

1 **Exploiting bacterial whole-genome sequencing data for the evaluation of**
2 **diagnostic assays: *Campylobacter* species identification as a case study**

3

4 Melissa J. Jansen van Rensburg^{a,b}, Craig Swift^{b,c}, Alison J. Cody^{a,b}, Claire Jenkins^{b,c},
5 and Martin C.J. Maiden^{a,b}

6

7 Department of Zoology, University of Oxford, Oxford, United Kingdom^a; NIHR Health
8 Protection Research Unit in Gastrointestinal Infections, UK^b; Gastrointestinal Bacteria
9 Reference Unit, Public Health England, London, United Kingdom^c;

10

11 Running title: *In silico* evaluation of diagnostic assays.

12

13 # Address correspondence to Martin C.J. Maiden, martin.maiden@zoo.ox.ac.uk

14 **ABSTRACT**

15 The application of whole-genome sequencing (WGS) to problems in clinical
16 microbiology has had a major impact on the field. Clinical laboratories are now using
17 WGS for pathogen identification, antimicrobial susceptibility testing, and epidemiological
18 typing. WGS data also represents a valuable resource for the development and
19 evaluation of molecular diagnostic assays, which continue to play an important role in
20 clinical microbiology. To demonstrate this application of WGS, the current study used
21 publicly available genomic data to evaluate a duplex real-time PCR (RT-PCR) assay
22 that targets *mapA* and *ceuE* for the detection of *Campylobacter jejuni* and
23 *Campylobacter coli*, leading global causes of bacterial gastroenteritis. *In silico* analyses
24 of *mapA* and *ceuE* primer and probe sequences from 1,713 genetically diverse *C. jejuni*
25 and *C. coli* genomes, supported by RT-PCR testing, indicated that the assay was
26 robust, with 1,707 (99.7%) isolates correctly identified. The high specificity of the
27 *mapA/ceuE* assay was the result of interspecies diversity and intraspecies conservation
28 of the target genes in *C. jejuni* and *C. coli*. Rare instances of a lack of specificity among
29 *C. coli* isolates were due to introgression in *mapA* or sequence diversity in *ceuE*. The
30 results of this study illustrate how WGS can be exploited to evaluate molecular
31 diagnostic assays using publicly available data, online databases, and open source
32 software.

33 INTRODUCTION

34 Accurate and timely diagnosis of infectious diseases is a cornerstone of clinical
35 microbiology. Notwithstanding the ongoing importance of conventional culture in many
36 settings, molecular diagnostics have markedly improved pathogen detection and
37 identification (1). The most recent development in this area is the application of whole-
38 genome sequencing (WGS) to problems in clinical microbiology (2-5). Although WGS is
39 transforming the field, genomics and rapid molecular tests have complementary roles to
40 play in diagnostic microbiology, particularly in resource-limited environments.

41

42 Since their introduction in the 1980s, nucleic acid amplification tests (NAATs), including
43 multiplex assays that facilitate syndrome-driven diagnosis, have come to be widely used
44 in bacteriology laboratories (1). In particular, multiplex NAATs are becoming
45 increasingly popular for the identification of gastrointestinal pathogens, which include a
46 wide range of viruses, bacteria, and parasites (6, 7). Many NAATs have, however, been
47 designed using representative nucleotide sequences from a limited number of isolates.
48 During the pre-WGS era, the performance of NAATs could not be examined at the
49 population level because the requisite large isolate collections were challenging to
50 assemble, and primer sequences were difficult to determine using Sanger sequencing.

51

52 The recent increase in WGS has generated an abundance of publicly available genomic
53 data that has the potential to improve the development and evaluation of NAATs and
54 other molecular diagnostics (8). At the time of writing, growing numbers of assembled
55 bacterial genomes were becoming available in public repositories, such as NCBI

56 (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html); however,
57 the majority of WGS data were available only as unassembled short reads. This limited
58 their use to laboratories with bioinformatics expertise and resources. The PubMLST
59 databases (<http://pubmlst.org>) address this issue by making large numbers of *de novo*
60 assembled bacterial genomes publicly available through a web interface with analysis
61 tools (9, 10). As of September 2016, the Ribosomal Multilocus Sequence Typing
62 (rMLST) Database (<http://pubmlst.org/rmlst/>) (11) contained over 180,000 assembled
63 bacterial genomes, which corresponded to more than 4,500 bacterial species. Similarly,
64 an increasing number of species-specific PubMLST databases are also being populated
65 with WGS data.

66
67 The current study demonstrates how WGS data can be exploited to evaluate diagnostic
68 assays. *Campylobacter*, a leading cause of bacterial gastroenteritis (12), was used as
69 an exemplar. As *Campylobacter* is difficult to culture and identify, NAATs have become
70 a popular tool for the diagnosis of campylobacteriosis (7, 13, 14); however, the extent to
71 which existing assays are affected by the high levels of genetic diversity common
72 among clinical isolates (15), or introgression, that is the transfer of DNA between
73 *Campylobacter* species (16-18), is unknown. The case study presented here is a duplex
74 Taqman real-time PCR (RT-PCR) for identification of *Campylobacter jejuni* and
75 *Campylobacter coli* (19). Developed in the pre-WGS era using single gene sequences,
76 the assay and variations thereof have been used for routine species identification (19-
77 21), studies of *Campylobacter* from humans (22-25), animals (26-35), and the
78 environment (36), and for outbreak investigations (37). For *C. jejuni*, the RT-PCR target

79 is *mapA*, which encodes a putative outer membrane lipoprotein (38) shown to be
80 immunogenic in chickens (39, 40). For *C. coli*, the target is *ceuE*, which encodes a
81 periplasmic binding protein involved in iron scavenging (41). As *mapA* and *ceuE* are
82 present in both organisms, the species specificity of the assay is contingent on the
83 conservation of primer- and probe-binding sequences. Accordingly, the *mapA/ceuE* RT-
84 PCR provided an opportunity to explore the utility of genomics and population genetics
85 approaches for *in silico* evaluations of diagnostic assays.

86 **MATERIALS AND METHODS**

87 **WGS data**

88 Best and colleagues (19) validated the *mapA/ceuE* assay using clinical *Campylobacter*
89 isolates from the United Kingdom (UK). As *Campylobacter* genotypes circulating in
90 Oxfordshire are representative of the UK (15, 25, 42, 43) and other high-income
91 countries (<http://pubmlst.org/campylobacter/>), the Oxfordshire sentinel surveillance
92 collection (15) was identified as an appropriate source of WGS data for this study. WGS
93 data from 1,724 *Campylobacter* isolates were accessed via the *Campylobacter*
94 *jejuni/coli* PubMLST database (<http://pubmlst.org/campylobacter/>). These
95 *Campylobacter* comprised all single patient isolates recovered in Oxfordshire between
96 June 2011 and June 2013.

