the amplification of the bacterial ITS region

610

Primer	Sequence (5' > 3')	Comment	Reference
1406F	TGYACACACCGCCCGT	Degenerate RISA	Fisher & Triplett
23SR	GGGTTBCCCCATTCTRG	primers	(1999)

611 All primers are synthesised and supplied by Eurofins Genomics

612

613

614

615

616

617

618 **Table 2.** PCR reagents for the amplification of the bacterial ITS region
 619

Reagent	Final concentration (25 ul)	Volume	1 reaction (ul)
H ₂ O	-	13.8	621
PCR buffer (10X)	x1	2.5	622
Q-solution (5X)	x1	5.0	
Primers (10 pmol/ul)	0.4 uM	1.0	623
dNTPs (10 mM each)	240 um (each)	0.5	624
Taq	2U	0.2	625
DNA	-	2	626

627 All reagents are supplied by Qiagen, with the exception of nuclease free H₂O which is
 628 supplied by Severn Biotech Ltd.

629

630 **Table 3.** PCR reaction conditions for the amplification of the bacterial ITS region

PCR step	Temperature (°C)	Time (mins)	Cycles
Initial denaturation	95	5	1
Denaturation	95	1	
Annealing	54	0.5	35
Extension	72	1	
Final Extension	72	5	1
Indefinite hold	10	-	-

637

638

639 **B) Microfluidic separation and cluster analysis of RISA-PCR profiles**

640

- 641 1. Run RISA-PCR products on an Agilent BioAnalyzer (Agilent Technologies UK Ltd.,
 642 Cheshire, United Kingdom) using a DNA 7500 chip according to the manufacturer's
 643 instructions.
- 644 2. Import BioAnalyzer profiles into Bionumerics software (Applied Maths, Gent, Belgium) for
 645 analysis. A dedicated script has been provided to Cardiff University by Applied Maths to
 646 convert BioAnalyzer profiles to a format compatible with Bionumerics.
- 647 3. Cluster analysis is performed in Bionumerics to compare the RISA-PCR profiles.
 648 Similarities between RISA are calculated using the Pearson coefficient, and dendrograms
 649 constructed by the unweighted-pair group method using average linkages (UPGMA). It is
 650 also possible to view the profiles as a composite image to look at the patient samples in
 651 chronological order. Changes in overall diversity can be observed by looking at the number
 652 and sizes of bands within the profile.
- 653 4. The size of specific RISA amplicons can be determined using the BioAnalyzer and
 654 correlated to specific bacterial genera using the In Silico PCR database (website
 655 <http://insilico.ehu.es/PCR>)(Bikandi *et al.* 2004). The intergenic spacer sizes can be
 656 recorded for strains and species of interest (examples are shown in Table 4). It is not

657 always possible to identify bands within the RISA profile as certain species, but bands of
658 certain sizes can be useful markers to look out for.

659

660 **Table 4.** Intergenic transcribed spacer (ITS) sizes (bp) for a selection of bacterial species
661 and strains associated with cystic fibrosis

Species/strain	ITS size(s) (bp)
<i>Pseudomonas aeruginosa</i>	753
<i>Burkholderia ambifaria</i> AMMD	681, 827, 830, 871
<i>Burkholderia cenocepacia</i> J2315	677, 812, 816
<i>Burkholderia multivorans</i> ATCC 17616	829, 830, 883
<i>Haemophilus influenzae</i>	995-1015, 757-759
<i>Ralstonia mannitolytica</i>	805
<i>Ralstonia pickettii</i>	796-797
<i>Stenotrophomonas maltophilia</i>	777-782, 826-829
<i>Achromobacter xylosoxidans</i>	887-891, 1021

663

664

665 C) Reagents/consumables/equipment

Reagent/Consumable	Supplier	Product Code
Taq DNA polymerase (Master Mix kit)	Qiagen	201203
Molecular grade water	Severn Biotech Ltd.	20-9000-01
2100 Bioanalyzer instrument	Agilent	G2939B
Agilent DNA 7500 kit	Agilent	5067-1506
Agilent 7500 reagents	Agilent	5067-1507

666

667

668 References

669 Bikandi, J., San Millán, R., Rementeria, A. and Garaizar, J. (2004). In silico analysis of
670 complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction. *Bioinformatics*
671 **20**:798-799.

672 Fisher, M. M. and Triplett, E. W. (1999). Automated approach for ribosomal intergenic spacer
673 analysis of microbial diversity and its application to freshwater bacterial communities. *Applied*
674 *and Environmental Microbiology* **65**:4630-4636.

675

676

677

678

679 \

680

681 **Supplementary method S5**

682 Standardised protocol 5: Bacterial diversity analysis (16S rRNA gene sequencing and
683 analysis)

684

685 A) **Extract DNA** from the samples to be sent for 16S rRNA sequencing and bacterial diversity
686 analysis (Standardised protocol 1).

687

688 B) **Check DNA quality** by performing a RISA PCR (Standardised protocol 3). This will
689 determine whether there is sufficient DNA for a positive PCR result and will also give some
690 preliminary information about bacterial diversity.

691

692 C) **Sample submission** to the commercial company Research and Testing Laboratory Inc.
693 (Lubbock, Texas, USA; <http://www.researchandtesting.com>). Sample submission
694 guidelines can be found online and a submission spreadsheet will need to be completed
695 and sent back. Samples are sent via a courier to the company. Research and Testing
696 perform Illumina MiSeq sequencing of the 16S rRNA gene V1-V2 regions and on
697 completion will email a web link which can be used to retrieve the data.

