A single community dominates structure and function of a mixture of multiple methanogenic communities

Pawel Sierocinski, Kim Milferstedt, Florian Bayer, Tobias Grosskopf, Mark Alston, Sarah Bastkowski, David Swarbreck, Phil J Hobbs, Orkun S Soyer, Jérôme Hamelin, et al.

1   **A single community dominates structure and function of a mixture of multiple**
2   **methanogenic communities**
3

4   Pawel Sierocinski*[1,6], Kim Milferstedt[2], Florian Bayer[1], Tobias Großkopf[3], Mark Alston[4], Sarah
5   Bastkowski[4], David Swarbreck[4], Phil J Hobbs[5], Orkun S Soyer[3], Jérôme Hamelin[2], Angus
6   Buckling[1]

7   [1] *Biosciences, University of Exeter, Penryn, Cornwall, TR10 9FE, UK*
8   [2] *LBE, INRA, 11100, Narbonne, France*
9   [3] *School of Life Sciences, University of Warwick, Coventry,* CV4 7AL*, UK*
10 [4] *Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, UK*
11 [5] *Anaerobic Analytics Ltd, Okehampton, EX20 1AS, UK*
12 [6] Lead contact
13
14 *Correspondence should be directed to p.sierocinski@exeter.ac.uk

15

16 SUMMARY

17 **The ecology of microbes frequently involves the mixing of entire communities**
18 **(community coalescence), for example flooding events, host excretion and soil tillage**
19 [1,2]**, yet the consequences of this process for community structure and function are**
20 **poorly understood** [3–7]**. Recent theory suggests that a community, due to coevolution**
21 **between constituent species, may act as a partially cohesive unit** [8–11]**, resulting in**
22 **one community dominating following community coalescence. This dominant**
23 **community is predicted to be the one that uses resources most efficiently when grown**
24 **in isolation** [11]**. We experimentally tested these predictions using methanogenic**
25 **communities, for which efficient resource use, quantified by methane production,**
26 **requires coevolved cross-feeding interactions between species** [12]**. Following**
27 **propagation in laboratory-scale anaerobic digesters, community composition**
28 **(determined from 16S rRNA sequencing) and methane production of mixtures of**
29 **communities closely resembled that of the single most productive community grown**
30 **in isolation. Analysis of each community's contribution towards the final mixture**
31 **suggests that certain combinations of taxa within a community might be co-selected**
32 **as a result of coevolved interactions. As a corollary of these findings, we also show**
33 **that methane production increased with the number of inoculated communities. These**
34 **findings are relevant to the understanding of the ecological dynamics of natural**

35 **microbial communities, as well as demonstrating a simple method of predictably**

36 **enhancing microbial community function in biotechnology, health and agriculture** [13]**.**

37

38 RESULTS AND DISCUSSION

39 We wanted to determine if coalesced methanogenic communities were dominated by the

40 community that used resources most efficiently in isolation. We used methanogenic

41 communities primarily because methane production is a useful proxy for the ability of an

42 anaerobic community to fully exploit available resources: Methanogenesis results from the

43 conversion of H_2, CO_2 and short chain fatty acids produced by hydrolysis and fermentation of

44 more complex organic material, and is often the only thermodynamically feasible way of

45 actively removing inhibitory end-metabolites [12]. Moreover, methanogenic communities are

46 characterized by complex cross-feeding interactions [12, 14, 15]; hence, the importance of

47 community cohesion in shaping community performance is likely to be particularly important

48 [9]. To provide insight into the temporal dynamics of compositional and functional change

49 following community mixing, we first measured the methane production and composition of

50 two methanogenic communities derived from industrial Anaerobic Digesters (ADs) (Table 1)

51 grown in isolation or as a mixture in laboratory scale ADs. Both the individual communities

52 and mixes were grown in four replicates. To remove any potentially confounding effects

53 caused by differences in starting density of tested communities, we standardized microbial

54 density based on qPCR-estimated counts of 16S rRNA gene copies. We found that the

55 methane production of the mixed community was initially intermediate between the two

56 individual communities, but after 5 weeks propagation started to produce gas at a rate

57 indistinguishable from the more productive of the individual communities (Figure 1A). We

58 examined both the starting point and the endpoint composition of the single and mixed

59 communities by Illumina sequencing 16s rRNA gene amplicon libraries. Consistent with the

60 phenotypic data, the composition of the mixture was much more similar to the better than the

61 worse performing community at the endpoint (Figure 1B). This was despite the single

62 endpoint communities changing considerably from their ancestral composition over the 5

63 weeks (Figure 1B).

64

65 We next determined if a single community dominated when multiple communities were mixed.

66   To this end, we propagated 10 single communities (from either industrial ADs or sewage or

67   agricultural waste AD feedstocks, with each replicated three times), and ten replicates of a

68   mixture of all ten communities (Table 1). The results were consistent with those from the two-

69   community mixture. First, methane production in mixtures of ten communities was higher than

70   the average of the individual communities. However, methane production of the mixtures did

71   not differ from the best performing single community, P13, (Figure 2A), which, like each of the

72   single communities used, was a constituent of all the mixtures. Second, the community

73   composition of mixtures (which varied very little between replicates, presumably because they

74   all had the same 10 community starting inocula) most closely resembled the best performing

75   community, P13 (Figure 2B). More generally, the more compositionally similar an individual

76   community was to the replicated 10-community mixtures, the greater the gas production of

77   the community when grown in isolation (Figure S1). Other community characteristics that

78   positively correlated with methane production were bacterial cell densities and within-

79   community (alpha) diversity, but not methanogen density (Figure S2). In summary, the results

80   demonstrate that the community most efficient at using resources (which in these

81   experiments was also the most diverse) dominates when multiple communities are mixed

82   together, thus enhancing mixed community productivity beyond the average of the

83   component communities.

84

85   We next explored the ecological mechanism(s) underpinning the observed dominance by the

86   community that produced the most methane. One explanation is that multiple taxa from the

87   same community act as semi-cohesive units and are selected together. This might arise as a

88   result of coevolved mutualistic (or unidirectional) cross-feeding interactions, notably between

89   methanogenic Archaea and hydrogen/acetate producers, where each organism both provides

90   essential resources and removes damaging waste products for each other [12,15,16].

91   Moreover, coevolved resource partitioning can result in taxa being selected together, because

92   species are expected to coevolve to minimise competition with co-occurring taxa [17–19].

93   Note that the selection of multiple taxa together in these contexts does not require any form of

94   group selection [11, 20], but simply selection of particular individuals from a key taxon whose

95   presence provides an advantage for individuals from taxa they have coevolved with. This

96 process can be described as ecological co-selection, equivalent to genetic co-selection,

97 where a gene can hitchhike to high frequency purely as consequence of being linked to

98 genes under positive selection [21].

99

100 An alternative explanation is that coevolved interactions within individual communities are

101 relatively unimportant, and the dominant community simply contains more competitive taxa

102 (for any functional task/interaction) than other communities. This does not imply that

103 coevolved cross-feeding interactions are unimportant for methanogenic communities, but that

104 these co-evolved interactions are no more specific for taxa isolated from within a community

105 than taxa isolated from different communities. In other words, functionally equivalent taxa are

106 interchangeable between communities. These different scenarios, selection for the best

107 individual taxa and co-selection, are two extremes of a continuum. The distinction is important

108 because dominance by a single community is necessarily a more likely consequence of

109 community coalescence when co-selection operates. Figure 3 (ABC) provides an illustration

110 of the two extreme scenarios, no co-selection and co-selection of the entire community, and

111 an intermediate case where there are two groups of interacting taxa, or modules, and co-

112 selection occurs within each.

