Title: Genomic evidence for revising the *Escherichia* genus and description of *Escherichia ruysiae sp. nov.*

Running title (51/54 characters incl. spaces): Genomic evidence for revising the *Escherichia* genus

Boas C.L. van der Putten[a,b,#], Sébastien Matamoros[a], COMBAT consortium[†], Constance Schultsz[a,b]

[a]Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology, Meibergdreef 9, Amsterdam, Netherlands

[b]Amsterdam UMC, University of Amsterdam, Department of Global Health, Amsterdam Institute for Global Health and Development, Meibergdreef 9, Amsterdam, Netherlands

[#]Corresponding author, email: boas.vanderputten@amsterdamumc.nl

[†]Members listed in the appendix.

Abstract (223/250 words)

The *Escherichia*/*Shigella* genus comprises four *Escherichia* species, four *Shigella* species, and five lineages currently not assigned to any species, termed '*Escherichia* cryptic clades' (numbered I, II, III, IV and VI). Correct identification of *Escherichia* cryptic clades is strongly

hindered by the indeterminate taxonomy of this genus. Furthermore, little is known about the cryptic clades as reports of genomic data are scarce. Hence, we searched public databases for whole-genome sequences of *Escherichia* cryptic clades and characterized these. Following a genomic analysis of the *Escherichia/Shigella* genus, we also describe a new *Escherichia* species: *Escherichia ruysiae sp. nov* (type strain OPT1704^T = NCCB 100732^T = NCTC 14359^T) and provide a closed genome assembly based on Illumina and Oxford Nanopore Technologies sequencing. We screened 79,911 Sequence Read Archive *Escherichia* records and detected 357 cryptic clade strains (0.44%). Based on average nucleotide identity, these strains should be grouped in seven distinct species: 1) *E. coli*, *Shigella spp.* and clade I; 2) Clade II; 3) *Escherichia ruysiae sp. nov*. (formerly clades III and IV); 4) *E. marmotae* (formerly clade V); 5) Clade VI; 6) *E. albertii* and 7) *E. fergusonii*. Notably, half of the clade I strains carried genes encoding shiga toxin, while ESBL- and carbapenemase-encoding strains were also found.

In conclusion, we provide an improved overview of the *Escherichia/Shigella* genus and advance our understanding of *Escherichia* cryptic clades.

Importance (107/150 words)

Correct definition and identification of bacterial species is essential for clinical and research purposes. Groups of *Escherichia* strains - "*Escherichia* cryptic clades" - have not been assigned to species, which causes misidentification of these strains. The significance of our research is threefold. First, we detect 357 cryptic clade strains, many more than previously known. This can serve as a resource for other researchers. Second, we show how these cryptic clades should be

41 assigned to existing or newly defined species. This could improve identification of cryptic clade
42 strains and *Escherichia* species. Finally, we characterize the genomes in detail, revealing
43 virulence genes encoded in the cryptic clade I genomes.

44

45 **Main text excluding references, legends, tables: 3476/5000 words**

46 Introduction

47 Within the *Escherichia* genus, four species are recognized; *E. coli*, *E. fergusonii* (Farmer et al.,
48 1985), *E. albertii* (Huys, Cnockaert, Janda, & Swings, 2003) and most recently, *E. marmotae* (Liu
49 et al., 2015). *E. coli*, *E. fergusonii* and *E. albertii* have been associated with animal and human
50 disease (Ooka et al., 2012; Russo & Johnson, 2003; Savini et al., 2008), while little is known
51 about the clinical relevance and characteristics of the most recently described *Escherichia*
52 species *E. marmotae*. Some *Escherichia* strains cannot be assigned to any of the four existing
53 species. Based on genetic data, these strains cluster into six groups, which were termed
54 '*Escherichia* cryptic clades' (Gangiredla et al., 2018; S. T. Walk et al., 2009). Recently, cryptic
55 clade V was formally recognized as a separate species (*E. marmotae*) (Liu et al., 2015), leaving
56 five cryptic clades that have not been recognized as species. As these have not yet received
57 entries in taxonomic databases, isolates representing cryptic clades might be misidentified by
58 tools that identify bacterial species based on whole-genome sequence (WGS). Also, the
59 abundancy of these cryptic clade strains in public databases such as RefSeq (curated repository
60 of reference sequences) or SRA (repository of raw sequence data) is unknown.

61 Therefore, we have addressed two questions in this study:

62 1) How abundant are *Escherichia* cryptic clades in the RefSeq and Sequence Read Archive (SRA)
63 databases?

64 2) Can the *Escherichia* cryptic clades be assigned to existing or new species based on genomic
65 analysis?

