1   **Metagenome-based comparisons of decay rates and host-specificity of fecal microbial**

2   **communities for improved microbial source tracking**

3

4   Brittany Suttner[1], Blake G. Lindner[1], Minjae Kim[1a], Roth E. Conrad[2], Luis M. Rodriguez[1b], Luis

5   H. Orellana[1c], Eric R. Johnston[1a], Janet K. Hatt[1], Kevin J. Zhu[1], Joe Brown[1d], and Konstantinos

6   T. Konstantinidis[1*]

7

8   [1] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA

9   30332, USA

10   [2] Ocean Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332,USA

11

12   Present address:

13   [a] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301, USA

14   [b] Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, 6020

15   Innsbruck, Tyrol, Austria

16   [c] Max-Planck-Institut für Marine Mikrobiologie, Celsiusstrasse 1, D-28359 Bremen, Germany

17   [d] Department of Environmental Sciences and Engineering, Gillings School of Global Public

18   Health, University of North Carolina at Chapel Hill, North, Carolina, NC 27599, United States

19

20   * To whom correspondence should be addressed.

21   Konstantinos T. Konstantinidis,

22   School of Civil & Environmental Engineering,

23   Georgia Institute of Technology.

24   311 Ferst Drive, ES&T Building, Room 3321,

25   Atlanta, GA, 30332.

26   Telephone: 404-639-4292

27   Email: kostas@ce.gatech.edu

28

29   **ABSTRACT**

30   Fecal material in the environment is a primary source of pathogens that cause waterborne

31   diseases and affect over a billion people worldwide. Microbial source tracking (MST) assays

32   based on single genes (e.g., 16S rRNA) do not always provide the resolution needed to attribute

33   fecal contamination sources. In this work, we used dialysis bag mesocosms simulating a

34   freshwater habitat that were spiked separately with cow, pig, or human feces to monitor the

35   decay of host-specific fecal signals over time with metagenomics, traditional qPCR, and culture-

36   based methods. Sequencing of the host fecal communities used as inocula recovered 79 non-

37   redundant metagenome-assembled genomes (MAGs) whose abundance patterns showed that the

38   majority of the fecal community signal was not detectable in the mesocosm metagenomes after

39   four days. Several MAGs showed high host specificity, and thus are promising candidates for

40   biomarkers for their respective host type. Traditional qPCR methods varied in their correlation

41   with MAG decay kinetics. Notably, the human-specific *Bacteroides* assay, HF183/BFDRev,

42   consistently under-estimated fecal pollution due to not being present in all hosts and/or primer

43   mismatches. This work provides new insights on the persistence and decay kinetics of host-

44   specific gut microbes in the environment and identifies several MAGs as putative biomarkers for

45   improved MST.

46

2

47   **KEYWORDS:** bioinformatics, comparative metagenomics, microbial ecology, water quality,

48   public health, gut microbiome

49

50   **SYNOPSIS:** We track cow, pig, and human fecal pollution in lake water over time with

51   metagenomics and benchmark these novel protocols against standard culture-based and qPCR

52   tests for water quality monitoring.

53

54   **INTRODUCTION**

55   Fecal indicator bacteria (FIB) are commonly used to assess microbial water quality and

56   identify recent fecal pollution events. Because culture-based efforts to count FIB are ineffective

57   for timely water management decisions, recent efforts have focused on rapid culture-independent

58   qPCR methods targeting traditional FIB or new biomarkers (1–3). Members of the genus

59   *Bacteroides*, or bacteriophages such as CrAssphage (4–7), are particularly suitable for microbial

60   source tracking (MST) because they tend to co-evolve with the host, are among the most

61   abundant genera in stool, have a narrow host range exclusive to warm-blooded mammals, and

62   generally have poor survival rates outside their host (8). However, recent evidence also suggests

63   the potential for *Bacteroides* to persist, and even grow under some environmental conditions (9,

64   10), which brings the assumptions about their survival outside of the host into question. Further,

65   several studies report that these markers have some cross-reactivity with other (non-human)

66   hosts (11–13) and may be too abundant in sewage for monitoring highly polluted waters (12).

67   Clearly, more research is needed on the ecology of *Bacteroides* and other biomarkers (e.g.,

68   prevalence in human vs. animal hosts from different geographical regions), their persistence in

69   the environment, and how well they correlate to risk of infection with enteric pathogens.

70   Further, the qPCR-based assays have their own (known) limitations (14). Specifically, it has

71   been challenging to reliably compare the performance of different assays across different studies

72   and environmental matrices because marker recovery efficiency and effect of PCR inhibitors can

73   vary significantly among environmental samples (15, 16). Furthermore, even the most commonly

74   used and studied human-associated markers (e.g., *Bacteroides* HF183) are not prevalent in all

75   human populations worldwide (17, 18), which suggests no single qPCR marker is likely to be

76   universally suitable for detecting human fecal contamination. Finally, fecal pollution of surface

77   waters is often the result of a complex mixture of multiple inputs further complicated by

78   environmental dispersion and deposition. The decay characteristics of different DNA markers is

79   apparently of high importance for MST investigations and evaluation of the associated public

80   health risk (19). Although an absolute gene count can be obtained via qPCR, estimates of the

81   relative contribution of various fecal sources in the natural environment cannot be quantitative

82   without this decay information. More comprehensive methods, such as metagenomics (20), can

83   help to improve biomarker discovery and overcome several of the limitations described above.

84   Most research efforts utilizing metagenomics and next-generation sequencing (NGS)

85   technologies thus far have focused on 16S rRNA gene amplicon sequencing to develop new

86   biomarkers (21). However, the 16S rRNA gene is highly conserved across *Bacteria* and

87   *Archaea*; as such, cross-reactivity with non-target hosts is common for all assays targeting even

88   the most variable regions of the 16S rRNA gene (8, 22). Functional, protein-coding genes that

89   are specific to a host's unique gut physiology (e.g. host-microbe interactions) are likely more

90   suitable targets for host-specific markers, but this represents a resolution level that 16S rRNA

91   gene amplicon data cannot offer. Clearly, more research is needed to establish the best meta-

92   omics and bioinformatic techniques as tools for identifying host-specific taxa and their genes for

93   MST applications (23). Such studies will also establish whether or not metagenomic methods

94   should be combined with conventional MST methods to obtain more accurate measures of fecal

95   pollution in watersheds since qPCR provides absolute (vs. relative for typical metagenomics

96   studies) abundances and generally has a lower limit of detection than metagenome shotgun

97   sequencing (24, 25).

98   In this study, we used dialysis bag mesocosms simulating a fecal pollution event in a

99   freshwater habitat and time-series metagenomics to track the decay of metagenome assembled

100   genomes (MAGs) from human, cow, and pig fecal inputs over time. Additionally, we used

101   traditional culture and qPCR-based MST markers and included a universal 16S rRNA gene

102   qPCR assay for translating metagenome-based relative abundances to absolute abundances in

103   order to directly compare against the traditional markers. Using the time-series abundance and

104   cross-reactivity information, we identified ~12 MAGs as candidate MST biomarkers and

105   compared their functional gene content to establish host-specific genomes and genes as potential

106   targets for improved water quality monitoring assays.

107

108   **MATERIALS AND METHODS**

109   *Lake water and fecal sample collection:* Lake water samples were collected from Lake Lanier

110   (Georgia, USA) in acid washed 10L carboys and transported immediately back to the lab and

111   stored in the dark at 4 °C until mesocosm set up the following day (within 24 hours). Human

112   fecal samples were collected from human volunteers who had not taken any antibiotics within

113   the past one month before sample collection. Since human gut microbiomes are known to vary

114   geographically, only samples from within the state of Georgia (USA) were used. All human

115   subjects in the study provided informed consent and the study was approved by the Georgia

116   Institute of Technology institutional review board (IRB) and carried out in accordance with the

117   relevant guidelines and regulations. See supplementary methods for further details on sample

118   processing.