113

114 The most direct way to demonstrate a role of co-selection would be to show that the outcome

115 of competition between single taxa from different communities does not predict the outcome

116 of competition at the community level [11]. Unfortunately, this is not feasible for such complex

117 communities, in which many taxa are very difficult to grow in isolation. However, there are

118 other testable predictions associated with the operation of co-selection or otherwise. If the

119 success of an individual taxon is independent of whether they are in the presence of taxa

120 from the same community (i.e. co-selection does not occur), communities that use resources

121 most efficiently and hence achieve the highest biomass per unit of time (productivity) will

122 contain the highest number of the best-performing taxa. It then follows that there will be a

123 positive relationship between the productivity of a community and the proportion of taxa it

124 contributes to the mixture (Figure 3A). If instead taxa are co-selected as modules, the

125 correlation between individual community contribution and productivity is likely to break down.

126 This is best illustrated by the extreme scenario whereby all taxa within a mixed community

127 are co-selected from a single community: the mixture will be entirely dominated by a single

128 constituent community, hence the contribution of all other communities will be independent of

129 their individual productivity (i.e. they will contribute null to the mixture's composition, even

130 though they have non-zero productivity individually; Figure 3B). The intermediate scenario,

131 where co-selection occurs within two independent modules can also break down this

132 correlation if one module contributes much more to community productivity than the other

133 (Figure 3C).

134

135 To determine if co-selection contributed to our findings, we first estimated the contribution of

136 each community to the 10-community mixtures using a non-negative least squares (NNLS)

137 approach. The community that had the most similar composition to the mixtures (and

138 produced the most methane) contributed an estimated 40% of its taxa to the mixtures, with

139 only two other communities contributing more than 10% of their taxa to the mixtures (Figure

140 3D). We then correlated the contribution each community made to the mixtures with two

141 measures of community productivity: methane production and cell densities (based on 16s

142 rRNA gene copy number), which themselves were positively correlated (Figure S2A). We

143 found no suggestion of a positive correlation between either measure of productivity and

144 contribution to the community (Figure 3D and E). These results suggest that co-selection of

145 taxa played an important role in dominance by the community that produced the most

146 methane.

147

148 That community coalescence results in the most productive individual community dominating

149 the mixed community has direct implications for biotechnological uses of microbial

150 communities. Given that the best performing community in isolation largely determined both

151 the composition and performance of mixtures of communities, methane production should

152 increase with increasing number of communities in a mixture. We therefore inoculated

153 laboratory-scale anaerobic digesters with 1, 2, 3, 4, 6 or 12 communities, ensuring that each

154 of the 12 starting communities was used an equal amount of times at each diversity level

155 ([22]; see Table S1). Cumulative methane production over a five-week period increased with

156 increasing number of communities used as an inoculum (Figure 2C). The positive correlation

157 between community function and the number of inoculating communities is analogous to the

158 commonly observed finding that community productivity increases with increasing species

159 diversity [23]. In this case, the mechanism underlying this positive relationship between the

160 number of communities and productivity is a "sampling effect": inoculating more communities

161 increases the chance that the best performing community will be present in the mix [24].

162 However, given that domination of mixtures by one community was not complete (Figure 3D

163 and E), it is possible that mixing communities could increase performance beyond that of the

164 maximum of single communities in some circumstances (transgressive over-yielding, [25]).

165

166 Here, we have shown that coalescence of microbial communities results in dominance of a

167 single community, the identity of which can be predicted from its efficiency of resource use in

168 isolation. This dominance is likely to be driven in part by co-selection of interacting taxa within

169 coevolved communities, which is likely to greatly increase the magnitude of dominance

170 following mixing [11]. It is unclear whether such effects would be apparent for aerobic

171 communities where cross-feeding interactions are less important for efficient resource use

172 [26], although studies to date [4] suggest asymmetric outcomes, although less extreme, may

173 be common. Our study has also identified a simple method to significantly improve methane

174 yield during anaerobic digestion: inoculate digesters with a broad range of microbial

175 communities, and the best performing community will dominate. However, further work under

176 a range of conditions is clearly required to determine the generality of these findings. Given

177 that resource use efficiency is often a desirable property of microbial communities, this

178 approach could be applied to a range of biotechnological processes driven by microbial

179 communities, as well as to manipulate microbiomes in clinical and agricultural contexts [13].

180

181 AUTHOR CONTRIBUTIONS

182 Methodology: PS, KM, FB, OSS, PJH, JH, AB; Formal analysis PS, SB, MA and AB;

183 Investigation: PS and FB; Writing - Original Draft: PS and AB; Writing - Review and Edition:

184 All authors; Visualisation PS and AB; Supervision: AB

185

191

192 REFERENCES

193 1.   Rillig, M.C., Antonovics, J., Caruso, T., Lehmann, A., Powell, J.R., Veresoglou, S.D.,

194   and Verbruggen, E. (2015). Interchange of entire communities: Microbial community

195   coalescence. Trends Ecol. Evol. *30*, 470–476.

196 2.   Rillig, M.C., Lehmann, A., Aguilar-Trigueros, C.A., Antonovics, J., Caruso, T., Hempel,

197   S., Lehmann, J., Valyi, K., Verbruggen, E., Veresoglou, S.D., *et al.* (2016). Soil

198   microbes and community coalescence. Pedobiologia (Jena). *59*, 37–40.

199 3.   Hausmann, N., and Hawkes, C. (2009). Plant neighborhood control of arbuscular

200   mycorrhizal community composition. New Phytol. *183*, 1188–1200.

201 4.   Livingston, G., Jiang, Y., Fox, J., and Leibold, M. (2013). The dynamics of community

202   assembly under sudden mixing in experimental microcosms. Ecology *94*, 2898–2906.

203 5.   Souffreau, C., Pecceu, B., Denis, C., Rummens, K., and De Meester, L. (2014). An

204   experimental analysis of species sorting and mass effects in freshwater

205   bacterioplankton. Freshw. Biol. *59*, 2081–2095.

206 6.   Adams, H.E., Crump, B.C., and Kling, G.W. (2014). Metacommunity dynamics of

207   bacteria in an arctic lake: The impact of species sorting and mass effects on bacterial

208   production and biogeography. Front. Microbiol. *5*.

209 7.   Calderón, K., Spor, A., Breuil, M.-C., Bru, D., Bizouard, F., Violle, C., Barnard, R.L.,

210   and Philippot, L. (2017). Effectiveness of ecological rescue for altered soil microbial

211   communities and functions. ISME J. *11*, 272–283.

212 8.   Gilpin, M. (1994). Community-level competition: asymmetrical dominance. Proc. Natl.

213   Acad. Sci. U. S. A. *91*, 3252–3254.

214 9.   Toquenaga, Y. (1997). Historicity of a simple competition model. J. Theor. Biol. *187*,

215   175–181.

216   10.   Wright, C.K. (2008). Ecological community integration increases with added trophic

217   complexity. Ecol. Complex. *5*, 140–145.

218   11.   Tikhonov, M. (2016). Community-level cohesion without cooperation. Elife *5*, e15747.

219   12.   Schink, B. (1997). Energetics of syntrophic cooperation in methanogenic degradation.

220   Microbiol. Mol. Biol. Rev. *61*, 262–280.

221   13.   Rillig, M.C., Tsang, A., and Roy, J. (2016). Microbial community coalescence for

222   microbiome engineering. Front. Microbiol. *7*, 6–8.