698

699 D) **Data processing and analysis in Mothur (version 1.33)**

700 Mothur is a bioinformatics program for analysing microbial communities that is being
701 developed by Schloss lab at the University of Michigan. This protocol has been adapted from
702 the MiSeq SOP found at http://www.mothur.org/wiki/MiSeq_SOP#OTU-based_analysis to
703 process 16S rRNA gene sequences that are generated using Illumina's MiSeq platform using
704 paired end reads. Please refer to this website for a more detailed description of the pipeline
705 steps.

706 **1. Preparing the files for Mothur**

707 Based on Illumina's MiSeq platform, there are two fastq files generated for each sample. The
708 R1 file corresponds to read 1 (forward read) and the R2 file corresponds to read 2 (reverse
709 read). It may be useful to rename these files (e.g. **example_R1**, **example_R2**) before starting
710 the Mothur pipeline as the original file names can be very long.

711 These fastq files then need to be paired up for each sample. Using Microsoft Excel in
712 windows, make a spreadsheet with 3 columns but no headings. Column A is sample name,
713 column B is R1 file name and column C is R2 file name. Save this as a .txt. file. This file
714 specifies which fastq files to pair together. Copy the .txt file and all of the sequence files over
715 to the directory that you will be using on the linux platform.

716 • **Move to the linux platform** and navigate to the directory where your sequence files and
717 the .txt file that you have just created is stored. In Mothur use the **make.contigs** command
718 to extract the sequence and quality score data from your fastq files, create the reverse
719 complement of the reverse read and then join the reads into contigs.

720

721 `mothur > make.contigs (file=example.txt, processors=4)`

722

723 This command also produces several files that you will need later: `example.trim.contigs.fasta`
724 and `example.contigs.groups`. These contain the sequence data and group identity for each
725 sequence. Note that 'groups' actually refers to samples in Mothur. The `example.contigs.report`
726 file will tell you something about the contig assembly for each read.

727 Look at a summary of the sequences:

```
728 mothur > summary.seqs (fasta=example.trim.contigs.fasta, processors=4)
```

729

730 Example output:

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	35	35	0	3	1
2.5%-tile:	1	342	342	0	5	214983
25%-tile:	1	343	343	0	5	2149830
Median:	1	344	344	0	5	4299659
75%-tile:	1	348	348	0	5	6449488
97.5%-tile:	1	367	367	8	5	8384334
Maximum:	1	502	502	331	249	8599316
Mean: 1	347.077	347.077	0.881615		5.26081	
# of Seqs:						8599316

731

732 2. Reducing sequencing and PCR errors

733

- 734 • Remove sequences with ambiguous bases and filter sequences based on different
735 criteria using the `screen.seqs` command. The criteria will vary depending on the
736 sequence data.

```
737 mothur > screen.seqs(fasta=example.trim.contigs.fasta, group=example.contigs.groups,  
738 summary=example.trim.contigs.summary, maxn=0, maxambig=0, maxhomop=7,  
739 minlength=342, maxlength=360)
```

- 740 • Re-summarise the sequences after filtering

```
741 mothur > summary.seqs (fasta=example.trim.contigs.good.fasta, processors=4)
```

742 Example output:

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	342	342	0	3	1
2.5%-tile:	1	342	342	0	5	173065
25%-tile:	1	344	344	0	5	1730650
Median:	1	344	344	0	5	3461300
75%-tile:	1	344	344	0	5	5191950
97.5%-tile:	1	355	355	0	5	6749535
Maximum:	1	360	360	0	7	6922599
Mean: 1	345.554	345.554	0	5.00096		
# of Seqs:						6922599

743

744

745 3. Processing improved sequences

746

- 747 • Remove duplicate sequences using the `unique.seqs` command.

748 `mothur > unique.seqs(fasta=example.trim.contigs.good.fasta)`

- 749 • Run the **count.seqs** command to generate a table where the rows are the names of the
750 unique sequences and the columns are the names of the groups. The table is then filled
751 with the number of times each unique sequence shows up in each group.

752 `mothur > count.seqs(name=example.trim.contigs.good.names,`
753 `group=example.contigs.good.groups)`

- 754 • Re-summarise the sequences and observe the number of unique sequences

755 `mothur > summary.seqs (count=example.trim.contigs.good.count_table)`

756 Example output:

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	342	342	0	3	1
2.5%-tile:	1	342	342	0	5	173065
25%-tile:	1	344	344	0	5	1730650
Median:	1	344	344	0	5	3461300
75%-tile:	1	344	344	0	5	5191950
97.5%-tile:	1	355	355	0	5	6749535
Maximum:	1	360	360	0	7	6922599
Mean: 1	345.554	345.554	0	5.00096		
# of unique seqs:		840523				
total # of seqs:		6922599				

757

- 758 • View the number of sequences in each sample. The following command will output a list
759 of samples and the number of sequence reads per sample.

760 `mothur > count.groups(count=example.trim.contigs.good.count_table)`

- 761 • Align the sequences to a customised reference alignment of bacterial sequences. This
762 reference alignment (`silva.bacteria.fasta`) must be downloaded into the directory you are
763 working from and is available here: http://www.mothur.org/wiki/Silva_reference_files

764 `mothur > align.seqs(fasta=example.trim.contigs.good.unique.fasta,`
765 `reference=silva.bacteria.fasta, processors=10)`

- 766 • Summarise the quality of the sequences after making the alignment

767 `mothur > summary.seqs(fasta=example.trim.contigs.good.unique.align,`
768 `count=example.trim.contigs.good.count_table, processors=10)`

769

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1044	1062	1	0	1	1
2.5%-tile:	1044	6424	323	0	5	173065
25%-tile:	1044	6424	325	0	5	1730650
Median:	1044	6424	325	0	5	3461300
75%-tile:	1044	6424	325	0	5	5191950
97.5%-tile:	1044	6424	336	0	5	6749535
Maximum:	43116	43116	359	0	7	6922599
Mean: 1044.1	6424.06	326.565	0	5.00094		
# of unique seqs:		840523				
total # of seqs:		6922599				