119

120   *Mesocosm set-up:* Sterile glass bottles were filled with 1.6 L of lake water and inoculated with

121   feces to a final concentration of 2.5 g/L and shaken well to thoroughly mix the feces:lake water

122   mixture before dispensing into dialysis bags. The dialysis bag's pore size (6-8 kDa molecular

123   weight cutoff) allows passage of small molecules and ions but prevents the passage of bacterial

124   cells and viral particles. The dialysis bags were filled to a total volume of 110 mL (~21 cm

125   length of 32 mm diameter dialysis tubing) and closed on both ends using polypropylene

126   Spectra/Por clamps (Spectrum Laboratories). Enough dialysis bags were filled to sample each

127   biological replicate in triplicate at each time point, i.e., 36 dialysis bags per host type (three

128   technical replicates per three biological replicates at 4 sampling time points). Additionally, four

129   uninoculated lake water negative control and two sterile milliQ water dialysis control bags were

130   included for both of the two mesocosm experiment batches. The dialysis bags were suspended in

131   ten-gallon aquarium tanks filled with lake water and stored in environmentally controlled rooms

132   at 22 °C in the dark. A small water pump was included in each tank for aeration and nutrient

133   distribution. A small headspace of air was left in each bag when sealing with the clamps so that

134   they could float freely in the tanks.

135

136   *Mesocosm sampling:* On the day of mesocosm set up, initial day zero (D0) reference community

137   lake water samples were collected by filtering five separate 250 mL aliquots of uninoculated lake

138   water onto 0.45 μm poly-carbonate (PC) membranes, three of which were stored at -80 °C in

139   PowerFecal (Qiagen) 2 mL screw-cap bead tubes until ready for DNA extraction (within 1-3

140   months); two others were stored at -80 °C in sterile 2 mL screw-cap tubes filled with acid-

141   washed 0.1 mm glass beads until ready for analysis following the EPA Method 1611 (26).

142   Further, 100 mL of the lake water was filtered and cultured (in triplicate) on mEI medium

143   following the EPA Method 1600 for culture-based enumeration of *Enterococci* (27). Finally, the

144   feces:lake water mixtures were sampled following the same protocol for the un-inoculated lake

145   water except using a 25 mL filter volume and 10-fold serial dilutions in 1X phosphate-buffered

146   saline (PBS) for culture-based enumeration of *Enterococci* (27). All dilutions yielding

147   measurements within the acceptable range of quantification were averaged to estimate

148   CFUs/100mL of each biological replicate. To test for extraneous DNA and potential

149   contamination from sample handling, 50 mL of sterile PBS was also filtered onto PC membranes

150   and following the same DNA extraction at every sampling time point as described in the EPA

151   method 1611 above to serve as a water sample filtration blank.

152   The qPCR markers used in this study are described in Table 1 and included the human-

153   specific *Bacteroides* HF183/BFDRev (hereafter HF183; (2)), a ruminant-specific *Bacteroidetes*

154   BacR (hereafter RumBac; (28)), human mitochondrial DNA (hereafter HUMmt; (29)),

155   *Enterococcus faecalis* 16S rRNA gene (hereafter EF16S; (30)), the standard EPA Method 1611

156   assay targeting total *Enterococci* (hereafter EPA1611) and a universal 16S rRNA gene qPCR

157   assay (hereafter GenBac16S; (31)) to normalize metagenome datasets for differences in

158   microbial load. See supplementary methods for details on DNA extraction from feces and filters,

159   conditions for qPCR assays and calculations for determining qPCR marker copy number and

160   detection limits.

161

162   *Metagenomic relative abundance estimation:* Supplementary methods provide details on

163   metagenome library creation and sequencing, detection of differentially abundant (DA) genes

164   between host inocula or mesocosm time series samples, MAG recovery and dereplication at 95%

165   average nucleotide level (ANI), and MAG annotation. To track relative abundance of MAGs or

166   FIB reference genomes (Table 1) in dialysis bag mesocosm metagenomes, Magic-BLAST v1.4.0

167   ((32); options: -no_unaligned -splice F -outfmt tabular -parse_deflines T) was used to map

168   metagenomic short reads to MAG contigs in order to express MAG abundance as average

169   sequencing depth (base pairs recruited/genome length). Matches were filtered for single best

170   alignments, using a minimum 90% query cover alignment length and 95% nucleotide identity of

171   reads mapping against the reference genome (ANIr). In order to remove biases from highly

172   conserved regions and contig edges, the 80% central truncated average of sequencing depth of all

173   bases (TAD80) as described previously (33). MAG abundance in each metagenomic dataset (as

174   % of total community) was calculated as the quotient of the MAG's TAD80 value and the

175   genome equivalents (GE) from MicrobeCensus (34).

176

177   *Approach for estimating metagenome limit of detection (LOD), absolute abundances, and decay*

178   *rates:* Reads belonging to the 16S rRNA gene were extracted with SortMeRNA v2.1 ((35);

179   options: --log --fastx --blast 1 --num_alignments 1 -v -m 8336) and the SILVA 16S database

180   dereplicated at 90% identity included with the program. The average 16S rRNA gene sequencing

181   depth was estimated by summing the alignment length column from the SortMeRNA blast-like

182   tabular output file and dividing by the average 16S rRNA gene length (1540bp). The average

183   sequencing depth was then divided by the average genome sequencing depth from

184   MicrobeCensus (34) to estimate the average 16S rRNA gene copy number per genome in the

185   metagenome. The 16S rRNA gene copy number value was used to convert the GenBac16S

186   qPCR count estimates to total cell density (number of prokaryotic cells per mL or mg) by

187   dividing the qPCR count estimates by 16S rRNA gene copy number. With this information, it

188   was then possible to estimate the theoretical LODs for a *Bacteroides* genome in each

189   metagenomic dataset (in cells/mL) assuming at least 10% of a genome must be covered to

190   reliably detect it in a metagenome (36) and that the average *Bacteroides* genome is 6.5 Mbp (34).

191   The absolute abundance (cells/mL) of fecal MAGs and reference genomes (Table 1) in

192   the mesocosm metagenomes was estimated by multiplying its relative abundance (i.e. TAD80

193   >95% ANIr divided by GE) by the corresponding total cell density in each mesocosm

194   metagenome. The same protocol was followed for the human mitochondrial reference genome

195   for the HUMmt qPCR assay (Table 1), except sequencing depth was normalized using only the

196   metagenome dataset size (in Gbp). Since there is no known reference genome for the RumBac

197   assay, a 317bp contig from a cow fecal metagenome with a perfect match to the assay oligos

198   (cow5_scaffold246842) was used as a proxy to estimate the absolute abundance as described

199   above for genomes except no truncation was used when estimating sequencing depth (i.e.

200   TAD100) and a 99% identity threshold for read mapping was used instead of 95% due to the

201   high sequence conservation of the 16S rRNA gene relative to the rest of the genome (37, 38).

202   The resulting sequencing depth value was divided by the 16S rRNA gene copy number for

203   *Bacteroides* (n = 7) and genome equivalents (or GE) from MicrobeCensus (34) and subsequently

204   multiplied by total cell density to estimate total number of *Bacteroides* cells per mL. The

205   absolute abundances were used to calculate decay rates based on a first-order decay model, N_t/N_0

206    $= 10^{-kt}$ (39). The time needed to produce a 2-log reduction in abundance (t_{99}) was calculated

207   using the decay constant (k) in the following equation, $t_{99} = -2/k$.

208

209   *Data Availability*: Host fecal MAG assemblies and short reads for host fecal and mesocosm

210   metagenomes have been deposited to NCBI databases under BioProject ID PRJNA691978,

211   except the cow fecal metagenome short reads, which were deposited previously to the SRA

212   database (BioProject ID PRJNA545149).

213

214   **RESULTS AND DISCUSSION**

215   **Performance and decay of traditional culture-based and qPCR markers:**

216   Dialysis bag mesocosms simulating a natural freshwater environment were spiked with

217   cow, pig, or human feces to represent a pollution event and monitored over time. Three

218   biological replicate fecal samples were used per host and are referred to hereafter as hum1,

219   hum2, hum3, cow7, cow8, cow9, pig7, pig8, and pig9 to indicate the specific individual host

220   fecal sample that was used for DNA extraction and inoculation into the lake water mesocosms.