223   14.   Hillesland, K.L., and Stahl, D. a (2010). Rapid evolution of stability and productivity at

224   the origin of a microbial mutualism. Proc. Natl. Acad. Sci. U. S. A. *107*, 2124–2129.

225   15.   Embree, M., Liu, J.K., Al-Bassam, M.M., and Zengler, K. (2015). Networks of energetic

226   and metabolic interactions define dynamics in microbial communities. Proc. Natl.

227   Acad. Sci. U. S. A. *112*, 15450–15455.

228   16.   Großkopf, T., and Soyer, O. (2016). Microbial diversity arising from thermodynamic

229   constraints. ISME J. *10*, 2725–2733.

230   17.   Schluter, D. (2000). The ecology of adaptive radiation.

231   18.   Roughgarden, J. (1976). Resource partitioning among competing species - A

232   coevolutionary approach. Theor. Popul. Biol. *9*, 388–424.

233   19.   MacArthur, R.H. (1970). Species-packing and competitive equilibrium for many

234   species. Theorectical Popul. Biol. *1*, 1–11.

235   20.   Gardner, A., and Grafen, A. (2009). Capturing the superorganism: a formal theory of

236   group adaptation. J. Evol. Biol. *22*, 659–671.

237   21.   Baker-Austin, C., Wright, M.S., Stepanauskas, R., and McArthur, J. V. (2006). Co-

238   selection of antibiotic and metal resistance. Trends Microbiol. *14*, 176–182.

239   22.   Hodgson, D.J.D., Rainey, P.B., and Buckling, A. (2002). Mechanisms linking diversity,

240   productivity and invasibility in experimental bacterial communities. *269*, 2277–2283.

241   23.   Tilman, D., and Lehman, C. (1997). Plant diversity and ecosystem productivity:

242   theoretical considerations. Proc. Natl. Acad. Sci. U. S. A. *94*, 1857–1861.

243   24.   Tilman, D. (1999). The ecological consequences of changes in biodiversity: A search

244   for general principles. In Ecology (Ecological Society of America), pp. 1455–1474.

245   25.   Harper, D. (1977). The population biology of plants (Academic Press).

246   26.   Morris, B.E.L., Henneberger, R., Huber, H., and Moissl-Eichinger, C. (2013). Microbial

247   syntrophy: Interaction for the common good. FEMS Microbiol. Rev. *37*, 384–406.

248   27.   Einen, J., Thorseth, I.H., and Øvreås, L. (2008). Enumeration of Archaea and Bacteria

249   in seafloor basalt using real-time quantitative PCR and fluorescence microscopy.

250   FEMS Microbiol. Lett. *282*, 182–187.

251   28.   Ruijter, J.M., Ramakers, C., Hoogaars, W.M.H., Karlen, Y., Bakker, O., van den hoff,

252   M.J.B., and Moorman, A.F.M. (2009). Amplification efficiency: Linking baseline and

253   bias in the analysis of quantitative PCR data. Nucleic Acids Res. *37*, e45.

254   29.   Brankatschk, R., Fischer, T., Veste, M., and Zeyer, J. (2013). Succession of N cycling

255   processes in biological soil crusts on a Central European inland dune. FEMS

256   Microbiol. Ecol. *83*, 149–160.

257   30.   Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013).

258   Development of a dual-index sequencing strategy and curation pipeline for analyzing

259   amplicon sequence data on the MiSeq Illumina sequencing platform. Appl. Environ.

260   Microbiol. *79*, 5112–5120.

261   31.   Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and

262   Sogin, M.L. (2013). Oligotyping: differentiating between closely related microbial taxa

263   using 16S rRNA gene data. Methods Ecol. Evol. *4*, 1111–1119.

264   32.   Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello,

265   E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., *et al.* (2010). QIIME allows

266   analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

267   33.   Edgar, R.R.C. (2010). Search and clustering orders of magnitude faster than BLAST.

268   Bioinformatics *26*, 2460–2461.

269   34.   Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME

270   improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

271   35.   McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A.,

272   Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes

273   taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and

274   archaea. ISME J. *6*, 610–8.

275   36.   Mullen, K., and Van Stokkum, I. (2007). The Lawson-Hanson algorithm for non-

276   negative least squares (NNLS).

277   37.   Soetaert, K., Meersche, K.V.D., and Oevelen, D. V (2009). Package limSolve, solving

278   linear inverse models in R.

279

280   FIGURE LEGENDS

281   **Figure 1: Temporal dynamics of methane production and composition when two**

282   **communities are mixed.** A) Cumulative methane production in ml (±SEM) over time of:

283   community P01 (white circles), community P05 (black circles) and their mixes (grey circles).

284   Cumulative methane production differed between treatments (ANOVA: $F_{2,9} = 23.2$, $P <$

285   0.001), but did not differ between the mixed community and P05 (Tukey-Kramer HSD: $P =$

286   0.5). P01 was lower than both other treatments (Tukey-Kramer HSD: $P < 0.001$ in both

287   cases). B) NMDS plot of unweighted UniFrac of communities P01 (white), P05 (black) and

288   their mixes (grey). Ancestral samples are represented by squares with samples from the

289   endpoint of the experiment by circles. At the endpoint, P05 was compositionally more similar

290   to the mixtures than P01, based on both unweighted (*t-tests* of mean distance to each mixture

291   for each replicate single community: $t_6 = 8.3$, $P < 0.001$) and weighted ($t_6 = 2.3$, $P = 0.03$)

292   UniFrac distances.

293

294   **Figure 2: Methane production and community composition when multiple communities**

295   **are mixed.** A) Total methane production of mixed (grey) and individual communities (white),

296   with mean values shown as horizontal lines. Mean total methane production was greater for

297   mixtures than for individual communities (*t*-test: $P < 0.001$ in 9 cases), except when measured

298   against community P13 (the best performer). B) NMDS plot of unweighted unifrac of 10

299   mixtures (grey) and 9 individual communities (white). Numbers in circles refer to individual

300   community identifiers (Table 1). Community P13 was significantly closer in composition to the

301   10 mixed communities than any other community (weighted and unweighted UniFrac

302   distances; Paired *t*-tests; $P < 0.001$, in all cases). There was also a significant link between

303   the community composition and the difference in gas production between the communities

304   (see Figure S1). Note: DNA yield from community P06, which had the lowest gas production

305   of all communities, was insufficient for sequencing, therefore it is excluded from this and

306   following graphs. C) Individual communities (white circles) and their average methane

307   production (white line); mixes of communities (grey circles) and their averages (grey line).

308   There was a monotonic increase in methane production with number of communities used

309   (Regression: $F_{1,26} = 5.4$, $P = 0.03$). For community composition of the mixes, see Table 1 and

310   Table S1.

311

312

313   **Figure 3: The role of co-selection in explaining dominance by a single community**. A-C)

314   The top panels illustrate three hypothetical scenarios describing how communities contribute

315   to a mixture of communities, while the bottom panels show the expected relationships

316   between a community's contribution and its methane production. The letters within the top

317   panels indicate taxa that drive two biochemical processes (abcd and ef); capitalised letters

318   are the best representatives of a taxon among all the communities. A). No co-selection. B).

319   Co-selection of all taxa within a community. C). Co-selection of taxa within two independent

320   modules. D). Mean estimated relative contribution of each individual community (numbered)

321   towards the 10 coalesced communities calculated using the NNLS method, plotted against

322   mean cumulative methane production for each community; there is no significant relationship

323   (Regression; $F_{1,7} = 1.7$, $P > 0.2$). E) As D, but relative contribution is plotted against number of

324   bacterial and archaeal cells calculated based on the 16S rRNA gene copy number

325   (Regression; $F_{1,7} = 1.7$, $P > 0.5$. Note the relative contribution is not a fractional contribution

326   because some OTUs present in the mixture were not detected in the constituent

327   communities. This is presumably because they only reached detected frequencies in the

328   mixture, but we can't rule out that the community that we failed to get sufficient reads from

329   contributed to the composition of the mixtures. Mind that the cell densities of Archaea and

330   bacteria do not significantly correlate with the gas production (see Figure S2).

