Segal's Law, 16S rRNA gene sequencing, and the perils of foodborne pathogen detection within the American Gut Project

James B Pettengill [Corresp., 1] , **Hugh Rand** [1]

[1] Biostatistics and Bioinformatics Staff, Office of Analytics and Outreach, US Food and Drug Administration, College Park, Maryland, United States

Corresponding Author: James B Pettengill
Email address: fixtgear@gmail.com

Obtaining human population level estimates of the prevalence of foodborne pathogens is critical for understanding outbreaks and ameliorating such threats to public health. Estimates are difficult to obtain due to logistic and financial constraints, but citizen science initiatives like that of the American Gut Project (AGP) represent a potential source of information concerning enteric pathogens. With an emphasis on genera *Listeria* and *Salmonella*, we sought to document the prevalence of those two taxa within the AGP samples. The results provided by AGP suggest a surprising 14% and 2% of samples contained *Salmonella* and *Listeria*, respectively. However, a reanalysis of those AGP sequences described here indicated that results depend greatly on the algorithm for assigning taxonomy and differences persisted across both a range of parameter settings and different reference databases (i.e., Greengenes and HITdb). These results are perhaps to be expected given that AGP sequenced the V4 region of 16S rRNA gene, which may not provide good resolution at the lower taxonomic levels (e.g., species), but it was surprising how often methods differ in classifying reads – even at higher taxonomic ranks (e.g., family). This highlights the misleading conclusions that can be reached when relying on a single method that is not a gold standard; this is the essence of Segal's Law: an individual with one watch knows what time it is but an individual with two is never sure. Our results point to the need for an appropriate molecular marker for the taxonomic resolution of interest, and calls for the development of more conservative classification methods that are fit for purpose. Thus, with 16S rRNA gene datasets, one must be cautious regarding the detection of taxonomic groups of public health interest (e.g., culture independent identification of foodborne pathogens or taxa associated with a given phenotype).

1   Segal's Law, 16S rRNA gene sequencing, and the perils of foodborne pathogen detection within

2   the American Gut Project

4   James B Pettengill[1], Hugh Rand[1]

6   [1] Biostatistics and Bioinformatics Staff, Office of Analytics and Outreach, US Food and Drug

7   Administration, College Park, Maryland, United States

9   Running title: Segal's Law and 16S rRNA gene sequencing

11   Corresponding author: James B Pettengill

12 **Abstract**

13 Obtaining human population level estimates of the prevalence of foodborne pathogens is critical

14 for understanding outbreaks and ameliorating such threats to public health. Estimates are

15 difficult to obtain due to logistic and financial constraints, but citizen science initiatives like that

16 of the American Gut Project (AGP) represent a potential source of information concerning

17 enteric pathogens. With an emphasis on genera *Listeria* and *Salmonella*, we sought to document

18 the prevalence of those two taxa within the AGP samples. The results provided by AGP suggest

19 a surprising 14% and 2% of samples contained *Salmonella* and *Listeria*, respectively. However, a

20 reanalysis of those AGP sequences described here indicated that results depend greatly on the

21 algorithm for assigning taxonomy and differences persisted across both a range of parameter

22 settings and different reference databases (i.e., Greengenes and HITdb). These results are

23 perhaps to be expected given that AGP sequenced the V4 region of 16S rRNA gene, which may

24 not provide good resolution at the lower taxonomic levels (e.g., species), but it was surprising

25 how often methods differ in classifying reads – even at higher taxonomic ranks (e.g., family).

26 This highlights the misleading conclusions that can be reached when relying on a single method

27 that is not a gold standard; this is the essence of Segal's Law: an individual with one watch

28 knows what time it is but an individual with two is never sure. Our results point to the need for

29 an appropriate molecular marker for the taxonomic resolution of interest, and calls for the

30 development of more conservative classification methods that are fit for purpose. Thus, with 16S

31 rRNA gene datasets, one must be cautious regarding the detection of taxonomic groups of public

32 health interest (e.g., culture independent identification of foodborne pathogens or taxa associated

33 with a given phenotype).

34 **Introduction**

35   Obtaining estimates of the prevalence of foodborne pathogens is critical to public health

36 because the estimates affect policy, allocation of funding, and understanding of foodborne

37 outbreaks. Such figures are difficult to obtain but estimates do exist based on surveillance

38 programs. Estimates by the US Center for Disease Control of the number of cases in 2013 of

39 *Salmonella* and *Listeria* was 7,277 (incidence = 15.19) and 123 (incidence = 0.26), respectively,

40 per 100,000 individuals in the United States

41 (http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6315a3.htm). These numbers are based on

42 the active monitoring done through FoodNet (Foodborne Diseases Active Surveillance Network)

43 and come from laboratory-confirmed infections in 10 U.S. localities that account for about 15%

44 of the U.S. population. Such monitoring has provided extremely valuable data to help evaluate

45 prevention measures, understand the causes and transmission of food-borne diseases, and set

46 targets for improving the domestic food supply (Centers for Disease & Prevention 2000;

47 Watstein & Jovanovic 2003). However, monitoring of the type done by FoodNet is not without

48 caveats, chief among them are the underreporting of cases due to infected individuals not visiting

49 local health facilities when sickened with a foodborne pathogen.