221   H1, H2, H3, C7, C8, C9, P7, P8, and P9 hereafter refer to the feces:lake water mesocosm sample

222   for each individual host (e.g. H1 refers to the lake water mesocosm spiked with feces from

223   hum1). Mesocosm sampling occurred in triplicate on days 0, 1, 4, 7, and 14 (hereafter, D0, D1,

224   D4, D7, and D14), which included qPCR analysis using the markers described in Table 1 and

225   metagenome sequencing.

226   The qPCR markers were first tested against the host fecal DNA samples used as inocula

227   to assess their sensitivity and specificity. The fecal DNAs were diluted 10-fold with water prior

228   to running qPCR to reduce the effect of PCR inhibitors (see Supplementary Materials and

229   Methods). All of the MST markers were not detected (ND) in any non-target hosts and none

230   were quantifiable in the uninoculated lake water negative controls (Table S6). However, the

10

231   human-specific HF183 marker was not detected in the hum2 fecal metagenome. The EPA

232   Method 1600 culture-based test for *Enterococci* (EPA1600) showed that the dialysis bag

233   mesocosms exceeded the EPA's recreational water quality criteria (RWQC) of 36 CFU/100 mL

234   throughout the entire duration of the cow and pig experiment and in all of the human timepoints

235   except on D14 (Figure 1A). Furthermore, the EPA Method 1611 qPCR-based test for

236   *Enterococci* (EPA1611) showed that all time-series samples that could be quantified exceeded

237   the EPA RWQC of 10^3 calibrator cell equivalents (CCE) per 100 mL (Figure 1B). However, this

238   assay was not detectable in the cow and pig samples by D14 and was only quantifiable in two of

239   the three human samples on D1. Overall, the concentration of *Enterococcus* spp. was similar

240   based on culture-based (EPA1600) and qPCR (EPA1611) assays and the first order decay rate

241   constant was -0.20 d^{-1} for both methods (Table S5). Additionally, method blanks (sterile PBS

242   filter controls) were included at each sampling point and analyzed according to the EPA1611

243   method and yielded no detectable amplification (data not shown), indicating no significant

244   contamination during mesocosm sample handling.

245   When tested in the time-series dialysis bag mesocosm samples, the qPCR gene copy

246   estimates for all of the host-specific MST assays decreased with time and returned to very near

247   or below the lowest concentration in the standard curves by D14 (~2.1 gene copies/µL DNA;

248   Figure 1C). Consistent with the hum2 fecal DNA results, the HF183 marker was ND in any of

249   the H2 mesocosm samples. The abundance of the HF183 marker in H3 mesocosms was two to

250   four orders of magnitude larger than the abundances observed in H1 mesocosms (Figure 1C).

251   Accordingly, only the H3 samples on D0 and D1 exceeded the quantitative microbial risk

252   assessment (QMRA)-based water quality threshold of 41 copies/mL for HF183 as simulated for

253   raw sewage in (40). Furthermore, the decay rates for the HF183 assay were 10-fold different in

254 H1 and H3 (0.02 and -0.29 d^{-1}, respectively; Table S5). The concentration of HF183 in H1

255 mesocosms was near the assay LOD (~5 cells/mL) at all time points, and thus there was no

256 discernable decay for this marker resulting in the near-zero decay rate. The average gene

257 copies/mL were consistent across the three biological replicates and were detectable until D14

258 for both the HUMmt and RumBac assays in human and cow mesocosms, respectively (Figure

259 1C).

260   The total cell density in the mesocosms based on the universal 16S qPCR assay

261 (GenBac16S; Table 1) was ~10^8 cells/mL at the start of all mesocosm incubations and tended to

262 decrease with time, reaching ~1.5×10^7 cells/mL by D7. The opposite trend was observed in the

263 negative control bags, which started at ~10^6 cells/mL and increased by nearly an order of

264 magnitude by D7 (Figure 1D). These results indicated potential bottle effects during our

265 incubations, which were assessed more fully by population genome binning of the D7

266 metagenomes as described in the Supplementary Results and Discussion. Notably, the bottle

267 effect was consistent across all mesocosms and did thus, not affect the main results reported

268 below.

269

270 **Taxonomic and phenotypic description of host-specific fecal MAGs:**

271   Host fecal reads were assembled into contigs with total length and N50 values ranging

272 from 2.5×10^7 to 1.3×10^8 and 1,913 to 19,034 base pairs, respectively (Table S3; See also

273 Supplemental Results for additional details on the metagenome datasets and community

274 coverage). Contig binning from the inocula fecal datasets (not the mesocosm datasets) resulted in

275 an initial set of 30 cow, 13 human, and 82 pig high quality MAGs. The MAGs were first

276 dereplicated at 95% average nucleotide identity (ANI) within each host resulting in a new set of

12

277 18 cow, 13 human, and 50 pig MAGs. These MAGs were subsequently further de-replicated

278 against the high quality MAGs from all three hosts and a collection of 477 Lake Lanier (LL)

279 MAGs (33) to identify any MAGs that are non-host specific and/or found in the natural

280 environment. This resulted in a final set of 17 cow, 13 human, and 49 pig high quality MAGs

281 whose IDs are provided in Supplementary Data S1. MAGs were named according to the

282 individual fecal sample from which they were originally assembled followed by the closest

283 relative of the MAG and the lowest taxonomic rank the two share according to the MiGA

284 TypeMat/NCBI database (p<0.1 threshold), i.e., C:class, O:order, F:family. G:Genus, S:Species.

285 For instance, "cow4_20_Treponema_F" means MAG #20 assembled from cow4 fecal

286 metagenome that had a *Treponema* sp. as the closest relative and was classified (at the lowest

287 level with statistical confidence) to the family *Spirochaetaceae* (or, in other words, the MAG

288 represents a novel genus and species of this family). Overall, the MAGs were highly host

289 specific at the species level (ANI >95%) and there were only two instances

290 (cow4_20_Prevotella_F and pig7_9_Tolumonas_C) in which a cow and pig MAG had ANI

291 >95% with each other and were dereplicated into a single genomospecies (i.e., a cluster of

292 genomes that is roughly equivalent to most named bacterial or archaeal species) and thus, were

293 not used further as potential biomarkers. There was more overlap among hosts when evaluating

294 the average amino acid identity values (%AAI; Figure S4) of their corresponding MAGs,

295 revealing that these MAGs likely represent distinct but closely related species found in different

296 hosts.

297   In all three host types, the majority of MAGs were classified at the class level as

298 *Bacteroidia* (41%, 46%, and 33% for cows, humans, and pigs, respectively) followed by

299 *Clostridia* (24%, 23%, and 31% for cow, humans, and pigs, respectively). In humans, the

13

300   *Bacteroidia* MAGs were primarily assigned to the family *Bacteroidaceae,* whereas the cow

301   MAGs were primarily from the *Prevotellaceae*. The majority of the pig MAGs could not be

302   classified well below class level; i.e., they represented novel families (Supplementary Data S1).

303   Consistent with their class level taxonomy, none of the host fecal MAGs (except

304   pig6_25_Oscillibacter_O) were phenotyped as aerobes using Traitar (41) and the majority of

305   MAGs were predicted to be anaerobes (100% human , 96% pigs, 82% cow; Figure S5). The

306   oxidative stress enzyme catalase was not found in any of the cow or pig MAGs but was detected

307   in two of the human MAGs (hum1_013_Akkermansia_G and hum2_003_Rubritalea_C).

308   Glucose fermentation was the most common energy-yielding pathway in MAGs from all three

309   host types (59% of cow, 71% of pig, and 100% of human MAGs). In addition to glucose

310   fermentation, 44, 15, and 15 unique sugar substrates for growth were identified in the pig, cow,

311   and human MAGs, respectively, with lactose being the most common in the pig and human

312   MAGs (76% and 85% of total MAGs, respectively) and maltose being most common in the cow

313   MAGs (82%). These results were also consistent with the DESeq2 analysis at the individual gene

314   level (see Supplementary Results and Discussion).