331

332   Table 1: **List of individual communities used in this analysis and their source.** All

333   Anaerobic Digester (AD) communities were derived from industrial ADs in the South West of

334 England. Specific locations cannot be provided because of commercial sensitivity. Note that

335 experiment numbers correspond with figure numbers.

336

Sample name	Feed/Type	Temperature	Used in experiments
P01	Silage and Foodwaste Anaerobic Digester (AD)	44 - 42.5°C	1,2,3
P02	Silage + Food waste AD	44 - 42.5°C	2,3
P03	Maize/Cow Slurry/Chicken Manure AD	45°C	3
P04	Maize/Cow Slurry/Chicken Manure AD	45°C	2,3
P05	Sewage Sludge AD	36°C	1,2,3
P06	Raw Sewage	Ambient	2,3
P08	Thickened Sewage Sludge	Ambient	2,3
P09	Sewage Based AD	36°C	2,3
P10	Food Waste AD	36°C	2,3
P11	Cow Slurry	Ambient	3
P12	Silage, Slurry and Manure Pre-Digestate	Ambient	3
P13	Silage, Slurry and Manure AD	40°C	2,3
P15	Food waste AD	36°C	2

337

338

339

340

341 STAR METHODS

342 **CONTACT FOR REAGENT AND RESOURCE SHARING**

343

344 The authors are happy to share any further resources linked to the research involved with

345 qualified third parties. Further information and requests for resources and reagents should be

346 directed to and will be fulfilled by the Lead Contact, Pawel Sierocinski

347 (p.sierocinski@exeter.ac.uk).

348

349 **EXPERIMENTAL MODEL AND SUBJECT DETAILS**

350

351 **Methanogenic communities**

352 The communities used were collected from anaerobic digesters (AD plants: communities

353 P01,P02, P03, P04, P05, P09, P10, P13 and P15) and communities present in nature used to

354 seed the AD plants (communities P06, P08, P11, P12, see the details in Table 1). All the

355 communities have been collected in the South West area of United Kingdom from operating

356 Anaerobic Digesters and the seeding communities they use for the reactors. The

357   communities were operating at temperatures between ambient and 45°C in their places of

358   origin. Communities were stored at 4°C prior to use.

359

360   **METHOD DETAILS**

361

362   **Cultivation details**

363   For all experiments, communities were grown in 500 ml bottles (600ml total volume with

364   headspace; Duran) using the commercially available Automated Methane Potential Test

365   System (AMPTS, Bioprocess Control Sweden AB) to measure CO_2-stripped biogas

366   production (referred to as methane in this paper). Samples were fed weekly with 25 ml of

367   defined medium in a fed-batch mode using a defined medium (see below for media

368   composition).

369

370   The communities used in experiment 1 were equalised in terms of bacterial cells per gram of

371   sample before inoculation using M9 salts to dilute them to the community with the lowest cell

372   density, based on qPCR enumeration of 16S rRNA gene copies. For experiments 2 and 3,

373   starting 16S rRNA copy number was determined (but not equalised between communities)

374   and did not correlate with methane production. The fermenters were inoculated with 275 g of

375   sample and fed with 25 ml of defined medium: meat extract 111.1 gl^{-1}, cellulose 24.9 gl^{-1},

376   starch 9.8 gl^{-1} glucose 0.89 gl^{-1}, xylose 3.55 gl^{-1} (carbon to nitrogen ratio of 15:1) every week,

377   starting with t_0. Before the start of the fermentation, 0.3 mL of 1000x Trace Metal stock (1 gl^{-1}

378      $FeCl_2 \cdot 4H_2O$, 0.5 gl^{-1} $MnCl_2 \cdot 4H_2O$, 0.3 gl^{-1} $CoCl_2 \cdot 4H_2O$, 0.2 gl^{-1} $ZnCl_2$, 0.1 gl^{-1} $NiSO_4 \cdot$

379      $6H_2O$, 0.05 gl^{-1} $Na_2MoO_4 \cdot 4H_2O$, 0.02 gl^{-1} H_3BO_3, 0.008 gl^{-1} $Na_2 WO_4 \cdot 2H_2O$, 0.006 gl^{-1}

380      $Na_2SeO_3 \cdot 5H_2O$, 0.002 gl^{-1} $CuCl_2 \cdot 2H_2O$) was added to each fermenter.

381

382   **Experiment structure**

383   In Experiment 1 (results shown in Figure 1) we cultivated community P01 (four replicates) and

384   community P05 (four replicates) and a 1:1 mix of the two. It ran for 5 weeks before samples

385   were harvested for sequencing (see below). The initial community was sequenced at the

386   same time. In Experiment 2 (results shown in Figure 2), we cultivated 10 individual

387 communities (listed in Table 1), in triplicate, and 10 mixes of all 10 communities mixed in

388 equal volumes, at the same total volume as the single communities. After 6 weeks samples

389 were harvested for sequencing. In Experiment 3 (results shown in Figure 2C) we used 12

390 communities (detailed in Table 1). They were grown in isolation as well as pseudo-randomly

391 combined to create a gradient of number of starting communities, with each community used

392 only once for each number of communities. This resulted in 12 single communities, 6 pairs, 4

393 triplets, 3 quadruplets, 2 mixes of 6 and one mix of 12 communities. Specific details of mixing

394 can be seen in Table S1. The cultures were propagated for 5 weeks.

395

396 **Measuring methane content of Biogas**

397 All resulting lab-scale reactors inoculated with the samples were run at 37°C using the

398 Automatic Methane Potential Test System (AMPTS). The AMPTS is a setup of 15 simple

399 fermenters using a 0.5L lab bottle as the vessel with its own stirring system provided with a

400 butyl rubber stopper and sampling ports. It is connected to an online gas measuring system to

401 allow continuous gas measurements. The AMPTS system measures the volume of biogas

402 produced following stripping of CO_2 (by passing the gas through 50 ml of 3M NaOH solution)

403 from the produced gas. To reproduce our results, however, there is no need for a

404 sophisticated setup, some pilot experiments yielding similar results in terms of gas production

405 were conducted using anaerobic serum flasks. We confirmed that the measured biogas was

406 >95% methane using Gas Chromatography with Flame Ionisation Detection optimized for

407 methane detection.