66 Finally, we describe a novel *Escherichia* species, *Escherichia ruysiae sp. nov.*, encompassing the
67 former cryptic clades III and IV.

68

69 Results

70 **Assessing abundance of cryptic clade strains in RefSeq**

71 We downloaded all 10,824 available *Escherichia* genomes from the RefSeq database and
72 identified 92 cryptic clade genomes using ClermonTyper (Beghain, Bridier-Nahmias, Le Nagard,
73 Denamur, & Clermont, 2018). Since ClermonTyper is only able to detect strains belonging to
74 known cryptic clades, we constructed a neighbor-joining phylogeny of the 10,824 RefSeq
75 *Escherichia* strains using mash (Ondov et al., 2016) and rapidNJ (Simonsen, Mailund, &
76 Pedersen, 2008) to detect cryptic clade strains missed by ClermonTyper. This analysis revealed
77 two additional cryptic clade strains, bringing the total of cryptic clade strains in RefSeq to 94, or
78 0.9% of all *Escherichia* genomes in RefSeq.

79 We then screened 79,911 Sequence Read Archive (SRA) *Escherichia* entries for presence of
80 cryptic clade genomes. In short, we performed a rough assembly of the SRA data using SKESA
81 and used fastANI to detect cryptic clade strains, using RefSeq cryptic clade genomes as

82 references. 757 putative cryptic clade genomes were included in a neighborhood-joining
83 phylogeny and were checked using ClermonTyper. This led to the discovery of 357 cryptic clade
84 genomes in the SRA, or 0.45% of SRA *Escherichia* assemblies. We evaluated SKESA assemblies
85 using Quast (Gurevich, Saveliev, Vyahhi, & Tesler, 2013), and excluded 13 assemblies due to
86 inadequate assembly qualities or aberrant assembly length. Since there are likely duplicates
87 between SRA and RefSeq cryptic clade genomes, we chose to continue with the 344 cryptic
88 clade genomes from the SRA. An overview of the process is provided in figure 1.