315

316   **Decay kinetics of host fecal MAGs in the mesocosms:**

317   MAG abundance dynamics over the incubation time revealed that all 13 dereplicated

318   human MAGs were detected in at least one human mesocosm, while only 13 out of 17 total cow

319   and 41 out of 49 total pig MAGs were detected in all cow and pig mesocosms, respectively. For

320   the conditions tested here, the majority of fecal MAGs from all three hosts were not detectable in

321   the mesocosm metagenomes after D4 (Figures S7 & S8). Accordingly, it was only possible to

322   estimate decay rates for 8 human, 3 cow, and 17 pig MAGs, respectively (Table S5) because at

323   least three abundance data points (i.e., D0, D1, and D4) were required. For the MAGs with

324   sufficient data points, the average log-2 reduction time (t_{99}) was similar for cow and pig MAGs

325   (4.5 and 5.6 days, respectively) but was higher for the human MAGs (average t_{99} of 14.3 days;

326   Table S5). This result was largely consistent with a previous quantitative microbial risk

327   assessment (QMRA) analysis that predicted that the gastrointestinal infection risk from sewage

328   contamination in surface waters is not significant (<3% chance of infection) after 3.3 days (40) in

329   accordance with the EPA risk threshold for bathing water (42). Specifically, Boehm et al,

330   reported t_{99} of 1.4 d for protozoan, 2.5d for viral, and 4.7 d for *E. coli* O157:H7 and 11.8 d for

331   *Salmonella*. Thus, the emerging 3-4 day rule of thumb for acceptable risk-levels seems to apply

332   to many (but not necessarily all) fecal pathogens in aquatic environments. Importantly, the decay

333   rate of the fecal MAGs is greater than or similar to those reported previously for several

334   pathogens, suggesting that the MAGs may be suitable candidates for use as FIB with this respect.

335   Consistent with this conclusion, none of the fecal MAGs were detected in any of the

336   uninoculated lake water negative control metagenomes or matched closely any of the 477 LL

337   MAGs, i.e., they are absent in the nearby natural ecosystem (Figure S11). Caution is needed,

338   however, to extrapolate these results to all habitats, as some aquatic habitats or environmental

339   conditions are known to support long-term survival of both pathogens and FIB (43).

340   Furthermore, there are only a few studies to date reporting decay rates of viral markers and

341   pathogens in the environment (44, 45) for evaluation against the MAG decay kinetics reported

342   here; hence, viral pathogens may deviate from these decay patterns.

343   The human MAGs showed much higher individual host specificity than the cow and pig

344   MAGs, i.e., MAGs assembled from an individual human fecal metagenome were always the

345   most abundant in the mesocosms spiked with the feces from that individual and showed much

346 lower abundances in the other two biological replicates (Figure S7A). In particular, among the

347 hum2 MAGs, none were present in the H3 mesocosms and only two were detected in the H1

348 mesocosms; thus, none of the hum2 MAGs were selected as candidate biomarkers (see below).

349 Therefore, targeting a single biomarker (whether it be a whole genome or qPCR assay) for MST

350 can still be limiting due to the high individual variability observed in the human or animal gut,

351 consistent with previous literature (46, 47), and the whole-community metagenomic approach

352 employed here and/or targeting multiple biomarkers may be advantageous with this respect.

353 Obviously, this limitation is not as important for MST in cases where the fecal input represents

354 the composite excreta of many individuals such as in municipal sewage systems.

355

356 **Best-performing host fecal MAGs:**

357   Based on the decay and host specificity results, we identified five cow, three human, and

358 six pig MAGs that were present in all three biological replicates of the same host type, were

359 highly abundant on D0 (>0.1%) and were not detected in the metagenomes after D4 (Figures S7

360 and S8; Supplementary Data S1). We investigated these MAGs further as candidate biomarkers

361 for MST. Notably, although most of the cow and pig MAGs are *Bacteroidia* and *Clostridia*, two

362 of the cow biomarkers (cow4_001_Treponema_F and cow8_3_Treponema_F) were actually

363 classified in the family *Spirochaetaceae*, while an *Actinobacteria* (pig4_16_Cellulomonas_C)

364 and the archaeal phylum *Euryarchaeota* (pig4_38_Methanoplasma_F) were among the pig

365 biomarker MAGs (Supplementary Data S1). These results suggest that biomarkers may be found

366 in novel taxa not previously considered for MST.

367   Phenotype classification using Traitar (41) showed that none of the potential biomarker

368 MAGs were aerobes or facultative anaerobes like the most commonly used FIB such as *E.*

16

369   *faecalis* and *E. coli*, and all had primarily anaerobic phenotypes related to carbohydrate

370   fermentation (Figure S5). Accordingly, the best gene targets for MST assay development at the

371   individual gene level to detect relatively recent pollution events will likely be related to

372   anaerobic functions specific to the different host types rather than the 16S rRNA gene, which has

373   primarily been the target of most MST research to date. There were several functional genes that

374   were significantly enriched in one host compared to the others and thus, could be targets for

375   biomarker development (see also additional discussion in the Supplementary Material). These

376   patterns, and the accompanying high host-specificity of the MAGs recovered, are presumably

377   driven, at least to some extent, by the different selection pressures prevailing in the gut of each

378   animal, as also indicated by the type of fermenters present in the different hosts.

379

380   *Functional annotation for host-specific MAGs and gene functions:*

381   The 14 MAGs identified as potential markers based on their host sensitivity, abundance,

382   and decay kinetics in the mesocosms showed no clear clustering based on the KEGG modules

383   (48) found in their genome and no modules were clearly unique to a single host type (Figure S6).

384   Thus, DESeq2 differential abundance (DA) analysis (49) based on reads mapped to assembled

385   genes was used to identify specific functions that are enriched in the host fecal metagenomic

386   assemblies. Of the 2,080 total KEGG functions identified, 177 were significantly DA with $P_{adj} <$

387   0.05 and $\log_2$ fold change (L2FC) > 3 using pairwise comparisons between human, cow, and pig

388   fecal samples (Supplementary Data S2; Figure S13). Most of these gene functions were also

389   recovered in the corresponding MAGs for the host type, further corroborating that the candidate

390   host-specific MAGs are robust biomarkers (see Supplementary Results and Discussion for

391   further details).

392   Most notably, seven genes for a type IV secretion system (T4SS) were highly abundant

393   and specific to the cow gut metagenomes (Figure S13). Evidence has shown that T4SS proteins

394   are important for shaping community composition in the gut (50), which suggests these proteins

395   could be viable targets as host-specific markers. These results were also consistent with another

396   study using competitive DNA hybridization to survey metagenomes and found host-specific

397   sequences related to secretion and surface-associated proteins (51). Furthermore, some of the DA

398   KEGG functions offered new insight on the fermentation pathways that distinguish cows and

399   pigs. Fumarate reductase subunit D (*frdD*), which is associated with the primitive electron

400   transport chain (ETC) of some fermenters (52), was more abundant in cows. The pig samples

401   were instead enriched for two genes associated with butyrate-producing fermentation (*atoA*;

402   acetoacetate CoA-transferase beta subunit, *bcd*; butyryl-CoA dehydrogenase) as well as the gene

403   associated with H_2-producing fermentation (*porD*; pyruvate ferredoxin oxidoreductase delta

404   subunit) (Supplementary Data S2). These results indicated that fermenting microbes inhabiting

405   the cow and pig gut carry out different strategies to sink excess reducing equivalents (the

406   primitive ETC or H_2, respectively). However, these trends were not discernable for the human

407   samples as fewer genes overall tended to be significantly enriched in human inocula, which

408   could be the result of sampling limitation (only 3 human fecal samples were compared against 6

409   cow and 6 pig samples) and the higher inter-person diversity described above.