408

409 **DNA extraction and qPCR quantification**

410 DNA for 16S rRNA gene amplicon sequencing was extracted using QIAamp DNA Stool Mini

411 Kit (QIAGEN) or FastDNA™ SPIN Kit for Soil (MP), depending on the experiment. Note that

412 DNA extraction for mixed community P06 from experiment 2 failed. The DNA for qPCR was

413 extracted with the QIAamp DNA Stool Mini Kit (QIAGEN), protocol for pathogen detection

414 with the 95ºC incubation step and the Powerlyzer Powersoil DNA KIT (MOBIO). DNA from

415 *Acinetobacter baylyi*, *Pseudomonas fluorescens* SBW25 for Bacteria and from *Halobacterium*

416 *salinarum* DSM 669 for Archaea was used as standards. The primers [27] used to quantify

417 Bacteria were 16S rRNA 338f - ACT CCT ACG GGA GGC AGC AG, 518r - ATT ACC GCG

418 GCT GCT GG for Archaea: 931f - AGG AAT TGG CGG GGG AGC A, m1100r - BGG GTC

419 TCG CTC GTT RCC. The reagents used were: 1x Brilliant III Ultra-Fast SYBR® Green QPCR

420 Master Mix; 150nM 338f and 300nM 518r or 300nM 931f and 300 nM m1100r; ROX 300nM;

421 and BSA 100 ng/µl final concentration. All samples were run in triplicate on a StepOnePlus

422 (Applied Biosystems) qPCR machine using a program with 3 minutes 95ºC initial denaturation

423 followed by 40 cycles of 5 seconds at 95ºC and 10 seconds at 60ºC, followed by a melting

424 curve 95ºC for 15 seconds; 60ºC for 1 minute ramping up to 95ºC in steps of 0.3ºC for 15

425 seconds each. The melting curve analysis and the confirmation of the negative controls was

426 done using StepOne Software v.2.3 (life technologies). The Cq values and the efficiencies of

427 the samples and standards was determined as previously using LinRegPCR version

428 2016.0[28]. The quantities were calculated using the one point calibration method as

429 described earlier[29].

430

431 **Amplicon library construction and sequencing**

432 16S rRNA gene libraries were constructed using primers designed to amplify the V4 region

433 and multiplexed [30]. Amplicons were generated using a high-fidelity polymerase (Kapa 2G

434 Robust) and purified using the Agencourt AMPure XP PCR purification system and quantified

435 using a fluorometer (Qubit, life technologies). The purified amplicons were then pooled in

436 equimolar concentrations by hand based on Qubit quantification. The resulting amplicon

437 library pool was diluted to 2 nM with sodium hydroxide and 5 µl transferred into 995 µl HT1

438 (Illumina) to give a final concentration of 10 pM. 600 µl of the diluted library pool was spiked

439 with 10% PhiX Control v3 and placed on ice before loading into Illumina MiSeq cartridge

440 following the manufacturer's instructions. The sequencing chemistry utilised was MiSeq

441 Reagent Kit v2 (500 cycles) with run metrics of 250 cycles for each paired end read using

442 MiSeq Control Software 2.2.0 and RTA 1.17.28.

443

444 **Analyses of sequenced samples**

445 MiSeq amplicon reads were merged using Illumina-utils software [31]. We quality-filtered only

446 the mismatches in the overlapping region between read pairs using a minimum overlap (--

447   min-overlap-size) of 200 nt and a minimum quality Phred score (--min-qual-score) of Q20. We

448   allowed no more than five mismatches per 100 nt (-P 0.05) over the 200 nt overlapping

449   region.

450

451   Reads that fulfilled the quality criteria were analysed using Quantitative Insights Into Microbial

452   Ecology (QIIME v.1.7) [32]. We removed chimera using the *identify_chimeric_seqs.py* script,

453   UCHIME reference 'Gold' database and USEARCH [33,34], which we also used to select

454   OTUs. We assigned the taxonomy of our reads with QIIME *pick_open_reference_otus.py*

455   function, using the Greengenes database version v13_8[35] as a reference with a minimum

456   cluster size of 2 (i.e., each OTU must contain at least two sequences). We collapsed the

457   technical replicates and filtered out the low abundance OTUs (<0.01% total,

458   *filter_otus_from_otu_table.py*) and samples rarefied to an even depth of 26702 for both

459   experiments where sequencing data is available. QIIME was used to calculate alpha and beta

460   diversity data and produce NMDS plots.

461

462   **QUANTIFICATION AND STATISTICAL ANALYSIS**

463   MacQiime was used to calculate alpha and beta diversity data and produce NMDS plots.

464   Data obtained with MacQiime was later combined with the gas production data and analysed

465   using JMP Pro 13 software (SAP) as described in the Figure legends.

466

467   For the NNLS analysis, following removal of low abundance OTUs and cumulative sum

468   scaling transformation, the resulting biom file was used to create a matrix $A \in \mathbb{Z}_{\geq 0}^{m \times n}$ (*m* rows

469   of OTUs by *n* sample columns) for all of the single communities, and a column vector $b \in \mathbb{Z}_{\geq 0}^{m}$

470   for each of the mixed communities; both A and b hold non-negative integers of OTU

471   abundances. Note that one of the individual samples contained a negligible number of reads

472   and was discarded from the analysis. The contribution, or weight, of each seed sample to the

473   pattern of OTUs observed in each of the mixed communities is given by the column vector $x \in$

474    $\mathbb{R}^{n}$ when solving a system of linear equations $Ax = b$. Written out this equation $(Ax = b)$

475   looks like this for each mixture:

476

477

$$\begin{pmatrix} \text{OTU}_{1,S1} & \cdots & \text{OTU}_{1,Sn} \\ \vdots & \ddots & \vdots \\ \text{OTU}_{m,S1} & \cdots & \text{OTU}_{m,Sn} \end{pmatrix} \text{x} \begin{pmatrix} x_{S1} \\ \vdots \\ x_{Sm} \end{pmatrix} = \begin{pmatrix} OTU_{mix\_1} \\ \vdots \\ OTU_{mix\_m} \end{pmatrix}$$

479
480
481   where S refers to each single community.
482

483   When modelling count data for environmental samples the fitted parameters of x will

484   also be non-negative and the number of OTUs will usually exceed the number of samples (m

485   > n). The task is to solve an over-determined system of linear equations where there are

486   more equations than unknowns. It is likely that some of the linear equations will 'disagree' and

487   there will be no exact solution. Geometrically, this may be interpreted as b not lying in the

488   column space of A, a (hyper)plane holding the column vectors of A, or $Ax - b \neq 0$. A least-

489   squares approach may find the non-negative vector $\bar{x} = (A^T A)^{-1} A^T b$ which is the projection of

490   b back onto the column space of A that minimises the least-squares error 'distance' $\|A\bar{x} - b\|$.

491   For our study the non-negative least-squares (NNLS) solution, $\bar{x}$, and least-squares errors

492   were computed via the R packages 'nnls' [36] and 'limSolve' [37] for each of the mixed

493   samples, with the mean relative contribution of each community to the 10 mixtures shown in

494   Figure 2C (relative to the gas production)..

495

496   **DATA AND SOFTWARE AVAILABILITY**
497
498   The raw sequences obtained from our experiments are available at the European Nucleotide

499   Archive and may be accessed at http://www.ebi.ac.uk/ena/data/view/PRJEB21193

500   (Experiment 1) and http://www.ebi.ac.uk/ena/data/view/PRJEB21187 (Experiment 2). We also

501   included the R code that allows the user to calculate the contribution a single community has

502   in a mix of communities (see Method S1). Using this code, a NNLS analysis can be

503   conducted with the input of a pre-filtered OTU table.