97
98 **Genome annotation and data extraction**

99 The autotagger functionality within the PubMLST Bacterial Isolate Genome Sequence
100 Database (BIGSDB) software (9) was used to identify *mapA* (PubMLST locus id
101 CAMP0952) and *ceuE* (CAMP1271), and the seven multilocus sequence typing (MLST)
102 (44, 45) and 52 rMLST loci (11). Sequences with $\geq 98\%$ identity and $\geq 98\%$ alignment to
103 existing alleles were annotated automatically. Using curation tools available in
104 PubMLST, predicted sequences with 70-98% identity to existing alleles were aligned at
105 the nucleotide and amino acid levels to the closest match in the database. Following
106 visual inspection of the alignments, complete coding sequences were added to the
107 database. Those with internal stop codons were 'flagged', that is highlighted in the
108 database, and marked as 'visually checked'. Allelic data and corresponding nucleotide

109 sequences, MLST-defined sequence types (STs), clonal complexes, and ribosomal STs
110 (rSTs) were exported from the database using the BIGSdb data export plugin (9).

111

112 **Isolate diversity and species identification**

113 The allelic diversity of the MLST and rMLST data was determined using the bias-
114 corrected version of Simpson's index of diversity (D) (46, 47) with 95% confidence
115 intervals (CIs) (48). Possible values of D ranged from 0 (no diversity) to 1 (maximum
116 diversity). The distribution of MLST clonal complexes was compared to that observed
117 for 3,349 human disease isolates recovered in Oxfordshire between 2003 and 2009
118 (42). Study isolates were assigned to species groups using rMLST (11). For this
119 analysis, concatenated nucleotide sequences of unique rSTs (~20,780 bp) were aligned
120 with MAFFT version 7.037b (49). The memory requirements for Maximum Likelihood
121 (ML) analysis of the study dataset exceeded that of a standard installation of MEGA
122 version 5.05; therefore, a ML phylogeny was generated on a Linux server with MEGA-
123 CC version 7.0 (50), using the General Time Reversible model with gamma-distributed
124 rates plus invariant sites, with 500 bootstrap replicates (51). This analysis required
125 knowledge of the command line and took nine days. As usability and computational
126 speed were considered important factors in this study, the ML phylogeny was compared
127 to a Neighbor-joining tree (52) reconstructed in MEGA version 5.05 (51) with the Kimura
128 2-parameter model (53) using 1,000 bootstrap replicates. At the population level, *C. coli*
129 segregates into three clades (54), and additional rMLST analyses were carried out to
130 resolve the assignment of a subset of isolates to these groups. The approach described
131 above was used to compare rSTs of interest to a reference set of 15 *C. coli* genomes

132 representative of the three clades, with the ML phylogeny generated using the Tamura-
133 Nei model with gamma-distributed rates plus invariant sites, with 500 bootstrap
134 replicates (Table S2) (16, 55, 56) (Table S2).

135

136 ***In silico* assay evaluation**

137 Nucleotide sequence alignments of unique *mapA* and *ceuE* alleles were generated as
138 for the rMLST phylogeny, and regions corresponding to the forward primer, probe, and
139 reverse primer (19) were extracted. Primer and probe nucleotide sequence fragments
140 were aligned and concatenated, and unique combinations were assigned allele
141 numbers in order of discovery.

142

143 **RT-PCR confirmation of *in silico* evaluation results**

144 Archived genomic DNA and bacterial cultures were available for the study isolates (15),
145 which facilitated RT-PCR confirmation of the *in silico* evaluation results. Representative
146 isolates ($n = 124$) were chosen for RT-PCR such that each unique *mapA* and *ceuE*
147 forward primer, probe, and reverse primer combination was tested at least once, with
148 the subset also representative of the genetic diversity of the study dataset. For isolates
149 with insufficient archived genomic DNA ($n = 5$), glycerol stocks of single-colony cultures
150 were inoculated onto Columbia agar with horse blood (Oxoid Ltd., Basingstoke, UK) and
151 incubated in a microaerobic atmosphere at 42°C for 48 hours. Boiled cell lysates were
152 prepared from single colonies as previously described (19). RT-PCR was carried out
153 according to the method of Best *et al.* (19), and positive results were defined as those
154 with cycle threshold (C_T) values ranging from 12 to 30.

155

156 **Genetic diversity, introgression, and selection in RT-PCR targets**

157 Individual *mapA* and *ceuE* nucleotide sequence alignments and gene phylogenies were
158 generated as described for rMLST. The *mapA* ML phylogeny was constructed using the
159 Tamura 3-parameter model with gamma distributed rates with 500 bootstrap replicates,
160 and the same parameters were used for *ceuE*, with the addition of invariant sites.

161 Nucleotide sequences were translated using MEGA version 5.05 (51) and allele
162 numbers were assigned to unique protein sequences. STRUCTURE (57), a Bayesian
163 clustering algorithm, was used to characterise introgression in *mapA* and *ceuE*, as
164 previously described (17, 18). Isolates were probabilistically assigned to species using
165 the linkage model, which adjusts for linkage disequilibrium between nucleotides (58).

166 The model was run with default settings for 10,000 burn-in iterations and 10,000
167 additional iterations, assuming a population number (k) of 2. Putative mosaic alleles
168 were identified as those with ≤ 0.75 probability of belonging to either *C. jejuni* or *C. coli*
169 (18). Site-by-site frequencies generated by STRUCTURE were used to identify nucleotide
170 sequence fragments with different ancestries in putative mosaic alleles (18, 58). After
171 putative recombinant alleles were excluded, within- and between-group p -distances
172 were calculated for *C. jejuni*- and *C. coli*-specific gene and protein sequences using
173 DnaSP version 5.10 (59). Species-specific synonymous and non-synonymous
174 substitution rates (d_N/d_S) were calculated for *mapA* and *ceuE* alleles encoding full-length
175 protein sequences using SNAP version 2.1.1 (www.hiv.lanl.gov) (60).