50   A complementary approach to large-scale surveillance networks managed by state and

51 federal agencies is the solicitation of the general population to participate in a study or data

52 collection initiative; such information may provide an independent estimate of the prevalence of

53 enteric pathogens such as *Listeria* and *Salmonella*. One such initiative is the American Gut

54 Project (AGP; www.americangut.org), which describes itself as "…one of the world's largest

55 crowd sourced, citizen science projects in the country [USA]. We discover new information

56 daily to shed light on the connection between the human microbiome and health." The AGP has

57 performed sequencing of the V4 region of the 16S rRNA gene from thousands of individual

58 participant's skin, mouth, and/or feces. Such an approach is not without caveats, however. Chief

59 among them are the potential bias introduced by the non-random participation by citizens (i.e.,

60 convenience sampling), the computational and bioinformatic challenges associated with

61 analyzing the large amount of DNA sequence data, and the potential for misinterpretation

62 (Gonzalez et al. 2016). Nevertheless, the data that is produced may be quite informative about

63 public health issues related to foodborne pathogens by detecting carriers and asymptomatic cases

64 that hospital-centric sampling schemes would miss.

65   Here we describe our efforts to use the data produced by the AGP to determine the

66   prevalence of *Listeria* and *Salmonella* – two genera with species of public health interest as

67   causative agents of gastroenteritidis (or worse such as death and spontaneous abortion in

68   women). Given the rich metadata associated with the human subjects that participated in the

69   AGP, we were particularly interested in determining life style habits or other factors that

70   correlate strongly with individuals that are positive for particular pathogens; we were encouraged

71   by the publicly available results provided by the AGP that showed 14% and 2% of samples did

72   contain either *Salmonella* or *Listeria*, respectively (tp://ftp.microbio.me/AmericanGut//ag-July-

73   29-2016//07-taxa/notrim/otu_table_fecal_L6.biom). However, in attempting to validate those

74   results we found that taxonomic classification of reads varied greatly depending on the algorithm

75   and reference database used. Our objective here is to describe those additional analyses and

76   discuss the implications of our results for using such data for public health surveillance.

77

78   **Materials and Methods**

79   *American Gut Project Data*

80   American Gut Project data as of August 1, 2016 were analyzed. Given that our original

81   objective was to determine the percentage of samples that contained either *Salmonella* or *Listeria*

82   we queried the QIIME output files (i.e., ftp://ftp.microbio.me/AmericanGut//ag-July-29-2016/07-

83   taxa/notrim/ag-cleaned_L6.txt) provided by AGP for those taxa. The AGP data contained 10,294

84   fecal, skin, or oral samples, and yielded 188 samples positive for *Listeria* and 1495 positive for

85   *Salmonella*; 31 samples were positive for both. This yielded 1652 samples that were carried

86   forward and analyzed here to determine the robustness of these classifications across multiple

87   taxonomy assigners. The sequences from each of the 1652 were extracted from the fasta file

88   available from AGP (ftp://ftp.microbio.me/AmericanGut//ag-July-29-2016/02-

89   [filtered/sequences-notrim.fna](ftp://ftp.microbio.me/AmericanGut//ag-July-29-2016/02-filtered/sequences-notrim.fna)), which resulted in a 48,312,131 sequences to be classified. No

90   additional filtering of reads was done.

91

92   *Taxonomic assigners*

93   For each of the 1652 samples, we classified the reads associated with them to taxonomy

94   using four different classifiers – RDP (Wang et al. 2007), UCLUST (Edgar 2010), BLAST

95   (Altschul et al. 1990), and SortMeRNA (Kopylova et al. 2014) (Table 1). SortMeRNA was the

96   method used to produce the AGP results. The program QIIME v1.9.1 (Caporaso et al. 2010) was

97   used to run each of these classifiers. These assigners do not represent an exhaustive list of those

98   available and were chosen as they are among the most frequently used; additional classifiers are

99   likely to only increase the variability that we observe and, therefore, not change our findings.

100 Additionally, we make no statements about the accuracy of the methods as we do not know the

101 true taxonomy of the reads but rather are estimating the variability in classification across the

102 methods, which illustrates the degree to which conclusions about pathogen presence is dependent

103 on the classifier used.

104   Within each classifier a range of cutoffs were used for the parameter that measures

105 confidence of taxonomic assignment (Table 1). The cutoff settings for these parameters are

106 specific to each assigner and were determined through exploratory analyses that evaluated which

107 cutoff values result in a substantial difference in number of reads classified. An example is

108 UCLUST, which has a `min_consensus_fraction` parameter, which determines the

109 fraction of hits that must agree in the number of assignments considered (`max_accepts`). With

110 `max_accepts` set to 3 (the default), all that matters is whether

111 `min_consensus_fraction` is more or less than 2/3. Thus, we evaluated two settings 1)

112 `min_consensus_fraction` = 0.51 (the default, at least 2/3 agreement is required) and 2)

113   `min_consensus_fraction` = 0.68 (complete agreement is required). The latter threshold

114 is more conservative (higher specificity) but also runs the risk of a higher false-negative rate

115 (lower sensitivity).

116

117 *Reference databases*

118   Analyses were performed against two different reference databases. The first database was

119 the Greengenes (gg_13_8) database (DeSantis et al. 2006). This database was used to produce

120 the AGP results and we primarily focus on the results using this database. Within the Greengenes

121 database there are 99,322 sequences of which 4 are classified to *Salmonella* (1 to only genus and

122 3 to species *enterica*) and 5 sequences classified to *Listeria* (1 to each of the following species

123 *monocytogenes, seeligeri, grayi, fleischmannii, weihenstephanensis*).

124   The second database was the human microbiome specific reference database HITdb

125 described in Ritari et al. (2015). HITdb includes many fewer sequences than Greengenes (n =

126 2,472) of which a single sequence is classified to *Salmonella enterica* and zero are classified to

127 the genus *Listeria*. The fewer number of representatives at this taxonomic level for these

128 pathogens, and likely others, may be a desirable property as it could avoid false-positives.