770

- 771 • Some sequences may start and end at the same position indicating a very poor alignment,
772 which is generally due to non-specific amplification. To make sure that everything overlaps
773 the same region re-run **screen.seqs** to get rid of sequences that start at or before position
774 1044 and end at or after position 6424, i.e. the start and end values that correspond to the
775 25%-tile and 75%-tile.
- 776 • Include the count table to update the table for the sequences we are removing. Include
777 the summary file so we do not have to figure out again all the start and stop positions.

```
778 mothur > screen.seqs(fasta=example.trim.contigs.good.unique.align,  
779 count=example.trim.contigs.good.count_table,  
780 summary=example.trim.contigs.good.unique.summary, start=1044, end=6424,  
781 maxhomop=7)
```

- 782 • Summarise the sequence data

```
783 mothur > summary.seqs(fasta=current, count=current)
```

784 Example output:

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1044	6424	293	0	3	1
2.5%-tile:	1044	6424	323	0	5	172521
25%-tile:	1044	6424	325	0	5	1725204
Median:	1044	6424	325	0	5	3450407
75%-tile:	1044	6424	325	0	5	5175610
97.5%-tile:	1044	6424	336	0	5	6728292
Maximum:	1044	7694	355	0	7	6900812
Mean:	1044	6424	326.565	0	5.0008	
# of unique seqs:			827648			
total # of seqs:			6900812			

- 785
- 786 • Filter the sequences to remove the overhangs at both ends even though paired-end
787 sequencing should not be an issue. In addition, any column that contains gap characters
788 e.g. "-" will be removed
- 789 • The command **filter.seqs** removes columns from alignments based on a criteria defined
790 by the user. For example, alignments generated against reference alignments (e.g. from
791 RDP, SILVA, or greengenes) often have columns where every character is either a '.' or a
792 '-'. These columns are not included in calculating distances because they have no
793 information in them.
- 794 • Vertical = T: any column that only contains gap characters is ignored. This can be turned
795 off by setting vertical to F.
- 796 • Trump =.: trump option will remove a column if the trump character is found at that position
797 in any sequence of the alignment.

```
798 mothur > filter.seqs(fasta=example.trim.contigs.good.unique.good.align, vertical=T,  
799 trump=.)
```

800 Example output:

```
801 Length of filtered alignment: 804  
Number of columns removed: 49196  
Length of the original alignment: 50000  
Number of sequences used to construct filter: 827648
```

- 802 • According to the output, this means that our initial alignment was 50000 columns wide and
803 that we were able to remove 49196 terminal gap characters using `trump=.` and vertical gap
804 characters using `vertical=T`. The final alignment length is 804 columns. Because some
805 redundancy has been created across our sequences by trimming the ends, re-run
806 **unique.seqs**

```
807 mothur > unique.seqs(fasta=example.trim.contigs.good.unique.good.filter.fasta,  
808 count=example.trim.contigs.good.good.count_table)
```

- 809 • Further de-noise the sequences using the **pre.cluster** command, which implements a
810 pseudo-single linkage algorithm with the goal of removing sequences that are likely due
811 to sequencing errors

```
812 mothur > pre.cluster(fasta=example.trim.contigs.good.unique.good.filter.unique.fasta,  
813 count=example.trim.contigs.good.unique.good.filter.count_table, diffs=2, processors=10)
```

```
814 mothur >  
815 chimera.uchime(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.fasta,  
816 count=example.trim.contigs.good.unique.good.filter.unique.precluster.count_table,  
817 dereplicate=t)
```

```
818 mothur >  
819 remove.seqs(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.fasta,  
820 accnos=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.accnos)
```

821 Example output:

```
822 Removed 43166 sequences from your fasta file.
```

823

- 824 • Re-summarise the sequence data

```
825
```

	Start	End	NBases	Ambigs	Polymer	NumSeqs
Minimum:	1	804	293	0	3	1
2.5%-tile:	1	804	323	0	5	163245
25%-tile:	1	804	325	0	5	1632445
Median:	1	804	325	0	5	3264889
75%-tile:	1	804	325	0	5	4897333
97.5%-tile:	1	804	336	0	5	6366533
Maximum:	1	804	355	0	7	6529777
Mean:	1	804	326.424	0	4.99811	
# of unique seqs:			148203			
total # of seqs:			6529777			

- 826 • The **split.abund** command reads a fasta file and a list or a names file splits the sequences
827 into rare and abundant groups.

```
828 mothur >  
829 split.abund(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.fasta,  
830 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.count_table,  
831 cutoff=2)
```

- 832 • The **classify.seqs** command allows you to use several different methods to assign their
833 sequences to the taxonomy outline of their choice. Current methods include using a k-

834 nearest neighbor consensus and Wang approach. The command requires that you provide
835 a fasta-formatted input and database sequence file (`trainset10_082014.rdp.fasta`) and a
836 taxonomy file (`trainset10_082014.rdp.tax`) for the reference sequences. These can be
837 downloaded from the following website: http://www.mothur.org/wiki/RDP_reference_files

```
838 mothur >  
839 classify.seqs(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.  
840 d.fasta,  
841 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund.co  
842 unt_table, reference=trainset10_082014.rdp.fasta, taxonomy=trainset10_082014.rdp.tax,  
843 cutoff=80, processors=10)
```