504
505

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Bacterial and Virus Strains		
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
Brilliant III Ultra-Fast SYBR (R) Green QPCR Master Mix	Agilent Technologies	600882-51
meat extract	Sigma-Aldrich Co. LLC.	70164-500G
Xylose	Sigma-Aldrich Co. LLC.	W360100-1KG
Cellulose	Sigma-Aldrich Co. LLC.	C6288
Starch	Sigma-Aldrich Co. LLC.	33615-1KG
Glucose	Sigma-Aldrich Co. LLC.	G8270-1KG
Critical Commercial Assays		
QIAamp DNA Stool Mini Kit (QIAGEN)	Qiagen	ID: 51504
FastDNA™ SPIN Kit for Soil	MP Biomedicals, LLC	116560200
PowerLyzer® PowerSoil® DNA Isolation Kit	MO BIO Laboratories, Inc.	12855-100
Deposited Data		
Sequencing data Experiment 1	European Nucleotide Archive	http://www.ebi.ac.uk/ena/data/view/PRJEB21193

Sequencing data Experiment 2	European Nucleotide Archive	http://www.ebi.ac.uk/ena/data/view/PRJEB21187
Experimental Models: Cell Lines		
Experimental Models: Organisms/Strains		
Community P01	Silage and Foodwaste Anaerobic Digester (AD)	This paper
Community P02	Silage + Food waste AD	This paper
Community P03	Maize/Cow Slurry/Chicken Manure AD	This paper
Community P04	Maize/Cow Slurry/Chicken Manure AD	This paper
Community P05	Sewage Sludge AD	This paper
Community P06	Raw Sewage	This paper
Community P08	Thickened Sewage Sludge	This paper
Community P09	Sewage Based AD	This paper
Community P10	Food Waste AD	This paper
Community P11	Cow Slurry	This paper
Community P12	Silage, Slurry and Manure Pre-Digestate	This paper
Community P13	Silage, Slurry and Manure AD	This paper
Community P15	Food waste AD	This paper
Oligonucleotides		
338f - ACT CCT ACG GGA GGC AGC AG	[27]	
518r - ATT ACC GCG GCT GCT GG	[27]	
931f - AGG AAT TGG CGG GGG AGC A	[27]	
m1100r - BGG GTC TCG CTC GTT RCC	[27]	
Recombinant DNA		
Software and Algorithms		

StepOne Software v.2.3	life technologies	https://www.thermofisher.com/uk/en/home/technical-resources/software-downloads/StepOne-and-StepOnePlus-Real-Time-PCR-System.html#
LinRegPCR version 2016.0	[28]	linregpcr.nl
R version 3.4.0	R Core Team (2013).	Mac: https://cran.r-project.org/bin/macosx/ PC: https://cran.r-project.org/bin/windows/base/old/
macQIIME	[32]	http://www.wernerlab.org/software/macqiime/download
Other		
NNLS Method for assessing community contribution in a mix	This paper,	Method S1

Figure 1

Figure 2

Figure 3

Figure S1: The relationship between the difference in composition of single communities from the mixtures and the difference in gas production of single communities from the mixture related to Figures 2A, 2B.. The Y-axis shows the difference in gas production between the mean of three replicates of each individual community and the mean gas production of the mixes (the smaller the value, the more similar the gas production to the mixtures). The X-axis shows the mean unweighted UniFrac distance between individual communities and each of the ten mixtures (the smaller the value, the more similar the community composition to the mixtures). These two variables are positively correlated (Spearman $\rho = 0.86$, $P < 0.001$)). Qualitatively similar results were obtained using weighted UniFrac distances (Spearman $\rho = 0.75$, $P < 0.02$).
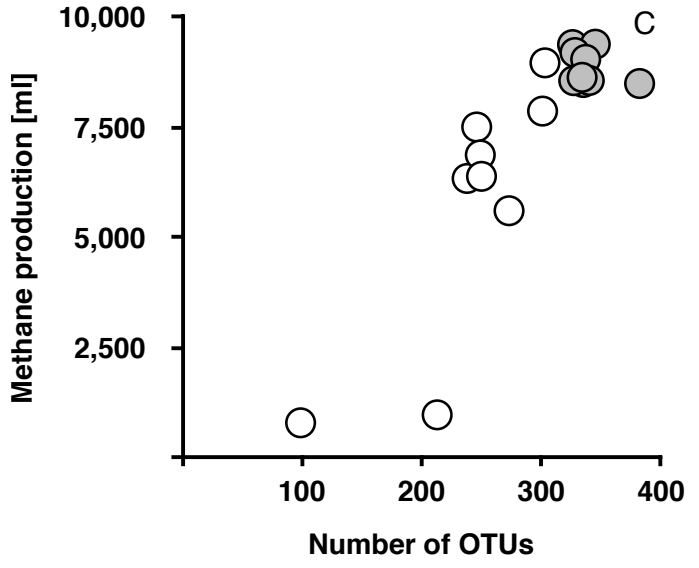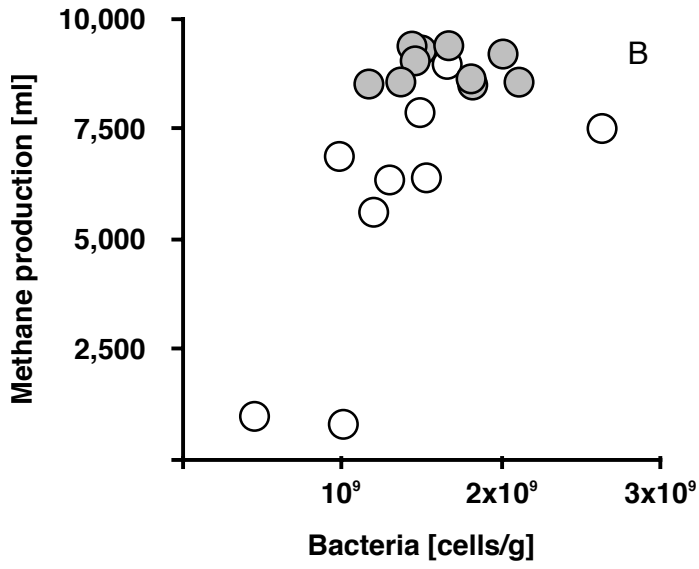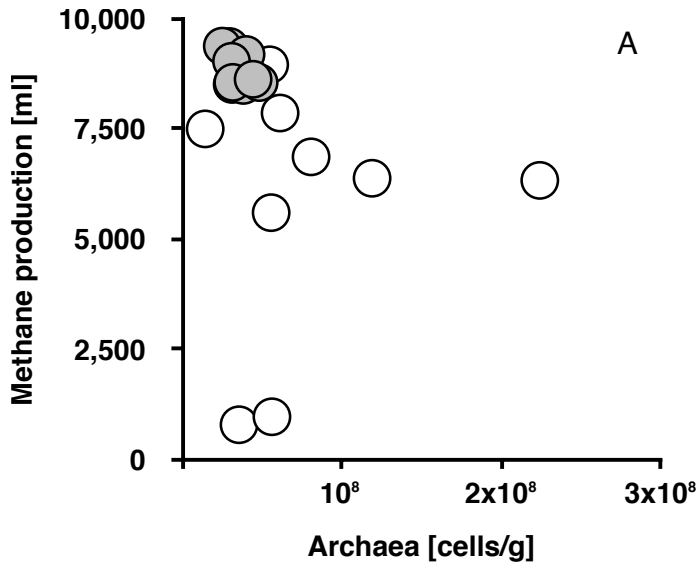
Figure S2: Within-community predictors of methane production from Experiment 2 related to Figure 3D, 3E. Relationships between methane production [ml] and: A) Archaeal densities [$\log_{10}$ cells/g] (Regression: $F_{1,15} = 0.32$, $P > 0.2$); B) Bacterial densities [$\log_{10}$ cells/g] (Regression: $F_{1,16} = 16.5$, $P < 0.001$); and C) number of OTUs (Regression: $F_{1,16} = 51.6$, $P < 0.001$). The reported statistics are based on combined mixed (grey circles) and individual (white circles) communities, but the same qualitative relationships were found when mixed communities were excluded form the analyses (Archaeal density: $F_{1,7} = 0.07$, $P > 0.2$; Bacterial density: $F_{1,7} = 92$, $P < 0.02$; OTUs: $F_{1,7} = 16.4$, $P < 0.01$).