410

411   **Decay of potential biomarker MAGs vs. reference FIB genomes in the mesocosms**

412   The absolute abundances (cells or viral particles/mL) of common FIB and MST

413   biomarkers over time were also compared to the candidate host-specific biomarker MAGs. The

414   former biomarkers included reference genomes associated with the qPCR assays used in this

415   study (Table 1) as well as genomes of the common commensal *E. coli* HS (NC_009800.1) and

416   CrAssphage (JQ995537). *E. faecalis* and *E. coli*, despite being "gold standard" FIB, performed

417   worse than the MST markers described here. *E. faecalis* was not detected in any human feces or

418   mesocosm samples by qPCR or metagenome-based methods and its abundance was too low in

419   the cow and pig feces to be detected upon dilution in the lake water mesocosms (data not

420   shown). Hence, this organism would not be able to indicate fecal contamination for any of the

421   hosts in this study. *E. coli* was detected in all of the host mesocosms and persisted for ~1 week

422   (Figure S16A), longer than the presumed fecal contamination risk of 4 days described above, and

423   maintained higher abundances over time compared to the fecal MAGs (Figure 2). Although it is

424   well known that *E. coli* and *E. faecalis* are not host-specific, and thus their usefulness for MST

425   may be limited, these results confirmed our expectations and provided further evidence against

426   the use of these organisms as FIB and highlight the need for improved standard indicators.

427   The *B. dorei* and CrAssphage genomes were not detected in any cow or pig mesocosm

428   metagenomes and they also had similar decay profiles to the human fecal MAGs (Figure 2A &

429   B); except the CrAssphage genome abundance increased from D0 to D1, whereas the *B. dorei*

430   genome (and fecal MAGs) abundance consistently decreased with time, which could possibly

431   indicate a predatory relationship between these two microbes (CrAssphage is predicted to be a

432   *Bacteroides* bacteriophage). Further, consistent with the qPCR results, neither of these genomes

433   were detected in any of the H2 mesocosm metagenomes. *Bacteroides* abundance based on the

434   HF183 qPCR assay tended to be lower than the MAGs and *B. dorei* reference genome (see

435   below). The human mitochondrial genome (mtGenome) was detected in all three human

436   mesocosm metagenomes until D7 and showed a steady decay in abundance with time (Figure 2A

437   & B and S16B), consistent with the HUMmt assay, which was detectable by qPCR until D14

438   (Figure 1C). The cow fecal MAGs were all ND by D7 and decayed faster compared to the *E. coli*

439   reference genome and *Bacteroides* abundance based on the RumBac qPCR assay (Figures 2C, S5

440   and S6).

441

442   *Correlation of MST qPCR markers to their metagenome counterpart:*

443   In order to more precisely evaluate the performance of the metagenome-based results

444   against those of traditional qPCR assays, absolute abundances (expressed as cells/mL) of the

445   RumBac and HF183 *Bacteroidetes* assays were compared to the abundance of the corresponding

446   reference genome in the mesocosm metagenomes. The correlation between *Bacteroides*

447   abundances based on qPCR and metagenomes estimates was not consistent between the two

448   assays (Figure 3A & B; R^2=0.18 and 0.76 for HF183 and RumBac, respectively). The RumBac

449   qPCR assay tended to give higher abundance estimates (linear regression slope = 0.16) than its

450   metagenome counterpart (Figure 3B). The HF183 qPCR assay consistently gave lower estimates

451   of *Bacteroides* abundance in the human mesocosms (linear regression slope= 10.26), especially

452   in H1, in which the HF183 qPCR assay estimated only about 6 *Bacteroides* cells/mL in the

453   mesocosms on D0, D1, and D4, well below the theoretical LOD for *B. dorei* in the metagenomes

454   ($\sim$3x10^4 cells/mL; see Materials and Methods for LOD estimation). However, the *B. dorei*

455   reference genome was well above this concentration based on metagenome abundance (Figure

456   3A). Further investigation showed that this was presumably caused by mismatches of the

457   forward HF183 primer to the dominant *Bacteroides* strains present in the host fecal inocula

458   (Figure S15). Specifically, the short reads from the fecal inoculum were searched against the 16S

459   rRNA gene of the reference *B. dorei* strain (which contains a perfect match to the HF183 assay

460   primers and probe) to calculate its 99% identity truncated average sequencing depth (TAD80).

461   For both hum1 and hum3 fecal metagenomes (there was no detection in hum2), the sequencing

462   depths of the probe and reverse primer were similar to the overall average sequencing depth for

463   the entire 16S rRNA gene (at about 42.0 and 247.0 for hum1 and hum3, respectively). However,

464   the sequencing depth of the forward primer region was 0 in hum1 and ~40 (6x less than the

465   average) in hum3. Furthermore, we manually checked the metagenomic reads for perfect

466   matches to the HF183 forward primer and found none in hum1 and only 17 in hum3, suggesting

467   that this region is not present in the dominant *Bacteroides* strain that was assembled from each

468   host.

469   Thus, our evaluation of traditional qPCR assays revealed several of the known limitations

470   of this approach such as mismatches of the PCR primers against the taxa present in the sample

471   (Figure S15), and lower decay rates of short DNA fragments (~120bp) targeted by PCR relative

472   to whole cells or the whole chromosome (see also Supplementary Results and Discussion). It is

473   important to note, however, that the PCR primer limitation is not expected to be as pronounced

474   in cases where the fecal input represents many individuals due to the high inter-individual

475   variability in the microbiome. In such cases, the PCR markers such as the HF183 are expected to

476   perform well for their purposes, as previously noted (40). Furthermore, there was some overlap

477   among the different host fecal MAGs at the genus level (>65% AAI; Figure S4), which, most

478   likely, accounts for the cross-reactivity commonly observed for the various 16S qPCR assays

479   targeting *Bacteroidales* at above the species level. (8, 16, 22).

480

481   **Using metagenomic methods for microbial source tracking**

482   Collectively, our findings suggest that the use of metagenomic methods to identify host-

483   specific MAGs and detect and track these MAGs in an environmental-like system is highly

21

484   promising and circumvents several of the limitations of traditional methods. Considering the

485   high individual host variability, especially among human hosts, more work is needed to

486   characterize the geographic stability of the putative biomarkers of human or animal hosts

487   reported here and the degree of their biogeography. All of the cow and pig fecal samples used

488   here were from the same farm in northern Georgia (USA) for the convenience in obtaining these

489   inocula as well as technical limitations in running the mesocosm incubations with a larger

490   number of samples. Thus, it will be important to determine if these MAGs are present in animals

491   from other herds across broader geographical regions. Many recent studies have made

492   considerable effort to sequence metagenomes and/or assemble MAGs from cow rumen (53–55)

493   as well a pig (56, 57) and chicken guts (58). However, this information has not yet been

494   synthesized together for MST marker development. Future work should leverage these datasets

495   to improve comparative functional gene analysis along with decay information to search for

496   better DNA markers. Furthermore, as high-throughput sequencing becomes more affordable and

497   routine, it may be possible to directly assess MST markers (and even pathogens) in

498   environmental metagenomes. To make regulatory standards based on metagenome data,

499   calculating absolute abundances of indicators (or pathogens) will be necessary. The

500   methodologies proposed here should be helpful in these directions.

501

502   **Acknowledgments**

507   (to B.S). The funding agencies had no role in the study design, data collection and analysis,

508   decision to publish, or preparation of the manuscript.

509

510   **Conflict of interest:** The authors declare no conflict of interest.

511

512

References

1. Kildare BJ, Leutenegger CM, McSwain BS, Bambic DG, Rajal VB, Wuertz S. 2007. 16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal *Bacteroidales*: A Bayesian approach. Water Research 41:3701–3715.

2. Haugland RA, Varma M, Sivaganesan M, Kelty C, Peed L, Shanks OC. 2010. Evaluation of genetic markers from the 16S rRNA gene V2 region for use in quantitative detection of selected *Bacteroidales* species and human fecal waste by qPCR. Syst Appl Microbiol 33:348–57.

3. McLellan SL, Eren AM. 2014. Discovering new indicators of fecal pollution. Trends in microbiology 22:697–706.

4. Stachler E, Bibby K. 2014. Metagenomic Evaluation of the Highly Abundant Human Gut Bacteriophage CrAssphage for Source Tracking of Human Fecal Pollution. Environmental Science & Technology Letters 1:405–409.

5. García-Aljaro C, Ballesté E, Muniesa M, Jofre J. 2017. Determination of crAssphage in water samples and applicability for tracking human faecal pollution. Microbial Biotechnology 10:1775–1780.