129 However, such a smaller database relative to Greengenes may run the risk of being too restrictive

130 where a read could be reliably classified at one of those taxa given a sufficient algorithm. The

131 taxonomy of within HITdb begins at the phylum level and, therefore, we do not present

132 congruence among assigners at that kingdom level when using HITdb.

133

134 *Congruence between classifiers*

135   The results from each of the classifiers with the various parameter settings were combined in

136 a pairwise fashion. To evaluate the congruence between the classifiers, we calculated the percent

137 of total reads within each of five different comparison categories (1) match: both classifiers

138 assigned the read to the same taxon, (2) mismatch: the classifiers differed in the taxonomy

139 assigned to a read, (3 & 4) method 1 unclassified and method 2 unclassified: cases in which one

140 classifier did not assign the read to a taxon but the other method did, and (5) noMatch: not

141 classified by either classifier. The small fraction of the reads in the noMatch category (~0.8% for

142 the default parameter settings using Greengenes) were not analyzed. As these categories suggest,

143 we include reads classified to the species level and reads partially classified (e.g., the lowest

144 taxonomic rank assigned to a read was only phylum or family). We also explored more explicit

145 methods for determining congruence between classifiers (i.e., inter-rater agreement; e.g.,

146 Cohen's κ (Cohen 1960)). Such statistics attempt to account for the fact that two classifiers could

147 agree by chance, which percent of reads that match (like that used here) does not. However,

148 preliminary analyses showed these statistics to be difficult to interpret as a result of the large

149 number of observations and at times small number of categories (e.g., at the kingdom or phylum

150 level). For example, at higher taxonomic ranks there is clearly better agreement between the

151 classifiers (see Results) but prevalence and bias adjusted Cohen's κ (PABAK) was substantially

152 less than that observed at the species level (Supplemental Material S1).

153   As an example of how the incongruence among classifiers impacts the detection of a

154 specific taxon, we focused on the number of reads under each method that were classified to the

155 genera *Salmonella* and *Listeria* and quantified differences between classifiers. This was done for

156 only the results produced with Greengenes. The congruence between methods in which reads

157 were classified to *Salmonella* and *Listeria* was evaluated using the Jaccard Similarity index (0

158   meaning two methods classified completely different reads to a given taxon; 1 denotes complete

159   overlap in the reads classified to a given taxon).

160

161   **Results**

162   *Default settings and Greengenes database*

163   The initial investigation into taxonomy assignments provided by the AGP indicated 188

164   samples positive for *Listeria,* 1495 samples positive for *Salmonella*, and 31 samples positive for

165   both, yielding 1652 samples. Taxonomic resolution was only to genus so the AGP results do not

166   include whether pathogenic *Listeria* (e.g., *Listeria monocytogenes*) was present. To determine

167   the robustness of these classifications we analyzed the sequences from those samples using

168   default parameter settings and the Greengenes database (similar to the AGP analyses) and found

169   little congruence among the four classifiers investigated in the assignment of reads to lower

170   taxonomic ranks. For example, in all but the highest taxonomic rank (Kingdom), there was an

171   average difference of approximately 30% in the number of unique taxa detected (Fig. 1). The

172   pattern of difference in number of taxa detected was also not consistent across the ranks where,

173   for instance, the RDP classifier detected more taxa than SortMeRNA at the genus and species

174   level but not at the phylum, class, order or family level (Fig. 1).

175   Focusing on congruence in the taxonomic identity classified to reads, we found that

176   congruence among the classifiers decreased with decreasing taxonomic rank (Fig. 2). This

177   suggests that it is more difficult and there is greater error associated with classifying reads to

178   lower taxonomic ranks, which has also been observed in other studies (Mizrahi-Man et al. 2013;

179   Wang et al. 2007). This incongruence was primarily driven by a read being classified to a

180   taxonomic group with one method but not the other rather than each method classifying the same

181   read to different taxonomic groups (Fig. 2). Certain methods also seemed to result in greater

182   incongruence. For example, RDP classified many more reads at the genus and species levels than

183   the other methods (Fig. 2).

184   When we focused on the congruence across the classifiers in assigning reads to *Salmonella*

185   or *Listeria*, we found that 447 and 1521 samples contained reads classified to *Listeria* or

186   *Salmonella*, respectively. The number of *Listeria* 'positive' samples reported by AGP was 188

187   and for *Salmonella* it was 1495 (Table 2). However, the different methods varied substantially in

188   the number of reads classified to each genus with BLAST consistently being the most liberal

189 (i.e., classified the most reads) followed by UCLUST; RDP was the most conservative (i.e., did

190 not classify reads) (Table 2). Perhaps of more importance is that there was little congruence

191 between the methods in which reads were classified to each taxonomic group as indicated by the

192 low Jaccard similarity indices between the pairwise comparisons of the methods (Fig. 3). The

193 metric suggests there is little consistency among the methods when classifying *Listeria* or

194 *Salmonella* (genus or species) to the reads. For example, only 22% of the reads classified by

195 UCLUST or BLAST to a species within *Salmonella* were actually classified to the same species

196 – some of those reads may not have been classified at all by one method while others could have

197 been classified to a different species (Fig. 3).