| Sample consisting of community(ies)	No. of communities in the mix	Gas production after 4 weeks	Gas production of the best component of the mix after 4 weeks	Average gas production of all components of a mix after 4 weeks	Difference in gas production between a mix and its best component	Difference in gas production between a mix and the average of its components
P06	1	494.9	N/A	N/A	N/A	N/A
P08	1	584.3	N/A	N/A	N/A	N/A
P09	1	600.8	N/A	N/A	N/A	N/A
P12	1	1626.2	N/A	N/A	N/A	N/A
P11	1	2262.8	N/A	N/A	N/A	N/A
P03	1	2776.8	N/A	N/A	N/A	N/A
P01	1	2935.5	N/A	N/A	N/A	N/A
P02	1	3640.8	N/A	N/A	N/A	N/A
P05	1	4243.0	N/A	N/A	N/A	N/A
P10	1	5490.1	N/A	N/A	N/A	N/A
P04	1	5891.3	N/A	N/A	N/A	N/A
P13	1	6060.3	N/A	N/A	N/A	N/A
P01+P11	2	4231.70	2935.50	2599.15	1296.20	1632.55
P02+P09	2	5018.40	3640.80	2120.80	1377.60	2897.60
P12+P03	2	604.70	2776.80	2201.50	-2172.10	-1596.80
P08+P05	2	4383.00	4243.00	2413.65	140.00	1969.35
P04+P06	2	5543.40	5891.30	3193.10	-347.90	2350.30
P13+P10	2	6081.40	6060.30	5151.65	21.10	929.75
P03+P12+P01	3	3276.30	2935.50	2446.17	340.80	830.13
P08+P02+P06	3	4980.60	3640.80	1573.33	1339.80	3407.27
P04+P11+P13	3	6429.10	6060.30	4738.13	368.80	1690.97
P05+P09+P10	3	4894.80	5490.10	3444.63	-595.30	1450.17
P01+P13+P10+P05	4	5834.60	6060.30	4682.23	-225.70	1152.38
P02+P04+P11+P06	4	6401.90	5891.30	3072.45	510.60	3329.45
P03+P08+P09+P12	4	1086.50	2776.80	1397.03	-1690.30	-310.53
P12+P02+P01+P09	6	5065.90	4243.00	2271.77	822.90	2794.13

+P08+P05					
P03+P04+P06+P10 +P11+P13	6	5689.20	6060.30	3829.37	-371.10	1859.83
All 12 communities' mix	12	6620.00	6060.30	3050.57	559.70	3569.43
AVERAGE FOR ALL MIXES		4758.8	4672.9	3011.6	85.9	1747.2

Table S1: Community mixing setup and detailed results of Experiment 3 related to Figure 2C. The details of the communities used can be found in Table 1.