176 RESULTS

177 Isolate diversity and species identification

178 Complete nucleotide sequences for *mapA* and *ceuE*, and the MLST and rMLST loci,
179 were obtained from 1,713/1,724 (99.4%) isolates (Table S1), excluding those with:
180 incomplete MLST and/or rMLST profiles ($n = 8$); misassembled *mapA* or *ceuE*
181 sequences ($n = 1$); or multiple alleles at any of the rMLST loci ($n = 2$), which is an
182 indicator that a mixed culture may have been sequenced. The isolates included can be
183 accessed via the *Campylobacter jejuni/coli* PubMLST isolate database and are grouped
184 in the “*mapA/ceuE* Evaluation” project. The collection comprised 293 STs ($D = 0.974$
185 [95% CIs 0.972-0.976]) and 597 rSTs ($D = 0.989$ [95% CIs 0.988-0.991]). The STs were
186 assigned to 33 clonal complexes, with proportions similar to those observed previously
187 in Oxfordshire (Fig. 1A) (15, 42). Species designations were inferred from the ML and
188 Neighbor-joining rMLST phylogenies (11), which aggregated rSTs into identical species
189 groups. As the same was true for all paired phylogenies, only the Neighbor-joining trees
190 are presented here. *C. jejuni* accounted for 1,521 (88.8%) isolates and *C. coli* for the
191 remaining 192 (11.2%) (Fig. 1B). Two *C. coli* rSTs, rST-398 ($n = 2$) and rST-4701
192 ($n = 1$), were distinct from the other *C. coli* sequences and occurred at the tip of a long
193 branch (Fig. 1B). Further rMLST analyses indicated that these isolates belonged to
194 *C. coli* clade 3 (Fig. 1C).

195

196 *In silico* assay evaluation

197 There were 72 *mapA* alleles of 645 bp and 126 *ceuE* alleles of 990-994 bp represented
198 in the isolate collection. Differences in *ceuE* allele lengths were mainly due to variation

199 in three homopolymeric tracts, which resulted in internal stop codons in 11 *C. jejuni*-
200 specific alleles ($n = 20$) (Table 1). These alleles were 'flagged' and marked as 'visually
201 checked' in the database. To evaluate assay specificity, primer- and probe-binding
202 sequences were extracted from *mapA* and *ceuE* and analysed in detail.

203

204 ***mapA***

205 Twenty-three unique *mapA* primer and probe combinations were identified among the
206 study isolates: twelve were present only in isolates designated as *C. jejuni*; nine in
207 *C. coli*; and two were present in both species. Two distinct groups, consistent with
208 microbiological species, were evident from the nucleotide sequence alignment of these
209 unique combinations (Fig. 2A). Primer and probe sequences were conserved among
210 *C. jejuni* isolates. The predominant primer and probe combination was detected in 1,166
211 (76.7%) isolates and was identical to the published sequences (19). Sequence variation
212 among divergent *C. jejuni* combinations was limited to between one and five
213 polymorphisms across the three regions. *C. coli* sequences were also conserved but
214 were divergent from *C. jejuni* combinations, differing from the published sequences (19)
215 at up to 18 sites (Fig. 2A); however, four *C. coli* isolates carried non-specific primer and
216 probe combinations (Table 2). Two combinations, each present in a single *C. coli*
217 isolate, corresponded to predominant *C. jejuni*-specific alleles 1 and 2 (Fig. 2A). The
218 remaining two non-specific combinations were composites of *C. coli* and *C. jejuni*
219 sequences (Table 2; Fig 2A).

220

221 *ceuE*

222 The 26 *ceuE* primer and probe combinations identified among the study isolates were
223 all species-specific: 21 were present only in *C. jejuni* and five in *C. coli*. Primer and
224 probe combinations were also stratified by species at the nucleotide sequence level
225 (Fig. 2B). *C. coli* sequences were highly conserved, with 186 (96.9%) isolates identical
226 to the published primer and probe sequences (19). Nucleotide variation was limited to
227 between one and eight polymorphisms per primer and probe combination, the majority
228 of which occurred in alleles 25 ($n = 2$) and 26 ($n = 1$), which were present in the clade 3
229 *C. coli* isolates (Table 2; Fig. 2B). In contrast, *C. jejuni* combinations were divergent
230 from the published sequences (19), containing between 10 and 13 nucleotide
231 differences across the three regions. *C. jejuni* were also more evenly distributed across
232 primer and probe sequences, with eight combinations accounting for 96.3% of isolates
233 in contrast to a single combination accounting for 96.9% of *C. coli* isolates (Fig. 2B).

234

235 Predicted assay performance and RT-PCR confirmation

236 Predicted species designations based on the results of the *in silico* evaluation were
237 consistent with rMLST species assignments for 1,707/1,713 (99.7%) isolates,
238 corresponding to 1,521 (100%) *C. jejuni* and 186 (96.9%) *C. coli*. These results were
239 confirmed by RT-PCR testing of 124 representative isolates. *C. coli* isolates with
240 complete *C. jejuni*-specific *mapA* primer and probe sequences were *mapA*-
241 positive/*ceuE*-positive (Table 2). RT-PCR results for the *C. coli* isolate carrying *C. jejuni*-
242 specific *mapA* forward primer and probe sequences and the three clade 3 *C. coli*
243 isolates were inconclusive, as C_T values for *mapA* and *ceuE*, respectively, ranged from

244 32 to 37, exceeding the assay cut-off of 30 (19) (Table 2; Table S3). Although the study
245 was not designed to quantify the effects of primer and probe mismatches on target
246 detection, there was a correlation between the number of polymorphisms and C_T values
247 (Table S3).

248

249 **Introgression, diversity, and selection in RT-PCR targets**

250 Additional analyses were carried out at the whole-gene level to explore the impact of
251 introgression, diversity, and selection on assay specificity. Individual gene phylogenies
252 confirmed that *mapA* and *ceuE* alleles were species-specific (Fig. 3), with the exception
253 of *mapA* alleles 20 and 88, which were present in the *mapA*-positive/*ceuE*-positive
254 *C. coli* isolates (Fig. 3A; Table 2). Clade 3 *C. coli* *mapA* and *ceuE* alleles clustered with
255 the other *C. coli* sequences; however, they were distinct from clade 1 sequences and
256 were at the end of a long branch in both phylogenies, indicative of genetic divergence
257 (Fig. 3). Also noteworthy were five *C. coli* alleles that occupied intermediate positions on
258 the *mapA* phylogeny (Fig. 3A). Taken together with the interspecies transfer of alleles
259 20 and 88, these findings indicated introgression in *mapA*, which supported the results
260 of the *in silico* evaluation.

261

262 Drawing on population genetics approaches, introgression in the RT-PCR target genes
263 was formally characterised using STRUCTURE, with mixed ancestry detected only in the
264 *mapA* gene of 17 (8.9%) *C. coli* isolates. In addition to the two previously identified
265 complete gene transfers, five putative mosaic alleles were detected ($n = 15$) (Fig. 4A).
266 All imported DNA was identical to the predominant *C. jejuni* sequence. Recombination

267 breakpoints occurred within the amplified region in four introgressed alleles, of which
268 alleles 19 and 111 corresponded to composite primer and probe sequences (Fig. 4B).
269 Alleles 22 ($n = 11$) and 93 ($n = 1$) were not identified as introgressed sequences during
270 the *in silico* evaluation because the putative breakpoints occurred at the 5' end of the
271 forward primer (Fig. 4B).