198   Some of these differences in classifications could likely be reconciled by changing the

199 confidence parameter. For example, with the default settings, read 10317.000001040_2356 was

200 clustered by UCLUST and BLAST to the genus *Salmonella,* by RDP to the species *Salmonella*

201 *enterica,* and by SortMeRNA to the family Enterobacteriaceae. If one reduced the confidence

202 setting for UCLUST, BLAST and SortMeRNA then the methods might converge on the

203 classification of *Salmonella enterica*. Alternatively, the confidence settings for UCLUST,

204 BLAST, and RDP could be increased so the methods would converge on the same classification

205 at the family level. However, looking at read 10317.000001127_16840, it was classified to the

206 genus *Enterococcus* within the order Lactobacillales by SortMeRNA and UCLUST, only to the

207 class level (Bacilli) with RDP (even with the low default confidence threshold of 0.5), and

208 *Listeria weihenstephanensis* within the order Bacillales by BLAST. This read, which is

209 representative of many of the reads classified to *Listeria* by at least one method, cannot be

210 congruently classified below the rank of class. Although these examples illustrate the ways in

211 which differences in classification may be reconciled, this would be problematic for an entire set

212 of reads (e.g., each read may require its own threshold settings and reads may be 'under'

213 classified if one takes a lowest common ancestor approach).

214

215 *Impact of varying confidence threshold and reference database*

216   The level of congruence among the classifiers varied substantially depending on the

217 parameter settings that each used to determine whether a read was classified to a taxon.

218 However, no parameter combinations produced complete congruence between two classifiers

219 and, in general, the default settings resulted in the highest congruence among the classifiers

220 (fewest number of reads classified by one method but not the other and highest number of reads

221 being classified to the same taxon (i.e., match; Figs. 2 & 4; Supplemental Material S2). An

222 example of classification at the Family rank with the RDP classifier and the Greengenes database

223 illustrates this. As the cutoff for the confidence parameter for RDP is increased from the default,

224 a greater number of reads were not classified; thus, fewer reads were classified to the same taxon

225 (Fig. 4a & 4b). This suggests that a higher threshold setting within RDP has the unwanted

226 consequence of producing many false negative assignments (a read is not classified to a taxon

227 when it should be).

228   The pattern that different threshold parameter cutoffs affected the degree of congruence

229 among the assigners and that no combination of threshold settings resulted in complete

230 congruence was also observed when classifying reads based on the HITdb database (Fig. 4;

231 Supplemental Material S2).

232

233 **Conclusions**

234   Citizen science projects, like the American Gut Project, represent a great opportunity to

235 further public health. We had initially thought to use the publicly available data, specifically the

236 assignment of reads to known foodborne pathogens and the rich metadata about each sample, to

237 generate hypotheses about the prevalence of such pathogens within the general population and

238 examine potential lifestyle and demographic correlates with carrier status. The results of

239 taxonomic assignment to each of the reads within the AGP showed this to be a promising avenue

240 to pursue, as ~15% of the samples were 'positive' for *Listeria* or *Salmonella*. Before making

241 claims about these pathogens that would be relevant to public health we sought to evaluate the

242 robustness of those classifications by re-analyzing the positive data and evaluating the taxonomy

243 assignments using additional classifiers. Although we do not know the true taxonomy of the

244 reads being classified, the results of those additional analyses that showed many reads being

245 classified differently by the classifiers raise questions about how reliable those classifications

246 are. Those results also caused us to forego any analyses looking into correlates in the metadata

247 that may explain the presence of *Listeria* or *Salmonella* within an individual's skin, mouth,

248 and/or fecal microbiome. The incongruence we observed among the methods (i.e., Segal's Law)

249 is due, in part, to the different underlying assumptions and algorithms employed in each

250 classifier. It is also partly due to the V4 16S rRNA marker potentially lacking the information

251   content (especially at lower taxonomic ranks like species) to be robust to those differences in

252   algorithms and assumptions. Interestingly, the disagreement between methods cannot be

253   removed or made insignificant by adjustments to the confidence threshold or by using a database

254   designed for human enterobacteria; we explored this possibility, and there were always an

255   appreciable number of reads with either mismatched taxon classification or no classification for

256   one of the methods (Fig. 4).

257   Our results are in agreement with those published elsewhere on the negative relationship

258   between accuracy in classifying V4 reads and decreasing taxonomic rank (Clarridge 2004; Liu et

259   al. 2008; Srinivasan et al. 2015). However, the V4 region has been shown to best capture the

260   information content contained in the full 16S rRNA gene and to have the best sensitivity for

261   bacterial and phylogenetic analyses (Wang et al. 2007; Yang et al. 2016). Our results are

262   reassuring in that read classification matches greatly outnumber mismatches (Figs. 2 & 4).

263   However, there are a large number of instances in which a read is classified by one method but

264   not another (Figs. 2 & 4); at the family level, 5% of reads were classified by one method but not

265   the other and this increases to 16% and 40% at the genus and species levels, respectively. These

266   problems make it difficult to use this data for research in *Salmonella* and *Listeria*, and certainly

267   detract from its use in a regulatory setting. Additionally, the results presented here showing that

268   even at the family level there can be 5% of reads classified by one method and not the other

269   show that conclusions within 16S rRNA gene studies and, likely, metagenome-wide association

270   studies (Wang & Jia 2016) will differ depending on the classifier being employed. Public health

271   policy should be set based on good empirical information, and misleading pathogen detection is

272   problematic. This highlights the difficulty of associating pathogens with a given phenotype (e.g.,

273   obesity; Sze & Schloss 2016) with the 16S rRNA gene V4 region using current taxonomic

274   assigners, and points toward problems that have been remarked upon elsewhere regarding the

275   publication of results obtained with a method that is not fit for purpose (e.g., Gonzalez et al.