- 844 • Next, since bacterial 16S rRNA sequences were amplified, anything classified as species
845 other than bacteria (e.g. mitochondria, archaea, etc.) have to be removed.
- 846 • The **remove.lineage** command reads a taxonomy file and a taxon and generates a new
847 file that contains only the sequences not containing that taxon. You may also include
848 either a `fasta`, `name`, `group`, `list`, `count` or `align.report` file to this command and Mothur will
849 generate new files for each of those that contains only the sequences not containing that
850 taxon.

```
851 mothur >  
852 remove.lineage(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.pic  
853 k.abund.fasta,  
854 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund  
855 .count_table,  
856 taxonomy=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.rdp  
857 .wang.taxonomy, taxon=Chloroplast-Mitochondria-unknown-Archaea-Eukaryota)
```

- 858 • The **cluster.split** command can be used to assign sequences to OTUs and outputs a
859 `.list`, `.rabund`, `.sabund` files. It splits large distance matrices into smaller pieces

```
860 mothur >  
861 cluster.split(fasta=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.ab  
862 und.pick.fasta,  
863 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund  
864 .pick.count_table,  
865 taxonomy=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.rdp  
866 .wang.pick.taxonomy, splitmethod=classify, taxlevel=4, cutoff=0.15, processors=10)
```

- 867 • Observe the number of sequences left

```
868 mothur > count.groups(count=current)
```

869

870 4. Preparing for analysis

871

- 872 • Create a shared file, which has samples in columns and OTUs in rows and displays the
873 count of each OTU in each sample. From this we can see how many sequences are in
874 each OTU from each group and tell Mothur that we're really only interested in the 0.03
875 cutoff level:

```
876 mothur >
877 make.shared(list=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.unique_list.list,
878 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund.pick.count_table, label=0.03)
```

- 881 • Get the consensus taxonomy for each OTU using the [classify.otu](#) command

```
882 mothur >
883 classify.otu(list=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.unique_list.list,
884 count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund.pick.count_table,
885 taxonomy=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.rdp.wang.pick.taxonomy, label=0.03)
```

- 889 • Check the number of sequences in each sample

```
890 mothur >
891 count.groups(count=example.trim.contigs.good.unique.good.filter.unique.precluster.uchime.pick.abund.pick.count_table)
```

- 893 • Subsampling the data ensures that each sample has the same number of reads for downstream analyses. This is done by **sub.sample**. At this point, you need to decide what to normalize all the reads in each sample to. If you do not specify “size=...” in the command, then Mothur will look at all your groups and find the one with the lowest number of reads and sub-sample everything to it, this is shown below:

```
898 mothur >
899 sub.sample(shared=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.unique_list.shared)
```

- 901 • In the case that there is a sample with an extremely low number of read, specify and subsample by a suitable size by changing **XXX** in the following command to the size that you prefer in this command (the minimum read number used should be >1000):

```
904 mothur >
905 sub.sample(shared=example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.unique_list.shared, size=XXX)
```

- 907 • Rename the very long file names to simpler ones. This can be done by the following command, which basically makes a copy of the file using the name you specified. The final files will all have .pick.pick near the end. Do this for the .groups, .list, and subample.shared files. The final .fasta file will come from a subsequent step (the get.oturep step in stage 5). The example below shows the command for renaming the .list file:

```
914 mothur > system (rename
915 example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.unique_list.list example.final.list)
```

917

- 918 • Alternatively, quit mothur and use the cp command to copy of the file and rename the
919 copy you have made. You will need to start the mothur software again once you have
920 copied the files. Example commands are shown below. The first command exits mothur,
921 the second command copies and renames the file (.list in this example), and the third
922 command re-starts mothur.

```
924 mothur > quit  
925 > cp  
926 example.trim.contigs.good.unique.good.filter.unique.precluster.pick.abund.pick.an.uniqu  
927 e_list.list example.final.list  
928 > mothur
```

931 5. Analysis

- 932
- 933 • Obtain a summary file. The **summary.single** command produces a table containing the
934 number of sequences, alpha diversity indices and the sample coverage. Although the
935 downstream analyses (alpha and beta diversity) will be performed using R statistical
936 software (R-Core-Team 2013)(<http://www.R-project.org>), this summary is useful to gain an
937 overview of the sample statistics.

```
938 mothur > summary.single(shared=example.final.shared, calc=nseqs-sobs-chao-ace-  
939 invsimpson-npshannon-coverage)
```

- 940 • The **get.oturep** command generates a fasta-formatted sequence file containing only a
941 representative sequence for each OTU.

```
942 mothur > get.oturep(phylip=example.final.phylip.dist, list=example.final.list,  
943 fasta=example.final.fasta, label=0.03)
```

- 944 • Transfer the following files from the linux platform to **windows**:

- 945 1. example.final.0.03.rep.fasta
- 946 2. example.final.groups.summary
- 947 3. example.final.shared

- 948
- 949 • The remaining actions are performed in **windows**
- 950 • Open example.final.shared in notepad++ , copy and paste in Excel, then copy and paste
951 (transpose) into a new sheet. Insert 1 column at the start labelled Genus. Save this
952 spreadsheet e.g. DiversityResults.xlsx
- 953 • Open the example.final.fasta file in notepad++ and remove all of the '-' symbols. This can
954 be done using the find and replace tool, replacing '-' with a backspace.
- 955 • Download RDP Classifier files from <https://sourceforge.net/projects/rdp-classifier/> and
956 save in a logical location. Copy and paste the example.final.0.03.rep.fasta into the same
957 file as the Classifier executable jar file
- 958 • **NOTE:** In order to run the following script you will need Java (download jre-7u60-
959 windows-x64.exe)
- 960 • Open dos terminal (type cmd into windows search) and navigate to the location of the
961 classifier exe jar file and the example.final.0.03.rep.fasta file

962	cd	To change directory within a drive
963	d:	To change drive (example shown here is to change to D drive)

- 964 • Run the classifier exe jar file to find the OTUs to the genus level. This creates two
965 new files (outfile in the script)