272

273 For both *mapA* and *ceuE*, between-species p -distances were at least an order of
274 magnitude greater than those within species, and lower levels of diversity were
275 observed for the targeted species (Table 3). Analyses carried out at the allele level
276 indicated that gene and protein diversity was primarily due to an abundance of rare
277 alleles (Fig. S1). Average d_N/d_S ratios for *mapA* and *ceuE* were <1 (0.077-0.14) in
278 *C. jejuni* and *C. coli*, consistent with both genes being under stabilising selection;
279 however, the distribution of synonymous and non-synonymous substitutions suggested
280 species-specific differences in *mapA* and *ceuE* evolution (Fig. S2).

281 **DISCUSSION**

282 The current study demonstrates how bacterial WGS data can be used indirectly to
283 support diagnostic laboratory activities. *In silico* analyses of primer and probe
284 sequences, in conjunction with RT-PCR, confirmed that the *mapA/ceuE* assay was
285 robust. Overall, 1,707 (99.7%) isolates were correctly identified, which was similar to the
286 97.7% level of accuracy reported during test validation using conventional approaches
287 (19). Assay specificity was attributable to a combination of interspecies diversity and
288 marked intraspecies conservation within primer- and probe-binding regions, in addition
289 to over-representation of sequences identical to published primers and probes (Fig. 2).

290

291 Experimental evidence suggests that the products of *mapA* and *ceuE* play roles in
292 colonization (39) and iron-acquisition (61), respectively, in *Campylobacter*. Whole-gene
293 analyses indicated that intraspecies diversity was low at the gene and protein levels and
294 that both targets were under stabilising selection (Table 3). One possible explanation for
295 these findings is that *mapA* and *ceuE* encode essential cellular components in *C. jejuni*
296 and *C. coli*. This is supported by the results of experimental studies, in which *mapA* and
297 *ceuE* mutants showed reduced potential for chicken colonisation (39, 62). Species-
298 specific differences in the distribution of synonymous and non-synonymous
299 substitutions in *mapA* and *ceuE* suggest divergent evolution in *C. jejuni* and *C. coli* post-
300 speciation (Fig. S2), perhaps due to host niche differences (35, 54, 63-65). Interestingly,
301 the predicted protein sequences of 11 *C. jejuni*-specific *ceuE* alleles ($n = 20$) were
302 truncated due to variation in three homopolymeric tracts, one of which occurred at the 5'
303 end of the gene (Table 1). Sequencing of NCTC 11168 demonstrated that

304 homopolymeric tracts are common in the *C. jejuni* genome, with multiple variants
305 detected in clones that were otherwise indistinguishable. These hypervariable
306 sequences regulate gene expression through phase variation (66). It is possible that
307 *ceuE* may also be phase variable, although confirmation of homopolymeric tract lengths
308 was beyond the scope of the current study.

309
310 Although a small proportion of *Campylobacter* could not be identified to the species
311 level using the *mapA/ceuE* assay, both targets were universally present and no isolates
312 were incorrectly identified. Only six (3.1%) *C. coli* could not be identified, including two
313 *mapA*-positive/*ceuE*-positive isolates and four isolates with inconclusive results due to
314 late detection of *mapA* ($n = 1$) or *ceuE* ($n = 3$) (C_T values 32-37) (Table 2; Table S3).
315 Introgression by horizontal genetic transfer (HGT) was the underlying cause of *mapA*
316 detection among *C. coli* isolates (Fig. 4). HGT can result in the transfer of complete
317 genes (whole-allele replacement), or the generation of mosaic alleles. While whole-
318 allele replacements were relatively uncommon (1%), they accounted for the *mapA*-
319 positive/*ceuE*-positive *C. coli* isolates. Mosaic alleles were more prevalent (7.8%), but
320 resulted in only one inconclusive RT-PCR result. The apparent lack of introgression in
321 *ceuE* may be due to functional and combinatorial epistasis, as the gene is part of the
322 *ceuBCDE* operon, the products of which form an inner-membrane ABC transporter
323 system (41). Those isolates that could not be conclusively identified due to late
324 detection of *ceuE* corresponded to clade 3 *C. coli*. While clade 1 *C. coli* account for the
325 majority of human disease and agricultural isolates, those belonging to clades 2 and 3
326 are generally from environmental sources (54). Phylogenetic analyses showed that

327 clade 3 *ceuE* sequences were divergent from other *C. coli*-specific alleles (Fig. 3B). An
328 accumulation of mutations in the forward primer region reduced amplification efficiency
329 (Fig. 2B; Table S3), indicating a lack of assay specificity for clade 3 *C. coli*.
330
331 Taken together, the results of the *in silico* evaluation showed that the *mapA/ceuE* RT-
332 PCR assay reliably identifies *C. jejuni* and *C. coli*, while whole-gene analyses provided
333 insights into underlying reasons for the specificity of the assay. The value of these
334 findings also extends to other diagnostic assays that use *mapA* or *ceuE* as targets (20,
335 21, 67-70). In the UK and other high-income countries, the impact of RT-PCR failures
336 observed in this study would be limited because: (i) *C. coli* accounts for a small
337 proportion of human campylobacteriosis cases (12); (ii) the RT-PCR result was
338 unaffected in the majority of isolates with introgressed *mapA* alleles; and (iii) clade 2
339 and 3 *C. coli* rarely cause human disease (54). Given that signals of host association
340 are more marked than geographic signals (35), it is likely that the assay will perform well
341 in other regions that consume similar food animals to the UK; however, laboratories in
342 regions with discernible differences in *Campylobacter* epidemiology should exercise
343 caution and validate the assay prior to use.
344
345 The *in silico* approach to assay evaluation used here could be extended to other NAATs
346 or molecular diagnostic tests. Compared to Sanger sequencing, WGS represents an
347 attractive alternative for studying primer sequences, particularly those with mismatches
348 that adversely affect amplification efficiency. The approach outlined in this study could
349 be used to evaluate existing assays, or it could be applied in conjunction with primer

- 350 design software during assay development, requiring only a personal computer with
- 351 Internet access and publicly available software.

352 **ACKNOWLEDGEMENTS**

353 The authors are grateful to Dr Frances Colles, Dr Keith Jolley, Dr James Bray, and Mr
354 Katriel Cohn-Gordon for their advice during the preparation of this work.

355

356 **AUTHOR CONTRIBUTIONS**

357 MJJvR, CJ, and MCJM designed the study. MJJvR, CS, and AJC did the laboratory
358 work. MJJvR analysed and interpreted the data. MJJvR and MCJM wrote the
359 manuscript, which CS, AJC, and CJ commented on.