276   2016). There is a clear need for improved methods that provide measures of taxonomic

277   assignment uncertainty and can support good public health policy decision making.

278

279

280   **Acknowledgements**

281   We thank A. Shpuntoff for assistance with conducting the analyses on the high performance

282   computing cluster and FDA Center for Food Safety and Applied Nutrition for supporting this

283   research. We also thank M. Ferguson for insightful discussion on methods for assessing inter-

284   rater agreement.

285   References

286

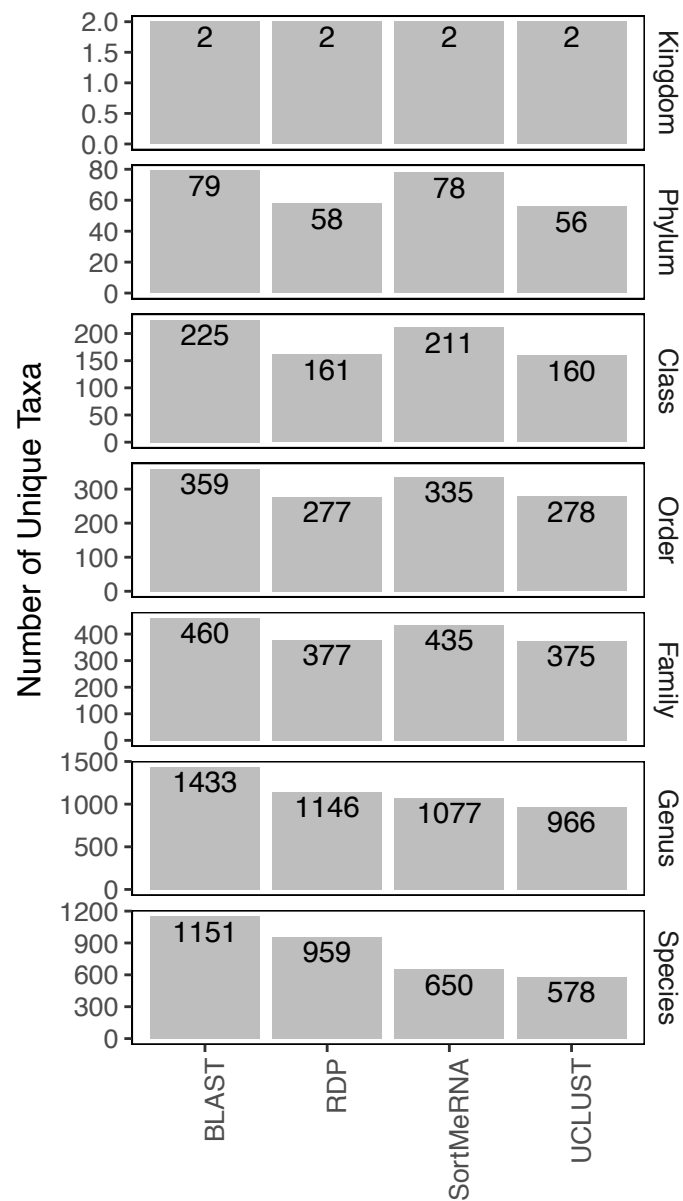
287 Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic Local Alignment
288   Search Tool. *J Mol Biol* 215:403-410.
289 Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena
290   AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE,
291   Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Tumbaugh
292   PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, and Knight R. 2010. QIIME
293   allows analysis of high-throughput community sequencing data. *Nature Methods* 7:335-
294   336.
295 Centers for Disease C, and Prevention. 2000. Preliminary FoodNet data on the incidence of
296   foodborne illnesses--selected sites, United States, 1999. *MMWR Morb Mortal Wkly Rep*
297   49:201-205.
298 Clarridge JE, 3rd. 2004. Impact of 16S rRNA gene sequence analysis for identification of
299   bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 17:840-862,
300   table of contents.
301 Cohen J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological*
302   *Measurement* 20:37-46.
303 DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P,
304   and Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and
305   workbench compatible with ARB. *Appl Environ Microbiol* 72:5069-5072.
306 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
307   26:2460-2461.
308 Gonzalez A, Vazquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, and Knight R. 2016.
309   Avoiding Pandemic Fears in the Subway and Conquering the Platypus. *mSystems* 1.
310 Kopylova E, Noé L, Pericard P, Salson M, and Touzet H. 2014. SortMeRNA 2: ribosomal RNA
311   classification for taxonomic assignation. Workshop on Recent Computational Advances
312   in Metagenomics, ECCB 2014.
313 Liu ZZ, DeSantis TZ, Andersen GL, and Knight R. 2008. Accurate taxonomy assignments from
314   16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids*
315   *Research* 36.
316 Mizrahi-Man O, Davenport ER, and Gilad Y. 2013. Taxonomic Classification of Bacterial 16S
317   rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs.
318   *Plos One* 8.
319 Ritari J, Salojarvi J, Lahti L, and de Vos WM. 2015. Improved taxonomic assignment of human
320   intestinal 16S rRNA sequences by a dedicated reference database. *BMC Genomics*
321   16:1056.
322 Srinivasan R, Karaoz U, Volegova M, MacKichan J, Kato-Maeda M, Miller S, Nadarajan R,
323   Brodie EL, and Lynch SV. 2015. Use of 16S rRNA Gene for Identification of a Broad
324   Range of Clinically Relevant Bacterial Pathogens. *PLoS One* 10.
325 Sze MA, and Schloss PD. 2016. Looking for a Signal in the Noise: Revisiting Obesity and the
326   Microbiome. *Mbio* 7.
327 Wang J, and Jia HJ. 2016. Metagenome-wide association studies: fine-mining the microbiome.
328   *Nature Reviews Microbiology* 14:508-522.