6. Cinek O, Mazankova K, Kramna L, Odeh R, Alassaf A, Ibekwe MU, Ahmadov G, Mekki H, Abdullah MA, Elmahi BME, Hyöty H, Rainetova P. 2018. Quantitative CrAssphage real-time PCR assay derived from data of multiple geographically distant populations. Journal of Medical Virology 90:767–771.

7. Liang Y, Jin X, Huang Y, Chen S. 2018. Development and application of a real-time
    polymerase chain reaction assay for detection of a novel gut bacteriophage (crAssphage).
    Journal of Medical Virology 90:464–468.

8. Ahmed W, Hughes B, Harwood VJ. 2016. Current Status of Marker Genes of Bacteroides
    and Related Taxa for Identifying Sewage Pollution in Environmental Waters. Water 8:231.

9. Green HC, Shanks OC, Sivaganesan M, Haugland RA, Field KG. 2011. Differential decay
    of human faecal *Bacteroides* in marine and freshwater. Environmental microbiology
    13:3235–3249.

10. Weidhaas J, Mantha S, Hair E, Nayak B, Harwood VJ. 2015. Evidence for Extraintestinal
    Growth of *Bacteroidales* Originating from Poultry Litter. Applied and Environmental
    Microbiology 81:196–202.

11. Ahmed W, Lobos A, Senkbeil J, Peraud J, Gallard J, Harwood VJ. 2018. Evaluation of the
    novel crAssphage marker for sewage pollution tracking in storm drain outfalls in Tampa,
    Florida. Water Research 131:142–150.

12. Stachler E, Akyon B, de Carvalho NA, Ference C, Bibby K. 2018. Correlation of
    crAssphage qPCR Markers with Culturable and Molecular Indicators of Human Fecal
    Pollution in an Impacted Urban Watershed. Environ Sci Technol 52:7505–7512.

13. Ahmed W, Payyappat S, Cassidy M, Besley C, Power K. 2018. Novel crAssphage marker
    genes ascertain sewage pollution in a recreational lake receiving urban stormwater runoff.
    Water Research https://doi.org/10.1016/j.watres.2018.08.049.

14. Savichtcheva O, Okabe S. 2006. Alternative indicators of fecal pollution: relations with pathogens and conventional indicators, current methodologies for direct pathogen monitoring and future application perspectives. Water Res 40:2463–76.

15. Ahmed W, Hughes B, Harwood VJ. 2016. Current Status of Marker Genes of *Bacteroides* and Related Taxa for Identifying Sewage Pollution in Environmental Waters. 6. Water 8:231.

16. Boehm AB, Van De Werfhorst LC, Griffith JF, Holden PA, Jay JA, Shanks OC, Wang D, Weisberg SB. 2013. Performance of forty-one microbial source tracking methods: a twenty-seven lab evaluation study. Water Res 47:6812–28.

17. Reischer GH, Ebdon JE, Bauer JM, Schuster N, Ahmed W, Åström J, Blanch AR, Blöschl G, Byamukama D, Coakley T, Ferguson C, Goshu G, Ko G, de Roda Husman AM, Mushi D, Poma R, Pradhan B, Rajal V, Schade MA, Sommer R, Taylor H, Toth EM, Vrajmasu V, Wuertz S, Mach RL, Farnleitner AH. 2013. Performance Characteristics of qPCR Assays Targeting Human- and Ruminant-Associated *Bacteroidetes* for Microbial Source Tracking across Sixteen Countries on Six Continents. Environmental science & technology 47:8548–8556.

18. Mayer RE, Reischer GH, Ixenmaier SK, Derx J, Blaschke AP, Ebdon JE, Linke R, Egle L, Ahmed W, Blanch AR, Byamukama D, Savill M, Mushi D, Cristóbal HA, Edge TA, Schade MA, Aslan A, Brooks YM, Sommer R, Masago Y, Sato MI, Taylor HD, Rose JB, Wuertz S, Shanks OC, Piringer H, Mach RL, Savio D, Zessner M, Farnleitner AH. 2018. Global Distribution of Human-Associated Fecal Genetic Markers in Reference Samples from Six Continents. Environ Sci Technol 52:5076–5084.

19. Cloutier DD, McLellan SL. 2017. Distribution and Differential Survival of Traditional and Alternative Indicators of Fecal Pollution at Freshwater Beaches. Applied and Environmental Microbiology 83.

20. Handelsman J, Tiedje JM, Alvarez-Cohen L, Ashburner M, Cann I, Delong E. 2007. The new science of metagenomics: revealing the secrets of our microbial planet. National Academies Press.

21. Unno T, Staley C, Brown CM, Han D, Sadowsky MJ, Hur H-G. 2018. Fecal pollution: new trends and challenges in microbial source tracking using next-generation sequencing. Environmental Microbiology 0.

22. Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol Rev 38:1–40.

23. Sharma M, Sharma NR. 2020. Metagenomic Applications of Wastewater Treatment, p. 157–166. *In* Chopra, RS, Chopra, C, Sharma, NR (eds.), Metagenomics: Techniques, Applications, Challenges and Opportunities. Springer, Singapore.

24. Suttner B, Johnston ER, Orellana LH, Rodriguez-R LM, Hatt JK, Carychao D, Carter MQ, Cooley MB, Konstantinidis KT. 2020. Metagenomics as a Public Health Risk Assessment Tool in a Study of Natural Creek Sediments Influenced by Agricultural and Livestock Runoff: Potential and Limitations. Appl Environ Microbiol 86:e02525-19.

25. Hong P-Y, Mantilla-Calderon D, Wang C. 2020. Mini Review: Metagenomics as a tool to monitor reclaimed water quality. Applied and Environmental Microbiology https://doi.org/10.1128/AEM.00724-20.

26. USEPA. 2012. Method 1611: Enterococci in Water by TaqMan Quantitative Polymerase Chain Reaction (qPCR) Assay.

27. USEPA. 2002. Method 1600: enterococci in water by membrane filtration using membrane-enterococcus indoxyl-B-D-glucoside agar (mEI), EPA 821-R-02-022. United States Environmental Protection Agency Washington, DC.

28. Reischer GH, Kasper DC, Steinborn R, Mach RL, Farnleitner AH. 2006. Quantitative PCR Method for Sensitive Detection of Ruminant Fecal Pollution in Freshwater and Evaluation of This Method in Alpine Karstic Regions. Applied and Environmental Microbiology 72:5610–5614.

29. Caldwell JM, Raley ME, Levine JF. 2007. Mitochondrial Multiplex Real-Time PCR as a Source Tracking Method in Fecal-Contaminated Effluents. Environmental Science & Technology 41:3277–3283.

30. Santo Domingo JW, Siefring SC, Haugland RA. 2003. Real-time PCR method to detect Enterococcus faecalis in water. Biotechnology Letters 25:261–265.

31. Ritalahti KM, Amos BK, Sung Y, Wu Q, Koenigsberg SS, Löffler FE. 2006. Quantitative PCR Targeting 16S rRNA and Reductive Dehalogenase Genes Simultaneously Monitors Multiple Dehalococcoides Strains. Applied and Environmental Microbiology 72:2765–2774.

32. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics 20:405.

33. Rodriguez-R LM, Tsementzi D, Luo C, Konstantinidis KT. 2020. Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. Environ Microbiol 22:3394–3412.

34. Nayfach S, Pollard KS. 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biology 16:51.

35. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217.

36. Castro JC, Rodriguez-R LM, Harvey WT, Weigand MR, Hatt JK, Carter MQ, Konstantinidis KT. 2018. imGLAD: accurate detection and quantification of target organisms in metagenomes. PeerJ 6.

37. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. 9. Nature Reviews Microbiology 12:635–645.

38. Konstantinidis KT, Rosselló-Móra R, Amann R. 2017. Uncultivated microbes in need of their own taxonomy. 11. The ISME Journal 11:2399–2406.

39. Crane SR, Moore JA. 1986. Modeling enteric bacterial die-off: A review. Water, Air, & Soil Pollution 27:411–439.

40. Boehm AB, Graham KE, Jennings WC. 2018. Can We Swim Yet? Systematic Review, Meta-Analysis, and Risk Assessment of Aging Sewage in Surface Waters. Environmental Science & Technology 52:9634–9645.

41. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. 2016. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems 1.

42. USEPA. 2012. Recreational Water Quality Criteria. EPA820-F-12–058.

43. Korajkic A, Wanjugi P, Brooks L, Cao Y, Harwood VJ. 2019. Persistence and Decay of Fecal Microbiota in Aquatic Habitats. Microbiology and Molecular Biology Reviews 83.

44. Boehm AB, Silverman AI, Schriewer A, Goodwin K. 2019. Systematic review and meta-analysis of decay rates of waterborne mammalian viruses and coliphages in surface waters. Water Research 164:114898.

45. Greaves J, Stone D, Wu Z, Bibby K. 2020. Persistence of emerging viral fecal indicators in large-scale freshwater mesocosms. Water Research X 9:100067.

46. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial Community Variation in Human Body Habitats Across Space and Time. Science 326:1694–1697.

47. Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLOS Biology 17:e3000102.

48. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36:2251–2252.

49. Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology 11:R106.

50. Verster AJ, Ross BD, Radey MC, Bao Y, Goodman AL, Mougous JD, Borenstein E. 2017. The Landscape of Type VI Secretion across Human Gut Microbiomes Reveals Its Role in Community Composition. Cell Host Microbe 22:411-419.e4.

51. Shanks OC, Domingo JWS, Lu J, Kelty CA, Graham JE. 2007. Identification of bacterial DNA markers for the detection of human fecal pollution in water. Appl Environ Microbiol 73:2416–2422.

52. Besten G den, Eunen K van, Groen AK, Venema K, Reijngoud D-J, Bakker BM. 2013. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. J Lipid Res 54:2325–2340.

53. Wilkinson T, Korir D, Ogugo M, Stewart RD, Watson M, Paxton E, Goopy J, Robert C. 2020. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. Genome Biology 21:229.

54. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature Biotechnology 1–10.

55. Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. 2019. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. 8. Nature Biotechnology 37:953–961.

56. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z, Feng Q, Liang S, Pedersen AØ, Kjeldsen NJ, Liu C, Maguin E, Doré J, Pons N, Le Chatelier E, Prifti E, Li J, Jia H, Liu X, Xu X, Ehrlich SD, Madsen L, Kristiansen K, Rogel-Gaillard C, Wang J. 2016. A reference gene catalogue of the pig gut microbiome. 12. Nature Microbiology 1:1–6.

57. Wang C, Li P, Yan Q, Chen L, Li T, Zhang W, Li H, Chen C, Han X, Zhang S, Xu M, Li B, Zhang X, Ni H, Ma Y, Dong B, Li S, Liu S. 2019. Characterization of the Pig Gut Microbiome and Antibiotic Resistome in Industrialized Feedlots in China. mSystems 4.

58. Gilroy R, Ravi A, Getino M, Pursley I, Horton D, Alikhan N-F, Baker D, Gharbi K, Hall N, Watson M. 2020. A Genomic Blueprint of the Chicken Gut Microbiome.

Table 1: qPCR markers used in this study and associated reference genomes. Host-specific MST markers include HF183, RumBac, and HUMmt; general FIB markers are EF16S and EPA1611. The GenBac16S assay was used for absolute quantification and LOD estimation for reference genomes in the metagenomes as described in the Materials and Methods section.

| Marker	Target	Reference	Reference genome	Accession
HF183	Human *Bacteroides* 16S	2	*Bacteroides dorei* CL03T12C01	NZ_CP011531.1
RumBac	Ruminant *Bacteroides* 16S	48	n/a	n/a
HUMmt	Human mtDNA NADH dehydrogenase subunit 5	49	Human mitochondrion genome	J01415.2
EF16S	*E. faecalis* 16S	50	*E. faecalis* ATCC29212	CP008816.1
EPA1611	*Enterococcus* 23S	46	*E. faecalis* ATCC29212	CP008816.1
GenBac16S	Universal 16S rRNA	51	n/a	n/a
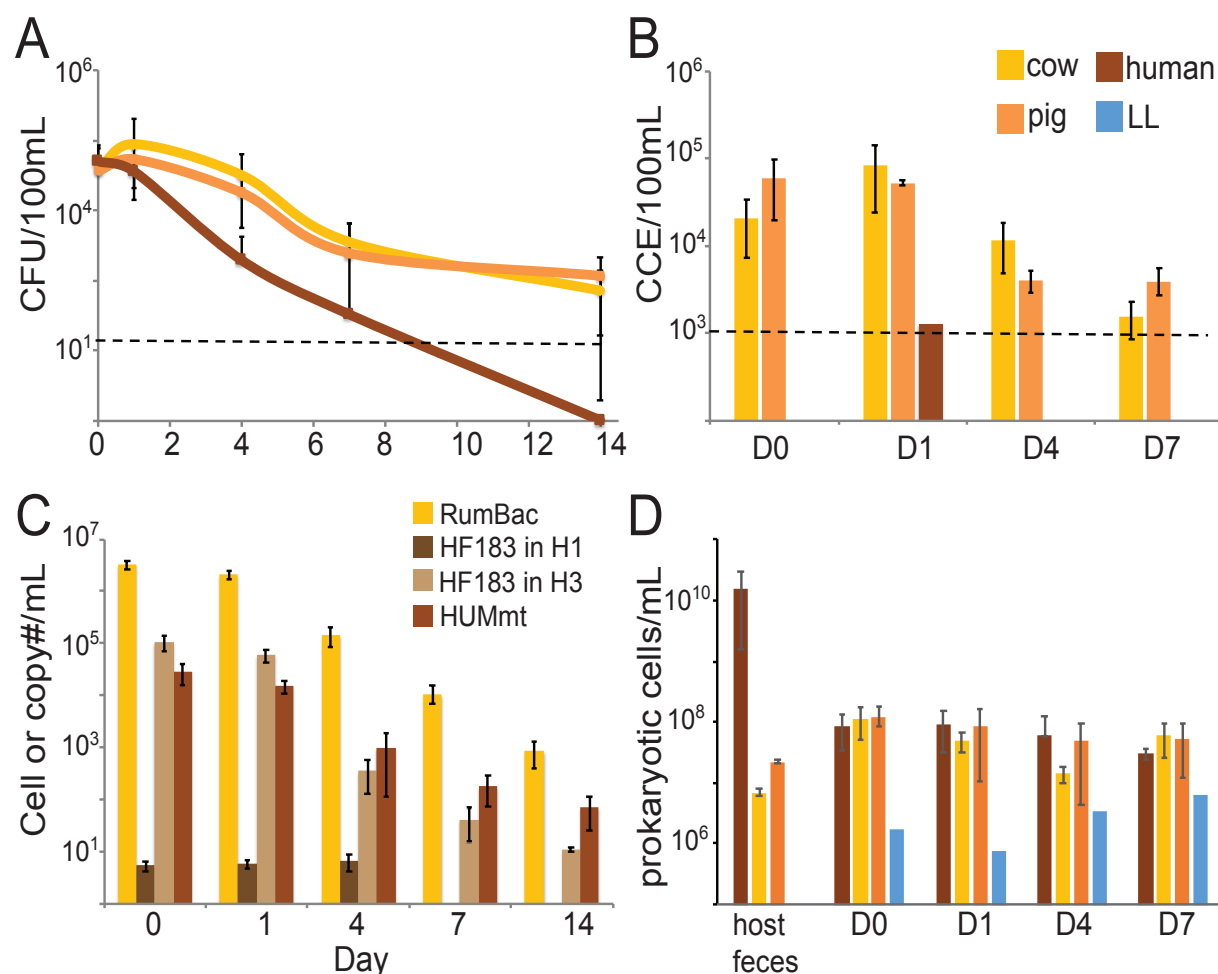
Figure 1: Traditional FIB, MST marker, and total bacterial cell abundances during the mesocosm incubations (A) EPA Method 1600 culture-based enumeration of *Enterococcus*. (B) EPA Method 1611 qPCR-based enumeration of *Enterococcus*. Black dotted lines show the EPA's recreational water quality criteria (RWQC) limit for impaired waters for each assay (CFU= colony forming units; CCE= calibrator cell equivalents). (C) Host-specific MST qPCR assays that could be detected in the dialysis bag mesocosms. The HUMmt is reported as #copies/mL and the rest are reported as #cells/mL. (D) Cell density in the mesocosms over time based on a universal 16S qPCR assay (GenBac16S). The average 16S rRNA gene copy number per genome was estimated from the corresponding metagenome for each sample by dividing average 16S gene sequencing depth by the average genome sequencing depth as described in the main text. In all figures, error bars are the standard deviation for averages that had more than three data points.
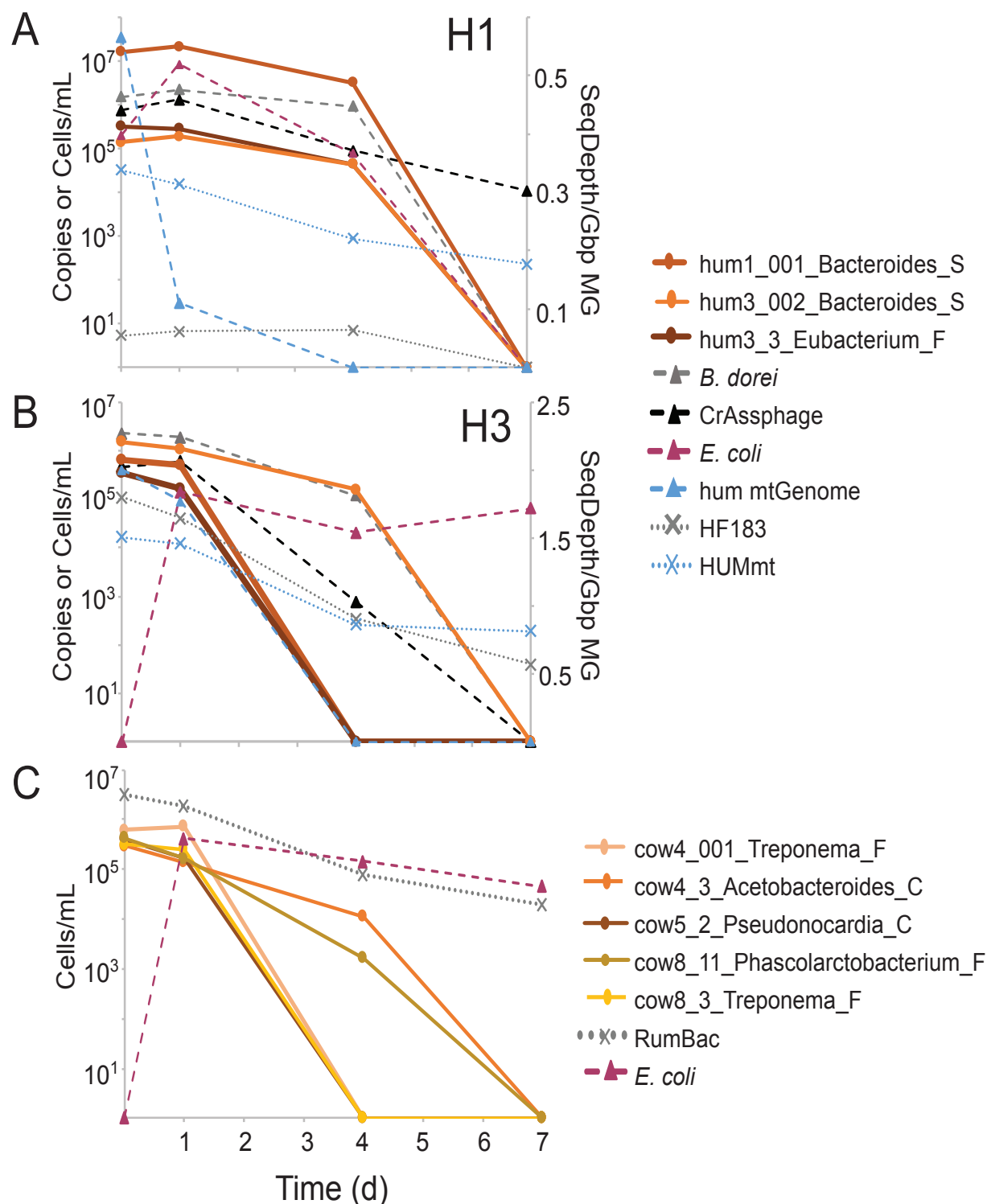
Figure 2: Compare absolute abundances of putative biomarker MAGs, traditional FIB and MST qPCR markers in (A) H1 mesocosms, (B) H3 mesocosms, and (C) the average of all 3 biological replicates of the cow fecal mesocosms. Absolute abundances (gene copies, cells or viral particles per mL) were determined for all targets except for the human mitochondrial