966	java -Xmx1g -jar Classifier.jar --conf=0.5 --hier_outfile=example_unclassified_hier.txt --	
967	assign_outfile=example_unclassified_detail.txt --format=fixrank example.final.0.03.rep.fasta	

- 968 • N.B. The '--assign_outfile=' command may need to be written as '--outputFile='
- 969 • Open example_unclassified_detail in Excel, copy and paste the genus column into the
970 DiversityResults.xlsx genus column. As the DiversityResults.xlsx genus column was
971 created from the subsampled sequences (OTUs) and the example_unclassified_detail
972 file contains genera identified from ALL of the OTUs found in the samples , there will
973 likely be more genera than OTUs. You will need to delete the genera corresponding to
974 removed OTUs for the spreadsheet to make sense. This is not difficult as the genera
975 and OTUs are organised numerically in Excel (smallest to largest).
- 976 • N.B. To further validate the genus IDs for each OTU the sequences from the
977 example.final.0.03.rep.fasta file can be used to search the RDP-II sequence database
978 (<http://rdp.cme.msu.edu/>)
- 979 • You now have a spreadsheet containing OTUs, genera and number of sequence reads
980 per sample. Remove any genera appearing with less than 10 reads across the sample
981 set.
- 982 • Consolidate OTUs to genus level.
- 983 • This spreadsheet can be used to calculate relative abundance of genera for a sample
984 by looking at the % of sequence reads each genus holds. Barcharts can be created in
985 Excel to visualise the data.
- 986 • Further data analysis (alpha and beta diversity, statistical differences between factors
987 for sample groups) is performed using R statistical software using this spreadsheet
988 and associated metadata.

989

990 References

991 R-Core-Team. (2013). R: A Language and Environment for Statistical Computing. Vienna,
992 Austria: R Foundation for Statistical Computing.

993

994

995

996

997

998

999

1000

1001

1002 **Supplementary method S6**

1003 R scripts for statistical analysis of microbiota data

1004

1005 The following R scripts can be used in conjunction with different spreadsheets to analyse
1006 microbiota data. The format of the spreadsheets is described, along with the R commands for
1007 each analysis performed in the paper. The examples used are for the Bcc trial data.

1008

1009 **1) R packages needed**

1010 The following R packages are needed for all the analysis and data visualisation

1011 `library(vegan)`

1012 `library(goeveg)`

1013 `library(car)`

1014 `library(psych)`

1015 `library(lattice)`

1016 `library(FSA)`

1017 `library(PMCMR)`

1018 `library(ggplot2)`

1019 `library(cowplot)`

1020 `library(gamlss)`

1021 `library(dplyr)`

1022 `library(funrar)`

1023 `library(NMF)`

1024 `library(RColorBrewer)`

1025 `library(colorspace)`

1026 `library(nlme)`

1027

1028 **2) Read in OTU tables and metadata spreadsheets**

1029 **#OTU table: Compile an excel spreadsheet with genus/OTU as column 1 without a**
1030 **heading. All the other columns have sample number as a heading and read numbers in**
1031 **rows corresponding to the genus/OTU in column 1. Save as a .txt file and read into R.**

1032 `data<-read.table(file.choose(), header=T)`

1033 **#Metadata table: Compile a spreadsheet with sample number as column 1. All other**
1034 **columns contain metadata (e.g. Patient number, Treatment group, qPCR total**
1035 **abundance, Shannon.diversity, Relative abundance of a key pathogen, FEV1). All**
1036 **columns have headings. Save as a .csv file and read into R.**

1037 `meta_table<-read.csv(file.choose(), row.names=1, check.names=FALSE)`

1038 **3) Calculate Shannon diversity and save as a new file**

```
1039 data_transpose<-t(data)
1040 matrix_data<-data.matrix(data_transpose)
1041 shannon.diversity <- diversity(matrix_data, "shannon")
1042 shannon.results<-as.data.frame(shannon.diversity)
1043 write.csv(shannon.results, "PATH/TO/FILE/Shannon_diversity_results.csv")
1044
1045 #Shannon diversity data: Shannon diversity values can be added into the metadata
1046 spreadsheet before reading to R for analysis.
```

1047

1048 **4) NMDS ordination plot of Bray-Curtis dissimilarity values**

1049 **#Format data**

```
1050 abund_table<-t(data)
1051 meta_table$Treatment<-factor(meta_table$Treatment, levels = c("Screening",
1052 "Start_OligoG", "End_OligoG","Start_Placebo", "End_Placebo","End"))
1053 env.treatment<-as.factor(meta_table$Treatment)
1054 env.treatment #check the levels of treatment
1055 env.patient<-as.factor(meta_table$Patient)
1056 env.patient #check the levels of patient
```

1057

1058 **#Calculate Bray-Curtis dissimilarity values**

```
1059 BC_OTU_table <- vegdist(abund_table, method="bray", binary = FALSE)
1060 mat_BC <- data.matrix(BC_OTU_table)
```

1061

1062 **#NMDS ordination of Bray-Curtis dissimilarity values**

```
1063 sol<-metaMDS(abund_table,distance = "bray", k = 2, trymax = 50)
```