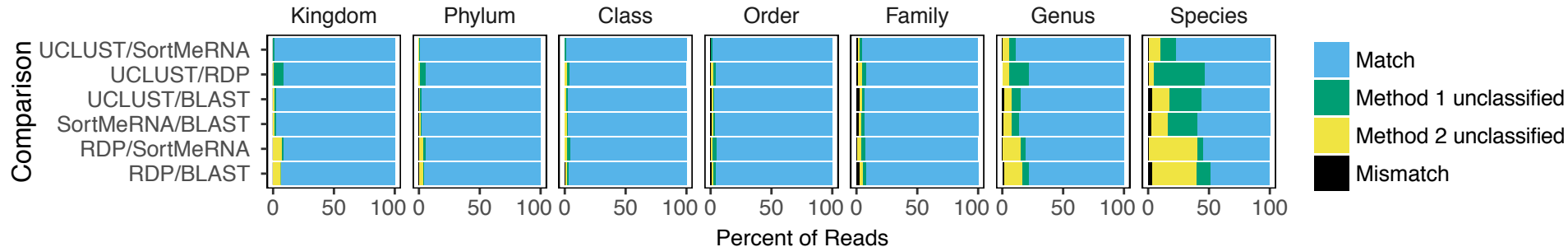
329   Wang Q, Garrity GM, Tiedje JM, and Cole JR. 2007. Naive Bayesian classifier for rapid
330   assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*
331   73:5261-5267.
332   Watstein S, and Jovanovic J. 2003. *Statistical Handbook on Infectious Diseases*: Greenwood
333   Publishing Group.
334   Yang B, Wang Y, and Qian PY. 2016. Sensitivity and correlation of hypervariable regions in
335   16S rRNA genes in phylogenetic analysis. *Bmc Bioinformatics* 17:135.
336

The number of unique taxa to which reads were assigned by each classification method.

The results are grouped by taxonomic rank.