genome (hum mtGenome), which is expressed as relative abundance (sequencing depth per Gbp metagenome) and is shown on the secondary axis for (A) and (B). MAGs are represented by solid lines with circle markers. Reference genomes are represented by dashed lines with triangle markers and include *Bacteroides dorei*, CrAssphage, *E. coli*, and the human mtGenome. The qPCR assays are represented by dotted lines with X markers and included the human-specific and ruminant specific *Bacteroides* assays (HF183 and RumBac, respectively) and the human mtDNA assay (HUMmt; reported as copies/mL). The human mesocosms are plotted separately because they were more variable among each other compared to the cows and also because neither *B. dorei*, CrAssphage, or HF183 were detected in any of the H2 mesocosms. Thus, H2 is not shown here.
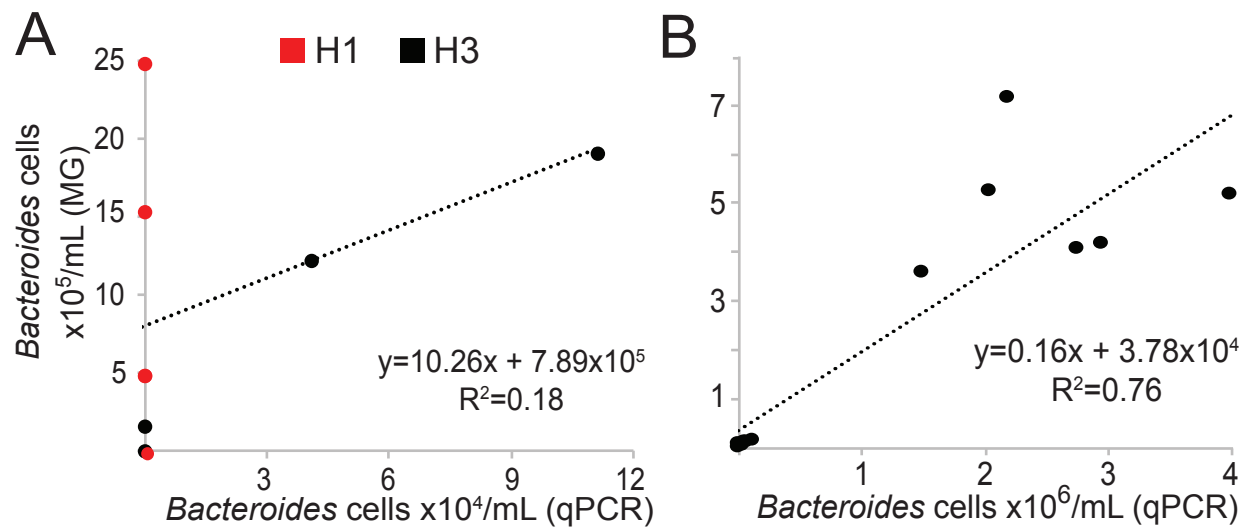

Figure 3: Correlation between qPCR and metagenome-based abundance estimates of MST markers and their reference genome counterparts. (A) Human-specific *Bacteroides* 16S (HF183) versus the absolute abundance of the reference genome *B. dorei* in the human mesocosm metagenomes. (B) Ruminant-specific *Bacteroides* 16S (RumBac) versus the absolute abundance of a contig recovered from the cow fecal inocula metagenomes carrying a perfect match to the RumBac assay in the cow mesocosm metagenomes. Absolute abundances for (A) and (B) are expressed as the number of *Bacteroides* cells/mL.