Method S1:

```
################################################################
## Solving a system of linear equations for a non-square [m > n]   ##
## matrix via the use of non-negative least squares [NNLS].      ##
## Return the solution 'weights' and residual sum of squares.     ##
##                              ##
## Mark Alston, Earlham Institute: mark.alston@earlham.ac.uk      ##
################################################################


## ======================================================================================= ##
## Install the R packages phyloseq, biomformat, nnls and limSolve.
## Now load as required...
## ======================================================================================= ##

library("phyloseq")
packageVersion("phyloseq")

library("biomformat")
packageVersion("biomformat")

library("nnls")
packageVersion("nnls")


## ======================================================================================= ##
## READ IN THE OTU table
## ======================================================================================= ##
## MAPPING FILE TO BE USED: 'mappingFile.txt'
## OTU TABLE TO BE USED:   'CSS_norm_rawValues.biom'

## SET YOUR WORKING DIRECTORY, e.g. point to where your files are located
setwd("/path/to/the/data")


## Quick look at the OTU table which is in biom format
read_biom("CSS_norm_rawValues.biom")


## load the biom file into phyloseq
data = import_biom("CSS_norm_rawValues.biom")
data


## ======================================================================================= ##
## Give meaningful names to the Taxon Column Headers and create a phyloseq object
## ======================================================================================= ##
myTaxTable <- tax_table(data)
colnames(myTaxTable) <- c("Kingdom","Phylum", "Class", "Order", "Family", "Genus","Species")

### take a look ###
head(myTaxTable)
rank_names(myTaxTable)

OTU = otu_table(data)
TAX = tax_table(myTaxTable)

myPhyloSeq_allData <- phyloseq(OTU,TAX)
myPhyloSeq_allData


## ======================================================================================= ##
## Incorporate some metadata about the samples ### i.e. first create a mappingFile.txt in a text editor
## ======================================================================================= ##
SAMPLES = import_qiime_sample_data("mappingFile.txt")
class(SAMPLES)


## ======================================================================================= ##
## MERGE the bits and bobs into a new phyloseq object
## ======================================================================================= ##
myPhyloSeq <- merge_phyloseq(myPhyloSeq_allData, SAMPLES)
myPhyloSeq


## inspect the OTU table
otu_table(myPhyloSeq)

## Checking the column headers and looking at the mapping file we see that
## columns 3,2,1,6,18,17,4,7,5 correspond to the 9 SINGLE communities [SAM11640-SAM11649]
## columns 11,12,10,19,15,16,13,14,8,9 correspond to the 10 MIXED communities [SAM11630-SAM11639]


## ======================================================================================= ##
## collapse OTU table at the 'family' level, and generate matrix 'A' [A.X = b]
## ======================================================================================= ##

bacteria_family <- tax_glom(myPhyloSeq, taxrank="Family")
bacteria_family_df <- as.data.frame(get_taxa(otu_table(bacteria_family))  )
bacteria_family_singleComm_df  <-  bacteria_family_df[,c(3,2,1,6,18,17,4,7,5)]

row.names(bacteria_family_singleComm_df) <- NULL
colnames(bacteria_family_singleComm_df) <- NULL
```

```
## ================================================================================================= ##
## get matrix 'A'
## ================================================================================================= ##
bacteria_family_matrix_A <- as.matrix(bacteria_family_singleComm_df)



## ================================================================================================= ##
## get vector 'b' [A.X = b], one for each of the 'mixed' samples [OTU table columns 11,12,10,19,15,16,13,14,8,9]
## ================================================================================================= ##
b_M01_bac_family_df  <- bacteria_family_df[,c(11)]
bf_M01  <-  as.matrix(b_M01_bac_family_df)

b_M02_bac_family_df  <- bacteria_family_df[,c(12)]
bf_M02  <-  as.matrix(b_M02_bac_family_df)

b_M03_bac_family_df  <- bacteria_family_df[,c(10)]
bf_M03  <-  as.matrix(b_M03_bac_family_df)

b_M04_bac_family_df  <- bacteria_family_df[,c(19)]
bf_M04  <-  as.matrix(b_M04_bac_family_df)

b_M05_bac_family_df  <- bacteria_family_df[,c(15)]
bf_M05  <-  as.matrix(b_M05_bac_family_df)

b_M06_bac_family_df  <- bacteria_family_df[,c(16)]
bf_M06  <-  as.matrix(b_M06_bac_family_df)

b_M07_bac_family_df  <- bacteria_family_df[,c(13)]
bf_M07  <-  as.matrix(b_M07_bac_family_df)

b_M08_bac_family_df  <- bacteria_family_df[,c(14)]
bf_M08  <-  as.matrix(b_M08_bac_family_df)

b_M09_bac_family_df  <- bacteria_family_df[,c(8)]
bf_M09  <-  as.matrix(b_M09_bac_family_df)

b_M10_bac_family_df  <- bacteria_family_df[,c(9)]
bf_M10  <-  as.matrix(b_M10_bac_family_df)



## ================================================================================================= ##
## 'weights' can be negative
## So try to solve giving only NON-NEGATIVE weights via non-negative least-squares (NNLS)
## see: 'nnls'  from  https://cran.r-project.org/web/packages/nnls/nnls.pdf
## ================================================================================================= ##


## ================================================================================================= ##
## FIRST: Inspect the 'Residual Sum of Squares' [RSS] values
## ================================================================================================= ##

soln_M01 <- nnls(bacteria_family_matrix_A,bf_M01)
soln_M02 <- nnls(bacteria_family_matrix_A,bf_M02)
soln_M03 <- nnls(bacteria_family_matrix_A,bf_M03)
soln_M04 <- nnls(bacteria_family_matrix_A,bf_M04)
soln_M05 <- nnls(bacteria_family_matrix_A,bf_M05)
soln_M06 <- nnls(bacteria_family_matrix_A,bf_M06)
soln_M07 <- nnls(bacteria_family_matrix_A,bf_M07)
soln_M08 <- nnls(bacteria_family_matrix_A,bf_M08)
soln_M09 <- nnls(bacteria_family_matrix_A,bf_M09)
soln_M10 <- nnls(bacteria_family_matrix_A,bf_M10)

solution <- cbind(soln_M01, soln_M02, soln_M03, soln_M04, soln_M05, soln_M06, soln_M07, soln_M08, soln_M09, soln_M10)

solution[2,]          ### Row 2 holds the values for the 'deviance', or 'Residual Sum of Squares' [RSS] values,
          ### that is the 'distance' of the solution vector from the projected vector, b



## ================================================================================================= ##
## SECOND: Grab the solution 'weight' values, or 'X' from [A.X = b]
## pass matrix 'A' and each vector 'b' in turn to the NNLS function
## ================================================================================================= ##

library("limSolve")
packageVersion("limSolve")

soln_M01 <- nnls(bacteria_family_matrix_A,bf_M01, verbose = TRUE)
soln_M02 <- nnls(bacteria_family_matrix_A,bf_M02, verbose = TRUE)
soln_M03 <- nnls(bacteria_family_matrix_A,bf_M03, verbose = TRUE)
soln_M04 <- nnls(bacteria_family_matrix_A,bf_M04, verbose = TRUE)
soln_M05 <- nnls(bacteria_family_matrix_A,bf_M05, verbose = TRUE)
soln_M06 <- nnls(bacteria_family_matrix_A,bf_M06, verbose = TRUE)
soln_M07 <- nnls(bacteria_family_matrix_A,bf_M07, verbose = TRUE)
soln_M08 <- nnls(bacteria_family_matrix_A,bf_M08, verbose = TRUE)
soln_M09 <- nnls(bacteria_family_matrix_A,bf_M09, verbose = TRUE)
soln_M10 <- nnls(bacteria_family_matrix_A,bf_M10, verbose = TRUE)

solution <- cbind(soln_M01$X, soln_M02$X, soln_M03$X, soln_M04$X, soln_M05$X, soln_M06$X, soln_M07$X, soln_M08$X, soln_M09$X, soln_M10$X)


## name columns and rows ##
```

```r
#########################
dimnames(solution) = list( c("P1","P4","P5","P8","P9","P10","P12","P13","P15"),c("M1","M2","M3","M4","M5","M6","M7","M8","M9","M10") )

solution  ## view the solution 'weights' for each mixed sample

write.table(solution, sep="\t", "solutionWeights.txt")


## =============================================================================================== ##
## THIRD: plot out the 'weights'
## Plot the weight of contribution for each of the single 'seed' samples to a mixture
## =============================================================================================== ##


## view weights for each mixed sample as barcharts  ##
## munge solution vectors into one vector          ##
## the following was adapted from: http://www.r-bloggers.com/using-the-svd-to-find-the-needle-in-the-haystack/ ##


library(lattice)
b_clr <- c("steelblue", "darkred")
b1 <- barchart(as.table(solution[,1]),
          main="M_01",
          horizontal=FALSE, col=ifelse(solution[,1] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b1, split=c(1,1,3,4), more=TRUE)    ### 'split' is used to lay out the barchart lattice
                                               ### e.g. split=c(1,1,3,2) means place plot b1 in col. 1, row 1

b2 <- barchart(as.table(solution[,2]),
          main="M_02",
          horizontal=FALSE, col=ifelse(solution[,2] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b2, split=c(2,1,3,4), more=TRUE)             ### 'split' is used to lay out the barchart lattice
                                                        ### e.g. split=c(2,1,3,2) means place plot b2  in col. 2, row 1
                                                        ### where the layout has 3 columns and 2 rows       x O x
                                                        ###                                                 x x x

b3 <- barchart(as.table(solution[,3]),
          main="M_03",
          horizontal=FALSE, col=ifelse(solution[,3] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b3, split=c(3,1,3,4), more=TRUE)


b4 <- barchart(as.table(solution[,4]),
          main="M_04",
          horizontal=FALSE, col=ifelse(solution[,4] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b4, split=c(1,2,3,4), more=TRUE)


b5 <- barchart(as.table(solution[,5]),
          main="M_05",
          horizontal=FALSE, col=ifelse(solution[,5] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b5, split=c(2,2,3,4), more=TRUE)


b6 <- barchart(as.table(solution[,6]),
          main="M_06",
          horizontal=FALSE, col=ifelse(solution[,6] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b6, split=c(3,2,3,4), more=TRUE)


b7 <- barchart(as.table(solution[,7]),
          main="M_07",
          horizontal=FALSE, col=ifelse(solution[,7] > 0,
                          b_clr[1], b_clr[2]),
          ylab="Weight",
          scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
          ) ###key = key)
     print(b7, split=c(1,3,3,4), more=TRUE)


b8 <- barchart(as.table(solution[,8]),
          main="M_08",
```

```
        horizontal=FALSE, col=ifelse(solution[,8] > 0,
                            b_clr[1], b_clr[2]),
        ylab="Weight",
        scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
        ) ###key = key)
    print(b8, split=c(2,3,3,4), more=TRUE)


b9 <- barchart(as.table(solution[,9]),
        main="M_09",
        horizontal=FALSE, col=ifelse(solution[,9] > 0,
                            b_clr[1], b_clr[2]),
        ylab="Weight",
        scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
        ) ###key = key)
    print(b9, split=c(3,3,3,4), more=TRUE)


b10 <- barchart(as.table(solution[,10]),
        main="M_10",
        horizontal=FALSE, col=ifelse(solution[,10] > 0,
                            b_clr[1], b_clr[2]),
        ylab="Weight",
        scales=list(x=list(rot=55, labels=rownames(solution), cex=1.1)),
        ) ###key = key)
    print(b10, split=c(1,4,3,4))
```

Method S1: **Code for the Non-Negative Least Squares (NNLS) analysis related to STAR Methods.** This annotated code can be run using R in order to obtain the contribution of individual communities towards the community mix. The input files needed are the .biom file containing the compositions of communities analysed and their phylogeny, preferably prepared using Cumulative Sum Scaling normalisation (see STAR methods). It also requires a mapping file as described in http://qiime.org/documentation/file_formats.html. The version of the code is suited for analysis of our dataset but can be readily adapted for any dataset containing multiple communities amplicon data.

The script was run in RStudio v1.0.143 using R version 3.4.0 (2017-04-21) and the following R package versions: limSolve_1.5.5.2, nnls_1.4, biomformat_1.4.0, Phyloseq_1.20.0.