1064

1065 **#Plot NMDS with ellipses around treatment groups and add in genus names (top 7%**
1066 **genera)**

```
1067 win.graph()
1068 plot(sol, display="sites", type="n", main="Treatment")
1069 points(sol, cex=1, pch=19, col=env.treatment)
1070 ordiellipse(sol, group=env.treatment, show.groups = "Screening",
1071 col="black", kind="sd", lwd=2)
1072 ordiellipse(sol, group=env.treatment, show.groups = "Start_OligoG", col=
1073 "red" , kind="sd", lwd=2)
```

```

1074 ordiellipse(sol, group=env.treatment, show.groups = "End_OligoG", col=
1075 "green3" , kind="sd", lwd=2)
1076 ordiellipse(sol, group=env.treatment, show.groups = "Start_Placebo", col=
1077 "blue" , kind="sd", lwd=2)
1078 ordiellipse(sol, group=env.treatment, show.groups = "End_Placebo", col=
1079 "cyan" , kind="sd", lwd=2)
1080 ordiellipse(sol, group=env.treatment, show.groups = "End", col= "magenta" ,
1081 kind="sd", lwd=2)
1082 legend("topright", c("Screening", "Start_OligoG",
1083 "End_OligoG","Start_Placebo", "End_Placebo","End"), cex=0.7, col=c
1084 ("black","red","green3","blue","cyan","magenta"), lwd=2)
1085 top_7_pc<-ordiselect(abund_table, sol, ablim = 0.07, choices = c(1, 2),
1086 method = "axes", freq = FALSE)
1087 ordipointlabel(sol, display="species", select = top_7_pc, col="black",
1088 cex=1, add = TRUE)
1089
1090 #Plot NMDS with ellipses around patient groups
1091 win.graph()
1092 env.patient
1093 plot(sol, display="sites", type="n", main="Patient")
1094 points(sol, cex=1, pch=19, col=env.patient)
1095 ordiellipse(sol, group=env.patient, show.groups = "27610001", col="red",
1096 kind="sd", lwd=2)
1097 ordiellipse(sol, group=env.patient, show.groups = "27610002", col=
1098 "#FF7600" , kind="sd", lwd=2)
1099 ordiellipse(sol, group=env.patient, show.groups = "27610003", col=
1100 "#FFEB00" , kind="sd", lwd=2)
1101 ordiellipse(sol, group=env.patient, show.groups = "27610004", col=
1102 "#9DFF00" , kind="sd", lwd=2)
1103 ordiellipse(sol, group=env.patient, show.groups = "27610005", col=
1104 "#27FF00" , kind="sd", lwd=2)
1105 ordiellipse(sol, group=env.patient, show.groups = "27610006", col=
1106 "#00FF4E" , kind="sd", lwd=2)
1107 ordiellipse(sol, group=env.patient, show.groups = "27610007", col=
1108 "#00FFC4" , kind="sd", lwd=2)
1109 ordiellipse(sol, group=env.patient, show.groups = "27610008", col=
1110 "#00C4FF" , kind="sd", lwd=2)
1111 ordiellipse(sol, group=env.patient, show.groups = "27610009",
1112 col="#004EFF", kind="sd", lwd=2)
1113 ordiellipse(sol, group=env.patient, show.groups = "27610011", col=
1114 "#2700FF" , kind="sd", lwd=2)

```

```

1115 ordiellipse(sol, group=env.patient, show.groups = "27611002", col=
1116 "#9D00FF" , kind="sd", lwd=2)
1117 ordiellipse(sol, group=env.patient, show.groups = "27611005", col=
1118 "#FF00EB" , kind="sd", lwd=2)
1119 ordiellipse(sol, group=env.patient, show.groups = "27611006", col=
1120 "#FF0076" , kind="sd", lwd=2)
1121
1122 #Permanova using adonis to determine whether groups are significantly different.
1123 Adonis cannot handle random effects (Patient) but can structure the formula ~ A + B to
1124 calculate the amount of variation explained by A (Patient) and then B (Treatment). This
1125 is not ideal as permanova can only really handle fixed effects, but the best that can be
1126 done this way.
1127 adonis(mat_BC ~ Patient + Treatment, data =meta_table, permutations = 999)
1128
1129 #Check beta dispersion. Groups should have the same variance to satisfy the
1130 conditions of permanova.
1131 beta <- betadisper(BC_OTU_table, meta_table$Treatment)
1132 permutest(beta) #not significant so datasets should have same variance
1133 beta2 <- betadisper(BC_OTU_table, meta_table$Patient)
1134 permutest(beta2)#not significant so datasets should have same variance
1135
1136 5) Statistical analysis of paired samples from different treatment groups
1137 #Subset the data into separate datasheets for combined (e.g. Start + End) and
1138 individual (e.g. Start) treatments
1139 subsetScreen_End<-subset(meta_table, Treatment=="Screening"|
1140 Treatment=="End")
1141 subsetOligoGstart_end<-subset(meta_table, Treatment=="Start_OligoG"|
1142 Treatment=="End_OligoG")
1143 subsetPlacebostart_end<-subset(meta_table, Treatment=="Start_Placebo"|
1144 Treatment=="End_Placebo")
1145
1146 subsetScreen<-subset(meta_table, Treatment=="Screening")
1147 subsetEnd<-subset(meta_table, Treatment=="End")
1148 subsetStartOligoG<-subset(meta_table, Treatment=="Start_OligoG")
1149 subsetEndOligoG<-subset(meta_table, Treatment=="End_OligoG")
1150 subsetStartPlacebo<-subset(meta_table, Treatment=="Start_Placebo")
1151 subsetEndPlacebo<-subset(meta_table, Treatment=="End_Placebo")
1152
1153 #Statistical analysis of differences in Shannon diversity/total abundance by qPCR
1154 between treatment groups using Wilcoxon signed-rank tests (example given for Screen

```

1155 - End group and Shannon diversity, but same process applies to other treatment group
1156 comparisons and qPCR data)