360

361 **FUNDING INFORMATION**

362 MJJvR was supported by the Clarendon Fund, Merton College (University of Oxford),
363 and Funds for Women Graduates. MCJM was funded by the Wellcome Trust (grant
364 number 087622). MJJvR, CS, AJC, CJ, and MCJM are affiliated to the National Institute
365 for Health Research Health Protection Research Unit (NIHR HPRU) in Gastrointestinal
366 Infections at the University of Liverpool in partnership with Public Health England (PHE),
367 in collaboration with University of East Anglia, University of Oxford and the Institute of
368 Food Research. MJJvR, AJC, and MCJM are based at the University of Oxford. CS and
369 CJ are based at PHE. The views expressed are those of the authors and not
370 necessarily those of the NHS, the NIHR, the Department of Health or Public Health
371 England. This work made use of data from the project 'Maintaining sentinel surveillance
372 for human campylobacteriosis in Oxfordshire: monitoring the impact of poultry industry
373 interventions on the burden of human disease'
374 (http://pubmlst.org/campylobacter/info/Oxfordshire_sentinel_surveillance.shtml).

376 REFERENCES

- 377 1. **Buchan BW, Ledeboer NA.** 2014. Emerging technologies for the clinical
378 microbiology laboratory. *Clinical Microbiology Reviews* **27**:783-822.
- 379 2. **Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW.** 2012. Transforming
380 clinical microbiology with bacterial genome sequencing. *Nature Reviews*
381 *Genetics* **13**:601-612.
- 382 3. **Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington**
383 **M, Holden MT, Dougan G, Bentley SD, Parkhill J, Peacock SJ.** 2012. Routine
384 use of microbial whole genome sequencing in diagnostic and public health
385 microbiology. *Plos Pathogens* **8**:e1002824.
- 386 4. **Köser CU, Ellington MJ, Peacock SJ.** 2014. Whole-genome sequencing to
387 control antimicrobial resistance. *Trends in Genetics* **30**:401-407.
- 388 5. **Robinson ER, Walker TM, Pallen MJ.** 2013. Genomics and outbreak
389 investigation: from sequence to consequence. *Genome Medicine* **5**:36.
- 390 6. **Platts-Mills JA, Liu J, Houpt ER.** 2013. New concepts in diagnostics for
391 infectious diarrhea. *Mucosal Immunology* **6**:876-885.
- 392 7. **Zhang H, Morrison S, Tang YW.** 2015. Multiplex polymerase chain reaction
393 tests for detection of pathogens associated with gastroenteritis. *Clinics in*
394 *Laboratory Medicine* **35**:461-486.
- 395 8. **Fournier PE, Dubourg G, Raoult D.** 2014. Clinical detection and
396 characterization of bacterial pathogens in the genomics era. *Genome Medicine*
397 **6**:114.
- 398 9. **Jolley KA, Maiden MC.** 2010. BIGSdb: Scalable analysis of bacterial genome
399 variation at the population level. *BMC Bioinformatics* **11**:595.
- 400 10. **Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley**
401 **KA, McCarthy ND.** 2013. MLST revisited: the gene-by-gene approach to
402 bacterial genomics. *Nature Reviews Microbiology* **11**:728-736.
- 403 11. **Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM,**
404 **Wimalaratna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC.** 2012.
405 Ribosomal Multi-Locus Sequence Typing: universal characterization of bacteria
406 from domain to strain. *Microbiology* **158**:1005-1015.
- 407 12. **Kaakoush NO, Castano-Rodriguez N, Mitchell HM, Man SM.** 2015. Global
408 Epidemiology of *Campylobacter* Infection. *Clinical Microbiology Reviews* **28**:687-
409 720.
- 410 13. **Fitzgerald C.** 2015. *Campylobacter*. *Clinics in Laboratory Medicine* **35**:289-298.
- 411 14. **On SL, Jordan PJ.** 2003. Evaluation of 11 PCR assays for species-level
412 identification of *Campylobacter jejuni* and *Campylobacter coli*. *J Clin Microbiol*
413 **41**:330-336.
- 414 15. **Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley S,**
415 **Parkhill J, Dingle KE, Bowler IC, Jolley KA, Maiden MC.** 2013. Real-time
416 genomic epidemiology of human *Campylobacter* isolates using whole genome
417 multilocus sequence typing. *J Clin Microbiol* **51**:2526-2534.
- 418 16. **Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly**
419 **DJ, Cody A, Colles FM, Strachan NJ, Ogden ID, Forbes K, French NP, Carter**
420 **P, Miller WG, McCarthy ND, Owen R, Litrup E, Egholm M, Affourtit JP,**

- 421 Bentley SD, Parkhill J, Maiden MC, Falush D. 2013. Progressive genome-wide
422 introgression in agricultural *Campylobacter coli*. *Molecular Ecology* **22**:1051-
423 1064.
- 424 17. Sheppard SK, McCarthy ND, Falush D, Maiden MC. 2008. Convergence of
425 *Campylobacter* species: implications for bacterial evolution. *Science* **320**:237-
426 239.
- 427 18. Sheppard SK, McCarthy ND, Jolley KA, Maiden MCJ. 2011. Introgression in
428 the genus *Campylobacter*: generation and spread of mosaic alleles. *Microbiology*
429 **157**:1066-1074.
- 430 19. Best EL, Powell NB, Swift C, Grant KA, Frost JA. 2003. Applicability of a rapid
431 duplex real-time PCR assay for speciation of *Campylobacter jejuni* and
432 *Campylobacter coli* directly from culture plates. *FEMS Microbiology Letters*
433 **229**:237-241.
- 434 20. Schuurman T, de Boer RF, van Zanten E, van Slochteren KR, Scheper HR,
435 Dijk-Alberts BG, Moller AV, Kooistra-Smid AM. 2007. Feasibility of a
436 molecular screening method for detection of *Salmonella enterica* and
437 *Campylobacter jejuni* in a routine community-based clinical microbiology
438 laboratory. *J Clin Microbiol* **45**:3692-3700.
- 439 21. Van Lint P, De Witte E, De Henau H, De Muynck A, Verstraeten L, Van
440 Herendael B, Weekx S. 2015. Evaluation of a real-time multiplex PCR for the
441 simultaneous detection of *Campylobacter jejuni*, *Salmonella* spp., *Shigella*
442 spp./EIEC, and *Yersinia enterocolitica* in fecal samples. *Eur J Clin Microbiol*
443 **34**:535-542.
- 444 22. Amar CF, East CL, Gray J, Iturriza-Gomara M, Maclure EA, McLauchlin J.
445 2007. Detection by PCR of eight groups of enteric pathogens in 4,627 faecal
446 samples: re-examination of the English case-control Infectious Intestinal Disease
447 Study (1993-1996). *Eur J Clin Microbiol* **26**:311-323.
- 448 23. Mason J, Iturriza-Gomara M, O'Brien SJ, Ngwira BM, Dove W, Maiden MC,
449 Cunliffe NA. 2013. *Campylobacter* infection in children in Malawi is common and
450 is frequently associated with enteric virus co-infections. *PLoS ONE* **8**:e59663.
- 451 24. O'Brien SJ, Rait G, Hunter PR, Gray JJ, Bolton FJ, Tompkins DS,
452 McLauchlin J, Letley LH, Adak GK, Cowden JM, Evans MR, Neal KR, Smith
453 GE, Smyth B, Tam CC, Rodrigues LC. 2010. Methods for determining disease
454 burden and calibrating national surveillance data in the United Kingdom: the
455 second study of infectious intestinal disease in the community (IID2 study). *BMC*
456 *Medical Research Methodology* **10**:39.
- 457 25. Sopwith W, Birtles A, Matthews M, Fox A, Gee S, Painter M, Regan M, Syed
458 Q, Bolton E. 2006. *Campylobacter jejuni* multilocus sequence types in humans,
459 northwest England, 2003-2004. *Emerg Infect Dis* **12**:1500-1507.
- 460 26. Bull SA, Allen VM, Domingue G, Jorgensen F, Frost JA, Ure R, Whyte R,
461 Tinker D, Corry JEL, Gillard-King J, Humphrey TJ. 2006. Sources of
462 *Campylobacter* spp. colonizing housed broiler flocks during rearing. *Appl Environ*
463 *Microbiol* **72**:645-652.
- 464 27. Griggs DJ, Johnson MM, Frost JA, Humphrey T, Jorgensen F, Piddock LJ.
465 2005. Incidence and mechanism of ciprofloxacin resistance in *Campylobacter*
466 spp. isolated from commercial poultry flocks in the United Kingdom before,