89

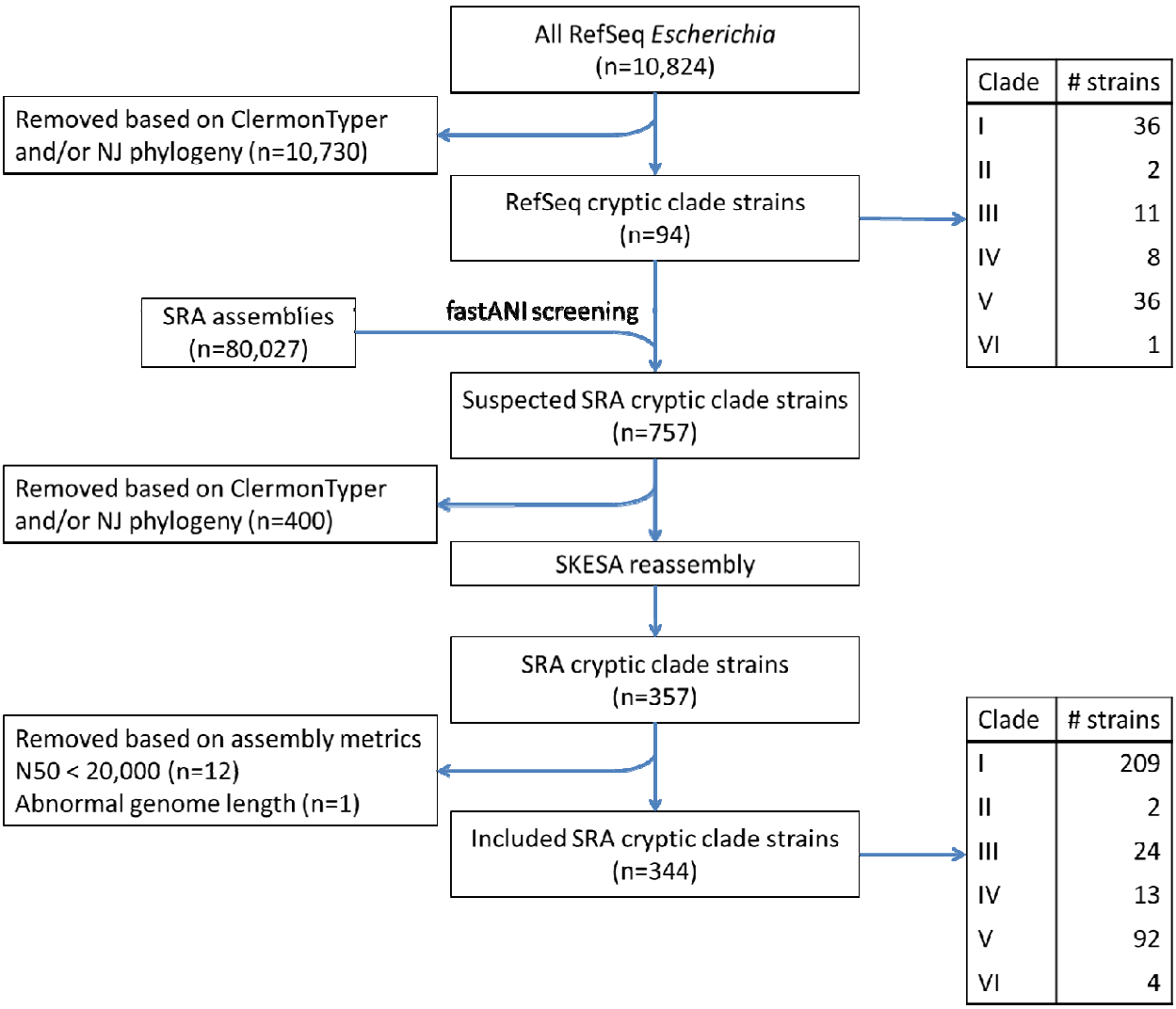

Figure 1. Flowchart of inclusion process of cryptic clade strains. Only the well-assembled SRA cryptic clade strains (lower table on right hand side) are used in subsequent analyses.

Genomic characterization of cryptic clade genomes from SRA

To visualize phylogenetic relationships between the closely related *Escherichia* and *Shigella* genera, we included – in addition to the 344 cryptic clade genomes - 50 *E. albertii* genomes, 10 *E. fergusonii* genomes, 72 *E. coli* genomes from the ECOR collection (Ochman & Selander, 1984)

98 and 13 *Shigella spp*. genomes, all from RefSeq. We constructed a maximum-likelihood
99 phylogeny using IQ-Tree (Nguyen, Schmidt, von Haeseler, & Minh, 2015) from a SNP alignment
100 produced by kSNP3 (Gardner, Slezak, & Hall, 2015) (fig. 2).

101

102 Based on species identification using Kraken2 (Wood & Salzberg, 2014) all cryptic clade I, II, III,
103 IV and VI strains were classified as *E. coli*. 82 out of 92 *E. marmotae* (formerly cryptic clade V)
104 strains were identified correctly while the remaining *E. marmotae* strains were identified as *E.*
105 *coli*.

106

107 **Resistance and virulence genes in cryptic clade genomes**

108 Genes encoding both subunits of shiga toxin were detected in 108/209 (52%) of cryptic clade I
109 strains. In 26 cryptic clade I strains, *stx1a* and *stx1b* were detected together, in 75 strains *stx2a*
110 and *stx2b* were detected together. In 7 clade I strains only one shiga toxin-encoding gene was
111 detected. Extended-spectrum beta-lactamase (ESBL) genes were detected in 8.1% and 2.2% of
112 cryptic clade I and cryptic clade V/*E. marmotae* genomes, respectively. One clade I strain and
113 two *E. marmotae*/clade V strains harbored carbapenemase-encoding genes.