1157 **#Check normality (shapiro.wilk), the spread of the data (boxplots) and homogeneity of**
1158 **variances (bartlett,test and leveneTest)**

```
1159 by (subsetScreen_End$shannon.diversity, subsetScreen_End$Treatment,  
1160 shapiro.test)  
1161 boxplot(subsetScreen_End$shannon.diversity~subsetScreen_End$Treatment)  
1162 bartlett.test(subsetScreen_End$shannon.diversity~subsetScreen_End$Treatment  
1163 )  
1164 leveneTest(subsetScreen_End$shannon.diversity~subsetScreen_End$Treatment)  
1165
```

1166 **#Further investigation of each treatment group individually for normality (histogram**
1167 **of data distribution and qqplot)**

```
1168 hist(subsetScreen$shannon.diversity)  
1169 qqnorm(subsetScreen$shannon.diversity)  
1170 qqline(subsetScreen$shannon.diversity)
```

1171

1172 **#In this case data were normal with homogeneous variances so used a Wilcoxon**
1173 **signed-rank test**

```
1174 wilcox.test(subsetScreen$shannon.diversity, subsetEnd$shannon.diversity,  
1175 paired = T)
```

1176

1177 **#Statistical analysis of differences in relative abundance of a key pathogen between**
1178 **treatment groups using GAMLSS-BEINF and (mu) logit links (example given for Screen**
1179 **- End group).**

1180 **#Check the distribution of the data, here the data follows a zero-one inflated beta**
1181 **distribution for both treatment groups being compared**

```
1182 hist(subsetScreen$Burkholderia)
```

```
1183 hist(subsetEnd$Burkholderia)
```

1184

1185 **#Perform the GAMLSS-BEINF analysis**

```
1186 subsetScreen_End$Burkholderia_proportion<-subsetScreen_End$Burkholderia/100  
1187 #make proportion and append column to dataset  
1188 subsetScreen_End$Patient<-as.factor(subsetScreen_End$Patient)  
1189 modBcc <- gamlss(Burkholderia_proportion ~ Treatment + random(Patient),  
1190 family = BEINF, data = subsetScreen_End)
```

```

1191 summary(modBcc)
1192
1193 #Plotting the results of all the comparisons as boxplots. There are three pairs of
1194 treatment comparisons (Screen – End, OligoG Start – OligoG End and Placebo Start –
1195 Placebo End) for three response variables (Shannon diversity, qPCR total abundance
1196 and relative abundance) giving nine plots.
1197 #Plotting the nine plots individually
1198 win.graph()
1199 plot1<-ggplot(subsetScreen_End, aes(x = Treatment, y = shannon.diversity,
1200 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1201 = "Set1") + theme(legend.position = "none") +
1202 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1203 ylim (0,2.5)
1204 plot2<-ggplot(subsetOligoGstart_end, aes(x = Treatment, y =
1205 shannon.diversity, fill=Treatment)) + geom_boxplot() + theme_bw() +
1206 scale_fill_brewer(palette = "Set1") + theme(legend.position = "none") +
1207 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1208 ylim (0,2.5)
1209 plot3<-ggplot(subsetPlacebostart_end, aes(x = Treatment, y =
1210 shannon.diversity, fill=Treatment)) + geom_boxplot() + theme_bw() +
1211 scale_fill_brewer(palette = "Set1") + theme(legend.position = "none") +
1212 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1213 ylim (0,2.5)
1214 plot4<-ggplot(subsetScreen_End, aes(x = Treatment, y = qPCR,
1215 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1216 = "Set1") + theme(legend.position = "none") +
1217 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1218 ylim (0,9)
1219 plot5<-ggplot(subsetOligoGstart_end, aes(x = Treatment, y = qPCR,
1220 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1221 = "Set1") + theme(legend.position = "none") +
1222 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1223 ylim (0,9)
1224 plot6<-ggplot(subsetPlacebostart_end, aes(x = Treatment, y = qPCR,
1225 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1226 = "Set1") + theme(legend.position = "none") +
1227 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1228 ylim (0,9)
1229 plot7<-ggplot(subsetScreen_End, aes(x = Treatment, y = Burkholderia,
1230 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette

```

```

1231 = "Set1") + theme(legend.position = "none") +
1232 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1233 ylim (0,100)
1234 plot8<-ggplot(subsetOligoGstart_end, aes(x = Treatment, y = Burkholderia,
1235 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1236 = "Set1") + theme(legend.position = "none") +
1237 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1238 ylim (0,100)
1239 plot9<-ggplot(subsetPlacebostart_end, aes(x = Treatment, y = Burkholderia,
1240 fill=Treatment)) + geom_boxplot() + theme_bw() + scale_fill_brewer(palette
1241 = "Set1") + theme(legend.position = "none") +
1242 theme(axis.title.x=element_blank()) + theme(axis.title.y=element_blank()) +
1243 ylim (0,100)

```

1244

1245 **#Combining all nine plots into one plot**

```

1246 plot_grid(plot1, plot2, plot3, plot4, plot5, plot6, plot7, plot8, plot9,
1247 nrow = 3, align = "h")

```

1248

1249

1250 **6) Linear regression FEV1 and Shannon diversity**

1251 **#In the Bcc trial 'End' samples did not have associated FEV1 data, subset the data to**
1252 **include all time points except 'End'.**

```

1253 subsetFEV1_data<-subset(meta_table, Treatment=="Screening" |

```

```

1254 Treatment=="Start_OligoG" | Treatment=="End_OligoG" |

```

```

1255 Treatment=="Start_Placebo" | Treatment=="End_Placebo")

```

1256

1257 **#Plot the two variables FEV1%predicted and Shannon diversity and perform linear**
1258 **regression to plot a trend line. This steps allows visualisation of the data and**
1259 **identification of potential correlations.**

```

1260 linearmod<-lm(subsetFEV1_data$shannon.diversity~

```

```

1261 subsetFEV1_data$FEV1_percent_predicted)

```

```

1262 print(linearmod)

```

```

1263 summary(linearmod)