- during, and after fluoroquinolone treatment. Antimicrobial Agents and Chemotherapy **49**:699-707.
28. **Humphrey TJ, Jorgensen F, Frost JA, Wadda H, Domingue G, Elviss NC, Griggs DJ, Piddock LJ.** 2005. Prevalence and subtypes of ciprofloxacin-resistant *Campylobacter* spp. in commercial poultry flocks before, during, and after treatment with fluoroquinolones. Antimicrobial Agents and Chemotherapy **49**:690-698.
29. **Kalupahana RS, Kottawatta KS, Kanankege KS, van Bergen MA, Abeynayake P, Wagenaar JA.** 2013. Colonization of *Campylobacter* spp. in broiler chickens and laying hens reared in tropical climates with low-biosecurity housing. Appl Environ Microbiol **79**:393-395.
30. **Kwan PS, Birtles A, Bolton FJ, French NP, Robinson SE, Newbold LS, Upton M, Fox AJ.** 2008. Longitudinal study of the molecular epidemiology of *Campylobacter jejuni* in cattle on dairy farms. Appl Environ Microbiol **74**:3626-3633.
31. **Rapp D, Ross CM, Pleydell EJ, Muirhead RW.** 2012. Differences in the fecal concentrations and genetic diversities of *Campylobacter jejuni* populations among individual cows in two dairy herds. Appl Environ Microbiol **78**:7564-7571.
32. **Ridley A, Morris V, Gittins J, Cawthraw S, Harris J, Edge S, Allen V.** 2011. Potential sources of *Campylobacter* infection on chicken farms: contamination and control of broiler-harvesting equipment, vehicles and personnel. J Appl Microbiol **111**:233-244.
33. **Ridley AM, Allen VM, Sharma M, Harris JA, Newell DG.** 2008. Real-time PCR approach for detection of environmental sources of *Campylobacter* strains colonizing broiler flocks. Appl Environ Microbiol **74**:2492-2504.
34. **Ridley AM, Morris VK, Cawthraw SA, Ellis-Iversen J, Harris JA, Kennedy EM, Newell DG, Allen VM.** 2011. Longitudinal molecular epidemiological study of thermophilic campylobacters on one conventional broiler chicken farm. Appl Environ Microbiol **77**:98-107.
35. **Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, Brick G, Meldrum R, Little CL, Owen RJ, Maiden MCJ, McCarthy ND.** 2010. Host Association of *Campylobacter* Genotypes Transcends Geographic Variation. Appl Environ Microbiol **76**:5269-5277.
36. **Oster RJ, Wijesinghe RU, Haack SK, Fogarty LR, Tucker TR, Riley SC.** 2014. Bacterial pathogen gene abundance and relation to recreational water quality at seven Great Lakes beaches. Environmental Science and Technology **48**:14148-14157.
37. **Edwards DS, Milne LM, Morrow K, Sheridan P, Verlander NQ, Mulla R, Richardson JF, Pender A, Lilley M, Reacher M.** 2014. Campylobacteriosis outbreak associated with consumption of undercooked chicken liver pâté in the East of England, September 2011: identification of a dose-response risk. Epidemiology and Infection **142**:352-357.
38. **Stucki U, Frey J, Nicolet J, Burnens AP.** 1995. Identification of *Campylobacter jejuni* on the basis of a species-specific gene that encodes a membrane protein. J Clin Microbiol **33**:855-859.