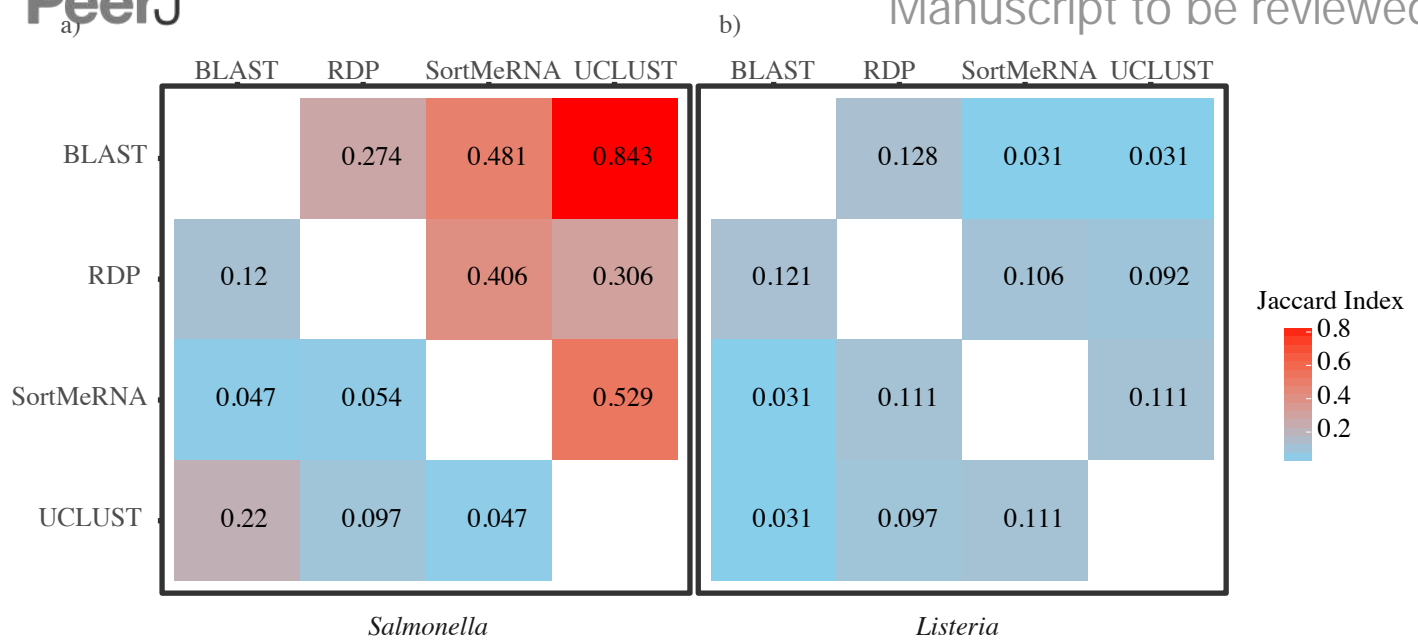
x

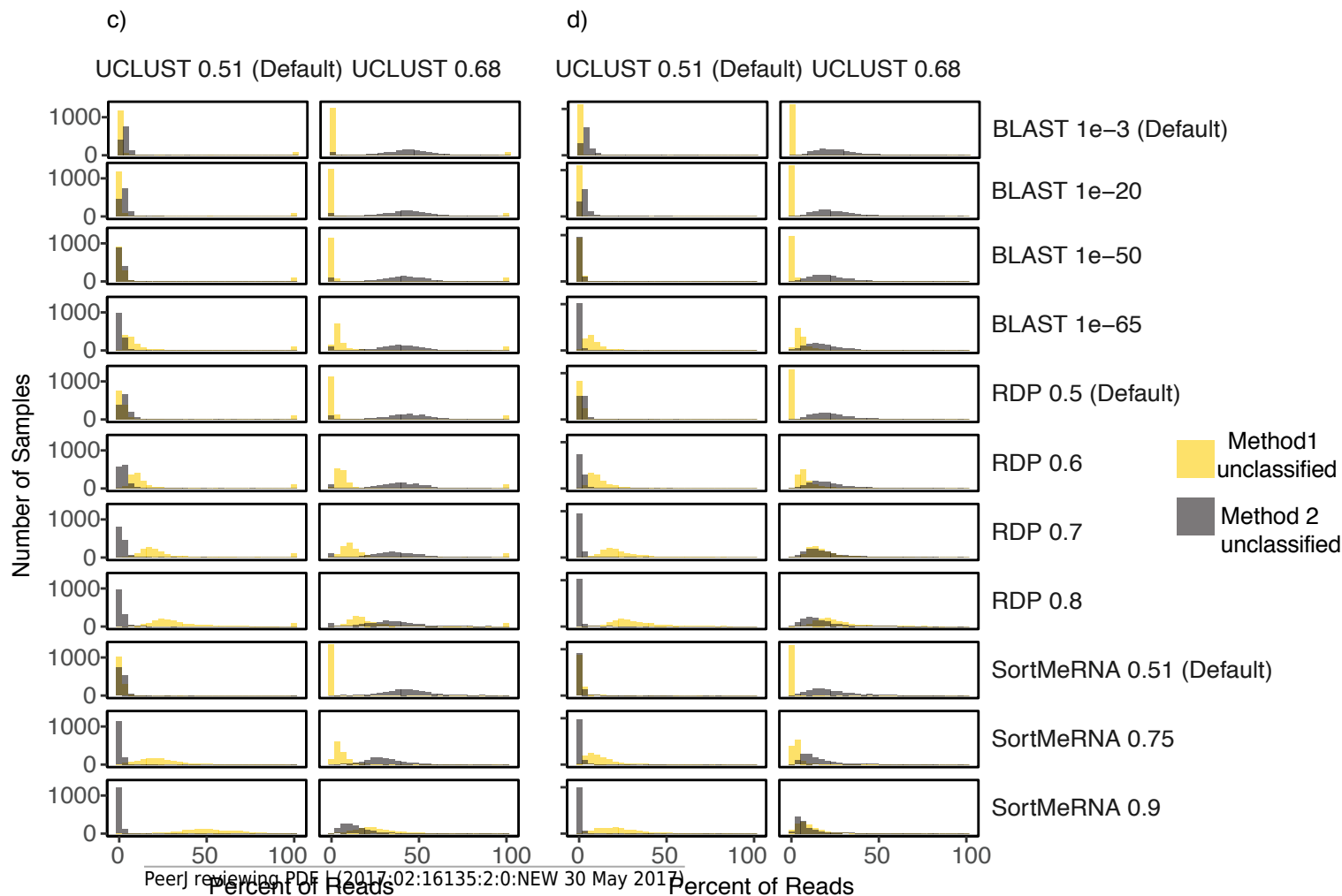
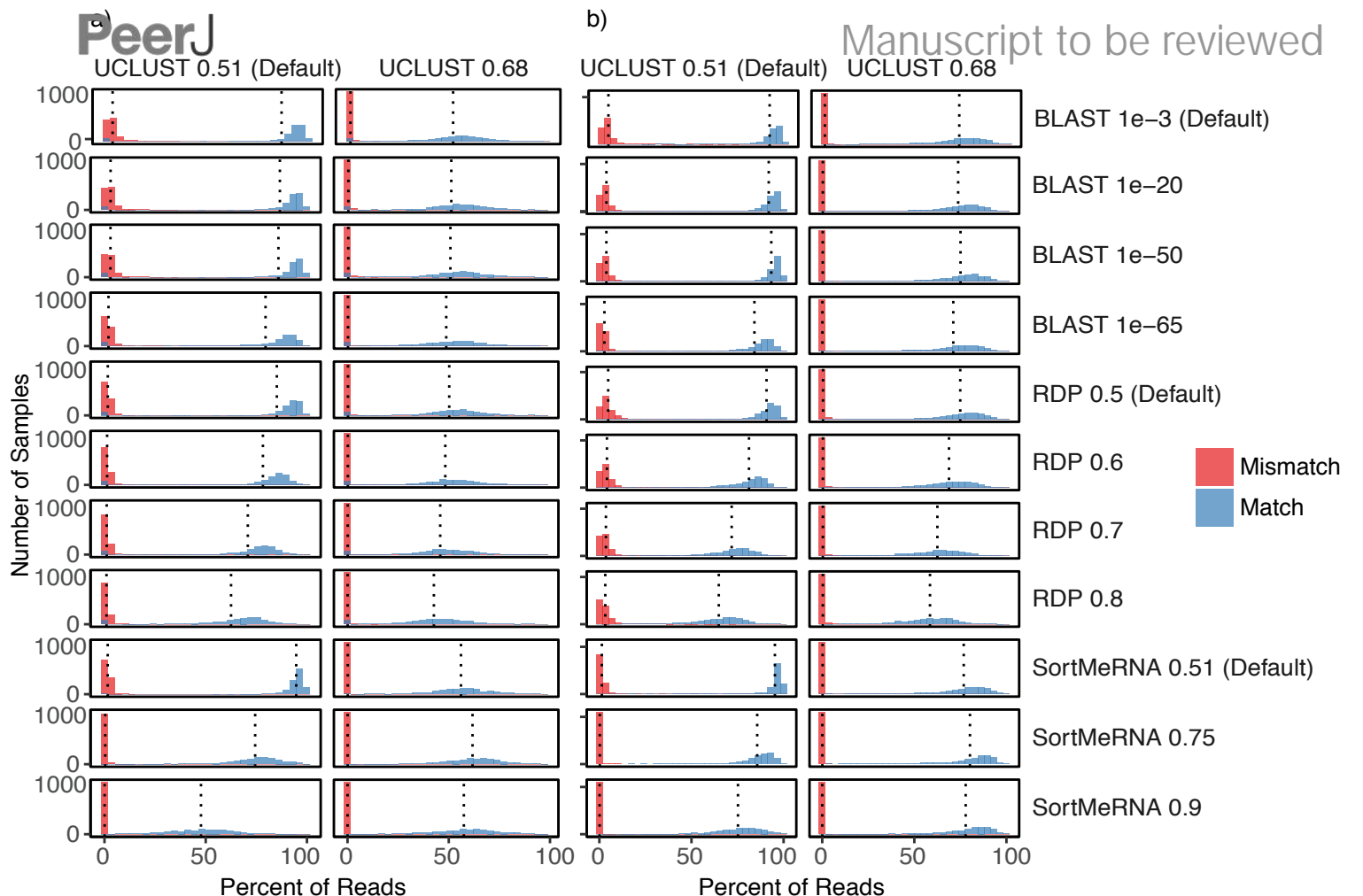
x