```

```

1264 plot(subsetFEV1_data $FEV1_percent_predicted,

```

```

1265 subsetFEV1_data$shannon.diversity, pch=16, col="black", main="Bcc FEV1 vs.
1266 diversity")

```

```

1267 abline(lm(subsetFEV1_data$shannon.diversity~

```

```

1268 subsetFEV1_data$FEV1_percent_predicted))

```

1269

1270 **#Patient needs to be taken into account as a random effect to determine statistical**
1271 **significance. Use a mixed model to do this (nlme package).**

1272

```
1273 mixedlm<-lme(shannon.diversity~FEV1_percent_predicted, random = ~1|Patient,  
1274 data=subsetFEV1_data)
```

```
1275 summary(mixedlm)
```

1276

1277 **#Check residuals of model, residuals need to be evenly distributed around 0 and to fall**
1278 **along the line in a qqplot.**

```
1279 plot(mixedlm)
```

```
1280 qqnorm(resid(mixedlm))
```

```
1281 qqline(resid(mixedlm))
```

1282

1283 **7) Analysis of V1 samples: hierarchical clustering**

1284 **#For V1 Screening samples hierarchical clustering was performed for the Bcc trial and**
1285 **the *P. aeruginosa* trial. NMDS ordination of Bray-Curtis dissimilarity values and**
1286 **boxplots of qPCR total abundance values were also used to investigate the data for the**
1287 **Bcc trial. As the R scripts for NMDS and plotting boxplots have been described**
1288 **previously they will not be shown again. Here hierarchical clustering of the Bcc V1 data**
1289 **is shown.**

1290 **#Read in OTU tables and metadata spreadsheets, the OTU table and metadata**
1291 **spreadsheets used previously need to be modified to remove all samples except those**
1292 **at screening (V1). This can be done in excel and the files read into R as before.**

```
1293 data_V1<-read.table(file.choose(), header=T)
```

```
1294 meta_table_V1<-read.csv(file.choose(), row.names=1, check.names=FALSE)
```

1295

1296 **#Format the OTU table to find the relative abundance of each genus in each sample.**

```
1297 OTU_total<-cbind(data_V1, total = rowSums(OTU_table_to_format))
```

```
1298 OTU_order<-OTU_total[order(-OTU_total$total),]
```

```
1299 OTU_order_no_total<-subset(OTU_order, select = -c(total))
```

```
1300 OTU_order_no_total_matrix<-as.matrix(OTU_order_no_total)
```

```
1301 transposed_matrix<-t(OTU_order_no_total_matrix)
```

```
1302 OTU_relabund<-make_relative(transposed_matrix)
```

```
1303 transposed_relabund<-t(OTU_relabund)
```

```
1304 percent_relabund<-(transposed_relabund)*100
```

1305

1306 **#Indicate which column annotations you need for the heatmap. Here the *Burkholderia***
1307 **species is the column annotation (in the column 'Species' of the metadata file)**

```
1308 env.species<-as.factor(meta_table_V1$Species)
```

```
1309 env.species
```

1310

1311 **#Choose the colour palette for the heatmap. The first command opens a colour palette**
1312 **so you can choose the colour scale, the second command reverses the colours in the**
1313 **palette (light-dark or dark-light)**

```
1314 mypalette<-choose_palette()
```

```
1315 col_rev <- rev(mypalette(25))
```

1316

1317 **#Draw the heatmap. These commands specify to only include the top 15 rows by**
1318 **highest OTU abundance. Hierarchical clustering is performed using Bray-Curtis**
1319 **dissimilarity values and Ward's clustering algorithm.**

```
1320 win.graph()
```

```
1321 Top15_rownames<-rownames(percent_relabund[1:15,])
```

```
1322 heatmap_Bcc_V1=ahitmap(percent_relabund, distfun = function(x) vegdist(x,  
1323 method = "bray"), hclustfun = function(x) hclust(x, method = "ward.D2"),  
1324 color = col_rev, treeheight = 30, Rowv=NA, annCol = env.species, annColors  
1325 = list(c("B. cenocepacia" = "#b2df8a", "B. multivorans" = "#fdbf6f", "No Bcc  
1326 detected" = "#fb9a99")), subsetRow = Top15_rownames)
```

```
1327 heatmap_Bcc_V1
```

1328

1329 **8) Pearson product-moment correlation coefficient (PPMCC) using the pearson**
1330 **correlation to compare bacterial community composition between paired**
1331 **samples**

1332

1333 **#OTU table: Compile an excel spreadsheet for each patient with genus/OTU as column**
1334 **1 without a heading. All the other columns have sample number as a heading and read**
1335 **numbers in rows corresponding to the genus/OTU in column 1. Save as a .txt file and**
1336 **read into R. An example for one patient is given below.**

1337

```
1338 Patient1<-read.table(file.choose(), header=T, check.names="FALSE")
```

1339

1340 **#Perform a correlation test for all of the different combinations of paired samples for**
1341 **each patient. An example for one patient is given below.**

1342

1343 `cor.test(Patient1$V1_1, Patient1$V1_2, method = "pearson")`

1344 `cor.test(Patient1$V2_1, Patient1$V2_2, method = "pearson")`

1345 `cor.test(Patient1$V4_1, Patient1$V4_2, method = "pearson")`

1346 `cor.test(Patient1$V5_1, Patient1$V5_2, method = "pearson")`

1347 `cor.test(Patient1$V7_1, Patient1$V7_2, method = "pearson")`

1348 `cor.test(Patient1$V8_1, Patient1$V8_2, method = "pearson")`

1349

1350 **#These values can be collated to determine key statistics such as mean and median,**

1351 **and the spread of the data visualised using boxplots**

1352

1353

1354

1355