- 512 39. **Johnson JG, Livny J, Dirita VJ.** 2014. High-throughput sequencing of
513 *Campylobacter jejuni* insertion mutant libraries reveals *mapA* as a fitness factor
514 for chicken colonization. *J Bacteriol* **196**:1958-1967.
- 515 40. **Shoaf-Sweeney KD, Larson CL, Tang X, Konkel ME.** 2008. Identification of
516 *Campylobacter jejuni* proteins recognized by maternal antibodies of chickens.
517 *Appl Environ Microbiol* **74**:6867-6875.
- 518 41. **Richardson PT, Park SF.** 1995. Enterochelin acquisition in *Campylobacter coli*:
519 characterization of components of a binding-protein-dependent transport system.
520 *Microbiology* **141 (Pt 12)**:3181-3191.
- 521 42. **Cody AJ, McCarthy NM, Wimalaratna HL, Colles FM, Clark L, Bowler IC,
522 Maiden MC, Dingle KE.** 2012. A longitudinal six-year study of the molecular
523 epidemiology of clinical *Campylobacter* isolates in Oxfordshire, UK. *J Clin*
524 *Microbiol* **50**:3193-3201.
- 525 43. **Sheppard SK, Dallas JF, MacRae M, McCarthy ND, Sproston EL, Gormley
526 FJ, Strachan NJ, Ogden ID, Maiden MC, Forbes KJ.** 2009. *Campylobacter*
527 genotypes from food animals, environmental sources and clinical disease in
528 Scotland 2005/6. *Int J Food Microbiol* **134**:96-103.
- 529 44. **Dingle KE, Colles FM, Falush D, Maiden MC.** 2005. Sequence typing and
530 comparison of population biology of *Campylobacter coli* and *Campylobacter*
531 *jejuni*. *J Clin Microbiol* **43**:340-347.
- 532 45. **Dingle KE, Colles FM, Wareing DRA, Ure R, Fox AJ, Bolton FJ, Bootsma HJ,
533 Willems RJL, Urwin R, Maiden MCJ.** 2001. Multilocus sequence typing system
534 for *Campylobacter jejuni*. *J Clin Microbiol* **39**:14-23.
- 535 46. **Hunter PR, Gaston MA.** 1988. Numerical index of discriminatory ability of typing
536 systems: an application of Simpson's index of diversity. *J Clin Microbiol* **26**:2465-
537 2466.
- 538 47. **Simpson EH.** 1949. Measurement of Diversity. *Nature* **163**:688.
- 539 48. **Grundmann H, Hori S, Tanner G.** 2001. Determining confidence intervals when
540 measuring genetic diversity and the discriminatory abilities of typing methods for
541 microorganisms. *J Clin Microbiol* **39**:4190-4192.
- 542 49. **Katoh K, Standley DM.** 2013. MAFFT multiple sequence alignment software
543 version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772-780.
- 544 50. **Kumar S, Stecher G, Peterson D, Tamura K.** 2012. MEGA-CC: computing core
545 of molecular evolutionary genetics analysis program for automated and iterative
546 data analysis. *Bioinformatics* **28**:2685-2686.
- 547 51. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011.
548 MEGA5: molecular evolutionary genetics analysis using maximum likelihood,
549 evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**:2731-
550 2739.
- 551 52. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for
552 reconstructing phylogenetic trees. *Mol Biol Evol* **4**:406-425.
- 553 53. **Kimura M.** 1980. A simple method for estimating evolutionary rates of base
554 substitutions through comparative studies of nucleotide sequences. *Journal of*
555 *Molecular Evolution* **16**:111-120.
- 556 54. **Sheppard SK, Dallas JF, Wilson DJ, Strachan NJ, McCarthy ND, Colles FM,
557 Rotariu O, Ogden ID, Forbes KJ, Maiden MCJ.** 2010. Evolution of an

- 558 agriculture-associated disease causing *Campylobacter coli* clade: evidence from
559 national surveillance data in Scotland. PLoS ONE **5**:e15708.
- 560 55. **Chen Y, Mukherjee S, Hoffmann M, Kotewicz ML, Young S, Abbott J, Luo Y,**
561 **Davidson MK, Allard M, McDermott P, Zhao S.** 2013. Whole-genome
562 sequencing of gentamicin-resistant *Campylobacter coli* isolated from U.S. retail
563 meats reveals novel plasmid-mediated aminoglycoside resistance genes.
564 Antimicrobial Agents and Chemotherapy **57**:5398-5405.
- 565 56. **Pearson BM, Rokney A, Crossman LC, Miller WG, Wain J, van Vliet AH.**
566 2013. Complete Genome Sequence of the *Campylobacter coli* Clinical Isolate 15-
567 537360. Genome Announcements **1**:e01056-01013.
- 568 57. **Pritchard JK, Stephens M, Donnelly P.** 2000. Inference of population structure
569 using multilocus genotype data. Genetics **155**:945-959.
- 570 58. **Falush D, Stephens M, Pritchard JK.** 2003. Inference of population structure
571 using multilocus genotype data: linked loci and correlated allele frequencies.
572 Genetics **164**:1567-1587.
- 573 59. **Librado P, Rozas J.** 2009. DnaSP v5: a software for comprehensive analysis of
574 DNA polymorphism data. Bioinformatics **25**:1451-1452.
- 575 60. **Korber B.** 2000. HIV signature and sequence variation analysis, p 55-72. *In*
576 Rodrigo AG, Learn GH (ed), Computational Analysis of HIV Molecular
577 Sequences. Kluwer Academic Publishers, Dordrecht, Netherlands.
- 578 61. **Miller CE, Williams PH, Ketley JM.** 2009. Pumping iron: mechanisms for iron
579 uptake by *Campylobacter*. Microbiology **155**:3157-3165.
- 580 62. **Palyada K, Threadgill D, Stintzi A.** 2004. Iron acquisition and regulation in
581 *Campylobacter jejuni*. J Bacteriol **186**:4714-4729.
- 582 63. **McCarthy ND, Colles FM, Dingle KE, Bagnall MC, Manning G, Maiden MC,**
583 **Falush D.** 2007. Host-associated genetic import in *Campylobacter jejuni*. Emerg
584 Infect Dis **13**:267-272.
- 585 64. **Rosef O, Gondrosen B, Kapperud G, Underdal B.** 1983. Isolation and
586 characterization of *Campylobacter jejuni* and *Campylobacter coli* from domestic
587 and wild mammals in Norway. Appl Environ Microbiol **46**:855-859.
- 588 65. **Waldenström J, Broman T, Carlsson I, Hasselquist D, Achterberg RP,**
589 **Wagenaar JA, Olsen B.** 2002. Prevalence of *Campylobacter jejuni*,
590 *Campylobacter lari*, and *Campylobacter coli* in different ecological guilds and
591 taxa of migrating birds. Appl Environ Microbiol **68**:5911-5917.
- 592 66. **Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D,**
593 **Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV,**
594 **Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM,**
595 **van Vliet AH, Whitehead S, Barrell BG.** 2000. The genome sequence of the
596 food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences.
597 Nature **403**:665-668.
- 598 67. **de Boer RF, Ott A, Guren P, van Zanten E, van Belkum A, Kooistra-Smid**
599 **AM.** 2013. Detection of *Campylobacter* species and *Arcobacter butzleri* in stool
600 samples by use of real-time multiplex PCR. J Clin Microbiol **51**:253-259.
- 601 68. **de Boer RF, Ott A, Kesztyus B, Kooistra-Smid AM.** 2010. Improved detection
602 of five major gastrointestinal pathogens by use of a molecular screening
603 approach. J Clin Microbiol **48**:4140-4146.