The agreement between taxon assignments, grouped by taxonomic rank.

Shown are the percentage of reads within each pairwise comparison among the classification methods that either matched, mismatched, or was not classified by one of the two methods at the default settings and using the Greengenes reference database. [For the unclassified designation method 1 refers to the first method listed in the comparison; the second method is the second one listed (e.g., in the RDP/BLAST comparison method 1 is RDP and method 2 is BLAST). Results are shown for all congruence categories (a) and, in order to better see the differences among the mismatch and unclassified categories, with match removed (b).

A heatmap of Jaccard Indices showing the low degree of congruence between the methods in which reads were classified to a) *Salmonella* and b) Listeria.

Values in the upper triangle are indices for comparison at the genus level; values in the lower triangle are for comparisons at the species level. A Jaccard Index of 1 indicates complete agreement in the reads classified to a given taxon for two methods; a Jaccard index of 0 indicates no overlap in the reads classified to a given taxon for two methods.

Salmonella

Listeria

Comparison of UCLUST classification results to other classifier results across the different parameter values and using both the Greengenes (a and c) and HITdb (b and d) databases.

Histograms are shown for the comparison metrics between samples at the match and mismatch categories (a and b) and for the unclassified by one method but not the other categories (c and d). In c and d Method 1 unclassified are row labels and Method 2 unclassified are column labels. For illustrative purposes, the UCLUST Default and SortMeRNA 0.9 comparison between unclassified reads and using the Greengenes database (d) resulted in virtually no reads assigned to a taxon with UCLUST that were not also assigned with SortMeRNA (Method2Nulls) but some reads were assigned to a taxon with UCLUST but not with SortMeRNA (which is to be expected as SortMeRNA 0.9 is very stringent). See Supplemental Material S1 for all comparisons.

Summary of the four different methods used to assign taxonomy to reads.

Default parameter settings are provided and "additional" values represent those used to determine the influence of settings on the congruence between methods..

Classifier	Description	Parameter Settings

BLAST	Assignment is based on the best hit from BLAST (Altschul et al. 1990)	blast_e_value
		  default: 0.001
		  additional: *1e−20, 1e−50, 1e−65*

RDP	Assignment via a Naïve Bayes Classifier under which reads are parsed into 8-mers (Wang et al. 2007)	confidence
		  default: 0.5
		  additional: *0.6, 0.7, 0.8*

SortMeRNA	Assignment is based on a variant of the Longest Increasing Subsequence for string matching within which alignment quality is based on E-values. (Kopylova et al. 2014)	min_consensus_fraction
		  default: 0.51
		  additional: *0.75, 0.9*
		similarity: 0.9
		sortmerna_e_value: 1.0
		sortmerna_coverage: 0.9
		sortmerna_best_N_alignments: 5

UCLUST	UCLUST clustering algorithm in conjunction with a reference database (Edgar 2010)	min_consensus_fraction
		  default: 0.51
		  additional: *0.68*
		similarity: 0.9
		uclust_max_accepts: 3

The number of reads assigned to species within the genera *Salmonella* and *Listeria* and the number of samples 'positive' for each.

Taxon	Classifier	Number of Reads Classified	Number of 'positive' Samples
*Listeria fleischmannii****	BLAST	7	7
	RDP	0	0
	SORTMERNA	0	0
	UCLUST	0	0
*Listeria grayi****	BLAST	85	39
	RDP	0	0
	SORTMERNA	4	4
	UCLUST	6	6
Listeria monocytogenes	BLAST	1	1
	RDP	0	0
	SORTMERNA	0	0
	UCLUST	0	0
*Listeria weihenstephanensis****	BLAST	3753	426
	RDP	80	54
	SORTMERNA	14	12
	UCLUST	9	8
Salmonella enterica	BLAST	457	204
	RDP	72	57
	SORTMERNA	43	40
	UCLUST	2090	522

1   * Non-pathogenic and non-hemolytic