- 604 69. **Fukushima H, Katsube K, Tsunomori Y, Kishi R, Atsuta J, Akiba Y.** 2009.
605 Comprehensive and rapid real-time PCR analysis of 21 foodborne outbreaks.
606 International Journal of Microbiology **2009**:917623.
607 70. **McAuliffe GN, Anderson TP, Stevens M, Adams J, Coleman R,**
608 **Mahagamasekera P, Young S, Henderson T, Hofmann M, Jennings LC,**
609 **Murdoch DR.** 2013. Systematic application of multiplex PCR enhances the
610 detection of bacteria, parasites, and viruses in stool samples. Journal of Infection
611 **67**:122-129.
612

613 **FIGURE LEGENDS**

614 **Fig. 1** Genetic diversity and species identification of 1,713 *Campylobacter* genomes
615 from Oxfordshire human disease isolates (2011-2013). (A) Frequency distribution of
616 major clonal complexes ($n \geq 10$) among 3,349 *Campylobacter* isolates from human
617 disease cases in Oxfordshire (2003-2009) typed using MLST (white) (42), and the 1,713
618 genomes included in this study (black). UA, STs unassigned to a clonal complex; Cj,
619 *C. jejuni* (blue); Cc, *C. coli* (yellow). (B) Neighbor-joining tree based on concatenated
620 nucleotide sequences of unique rMLST profiles ($n = 597$) identified among the study
621 isolates. ▲, putative clade 3 *C. coli*. (C) Neighbor-joining tree based on concatenated
622 nucleotide sequences of rMLST profiles of 15 representative isolates belonging to
623 *C. coli* clades 1 (yellow), 2 (orange), and 3 (pink), and three putative clade 3 isolates
624 identified in this study (▲).

625
626 **Fig. 2** Genetic diversity of *mapA* (A) and *ceuE* (B) primer and probe sequences.
627 Published primer (cream) and probe (orange) sequences (19) are shown above
628 concatenated nucleotide sequence alignments of unique combinations identified in
629 genomes from campylobacteriosis cases in Oxfordshire (2011-2013). Dots represent
630 conserved nucleotides. Numbers above the published primers and probes indicate
631 nucleotide positions relative to the complete gene, with breaks between regions marked
632 (▼). Adjacent histograms indicate frequencies of combinations in *C. jejuni* (blue) and
633 *C. coli* (yellow). *, complete *C. jejuni*-specific combination detected in a single *C. coli*
634 isolate; ●, composite non-specific combination detected in a single *C. coli* isolate.

635

636 **Fig. 3** Neighbor-joining trees showing relationships among 72 *mapA* (A) and 126 *ceuE*
637 (B) unique gene sequences from *Campylobacter* genomes from cases of human
638 disease in Oxfordshire (2011-2013). Adjacent Venn diagrams indicate the number of
639 species-specific and shared alleles. X →, *C. jejuni*-specific allele detected in *C. coli*,
640 where 'X' indicates the allele number; ●, putative introgressed alleles; ▲, putative clade
641 3 *C. coli* alleles. Key: *C. jejuni*, blue; *C. coli*, yellow; shared alleles, grey.

642
643 **Fig. 4** Characterisation of introgression in *mapA* with STRUCTURE. (A) Probabilistic
644 assignment of study isolates to species based on analysis of *mapA* nucleotide
645 sequences using the linkage model. Putative mosaic sequences (*m*) were identified as
646 those with ≤0.75 probability of belonging to either *C. jejuni* or *C. coli*. Each isolate is
647 represented by a vertical line, with shading indicative of the proportion attributed to
648 *C. jejuni* (black) or *C. coli* (grey) ancestry. The dashed white line indicates the species
649 boundary as determined by rMLST. (B) Recombination breakpoints in putative mosaic
650 alleles were inferred using site-by-site nucleotide ancestries generated by STRUCTURE.
651 Bar plots represent individual putative mosaic sequences, with whole-gene allele
652 numbers and corresponding primer and probe combinations shown in bold and
653 parentheses, respectively. Vertical lines represent individual nucleotides with shading
654 indicative of ancestry as in (A). Dashed white lines demarcate the region amplified by
655 *mapA* primers (19).

656 **TABLES**657 **Table 1** Details of *ceuE* alleles with internal stop codons identified among *C. jejuni*

658 isolates

Polymorphism	Nucleotide position	Putative effect on protein	PubMLST gene allele	ST/CC/rST ^a (n)
T(8 -> 7)	34	Truncation	288	5756/UA ^b /263 (1)
			290	2844/ST-460/325 (2)
			291	48/ST-48/106 (2);
				48/ST-48/98 (1);
				48/ST-48/99 (1);
				520/ST-21/377 (1)
			293	2274/UA/123 (1)
			294	443/ST-443/221 (1)
			296	5707/UA/3509 (1)
			298	1932/ST-460/4596 (2)
			300	464/ST-464/7025 (1)
T(8 -> 9)	34	Truncation	295	21/ST-21/538 (1)
A(5 -> 4)	202	Truncation	289	47/ST-21/510 (3);
				3633/ST-21/510 (1)
T(6 -> 5)	483	Truncation	292	53/ST-21/460 (1)
Deletion (C)	675	Truncation	297	257/ST-257/186 (2)

^a ST, sequence type; CC, MLST-defined clonal complex; rST, ribosomal ST.^b UA, not assigned to a clonal complex.

659

660 **Table 2** Details of *C. coli* isolates with atypical *mapA/ceuE* primer and probe sequences

Atypical target	Isolate	ST (CC) ^a	Gene allele / primer and probe combination (RT-PCR result) ^b		<i>mapA/ceuE</i>
			<i>mapA</i>	<i>ceuE</i>	RT-PCR result
<i>mapA</i>	OXC7352	6973	20/2	136/22	Mixed
		(ST-1150)	(+)	(+)	
	OXC6987	825	88/1	3/22	Mixed
		(ST-828)	(+)	(+)	
	OXC6395	1487 (ST-1150)	19/22 ^d (-)	17/22 (+)	<i>C. coli</i>
<i>ceuE</i>	OXC7615	6760 (UA ^c)	111/23 ^e (late +)	139/22 (+)	Inconclusive
	OXC7241	6698 (UA)	96/19	153/25	Inconclusive
			(-)	(late +)	
	OXC7243	6698 (UA)	96/19	153/25	Inconclusive
			(-)	(late +)	
	OXC7653	6975 (UA)	114/21 (-)	183/26 (late +)	Inconclusive

^a ST, sequence type; CC, clonal complex.

^b (+), target detected; (-), target not detected; (late +), target detected after cycle 30.

^c UA, ST unassigned to a clonal complex.

^d Forward primer *C. jejuni*-specific; probe and reverse primer *C. coli*-specific.

^e Forward primer and probe *C. jejuni*-specific; reverse primer *C. coli*-specific.

661

662 **Table 3** Intra- and interspecies diversity of *mapA* and *ceuE* gene and protein
663 sequences^a

	<i>mapA</i>	<i>ceuE</i>
Gene sequence diversity		
<i>C. jejuni</i>	0.014	0.018
<i>C. coli</i>	0.023	0.005
<i>C. jejuni</i> / <i>C. coli</i>	0.232	0.13
Protein sequence diversity		
<i>C. jejuni</i>	0.009	0.015
<i>C. coli</i>	0.020	0.009
<i>C. jejuni</i> / <i>C. coli</i>	0.232	0.089

^a Putative recombinant sequences and alleles encoding truncated peptide sequences were excluded. Values indicate *p*-distances.

664