114

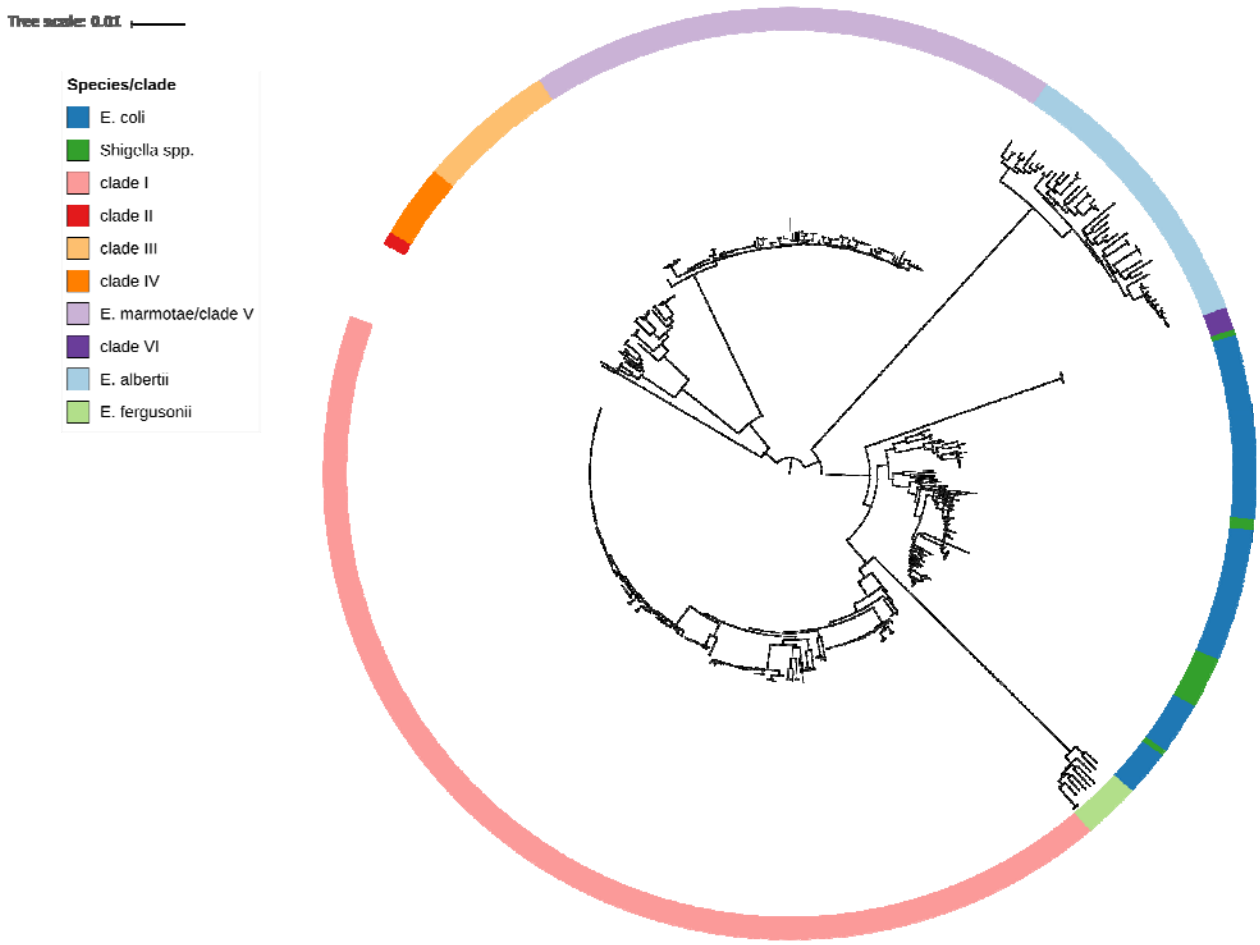

Figure 2. Midpoint-rooted maximum-likelihood phylogeny of 489 *Escherichia* and *Shigella* strains based on a core SNP alignment of 3445 positions. Nodes with bootstrap values under 95 were collapsed. Visualized in iTOL (Letunic & Bork, 2019).

Assignment of cryptic clades to existing or new species

Based on a 95% average nucleotide identity (ANI)(Chun et al., 2018; Goris et al., 2007) 489 of the included *Escherichia* and *Shigella* strains should be assigned to seven discrete species with no relationships between these groups (table 1 and fig. 3). *E. coli* and cryptic clade I showed mean ANI values of 96.0%, indicating *E. coli* and cryptic clade I should be assigned into a single

species, possibly separated into subspecies. Genomically, *Shigella spp*. belong to *E. coli* as a mean ANI is 97.6% with *E. coli*. This is in line with earlier reports(Pettengill, Pettengill, & Binet, 2016).

Cryptic clade III and IV should be assigned to a single species, as the mean ANI values between strains from these groups was 96.6%. Based on ANI analysis cryptic clade II, cryptic clade V/*E. marmotae*, cryptic clade VI, *E. fergusonii* and *E. albertii* should all be assigned to separate species.

Table 1. Mean ANI values between groups calculated using fastANI, for 489 included *Escherichia* and *Shigella* strains. Values higher than the threshold for species delineation (95%) are in bold.

	Clade I	Clade II	Clade III	Clade IV	Clade V/*E. marmotae*	Clade VI	*E. albertii*	*E. coli*	*E. fergusonii*	*Shigella*
Clade I	**98.5**	91.6	92.3	92.5	91.0	94.0	89.9	**96.0**	92.1	**95.8**
Clade II		**99.0**	92.1	92.0	91.3	90.9	89.7	91.8	89.2	91.7
Clade III			**98.8**	**96.6**	92.3	91.6	89.8	92.6	89.4	92.6
Clade IV				**99.0**	92.2	91.7	89.9	92.8	89.5	92.9
Clade V/*E. marmotae*					**99.3**	90.4	89.4	91.2	88.4	91.1
Clade VI						**99.9**	89.6	94.4	90.1	94.4
E. albertii							**98.6**	90.1	88.2	90.2
E. coli								**97.7**	91.1	**97.6**
E. fergusonii									**98.7**	91.0
Shigella spp.										**98.3**

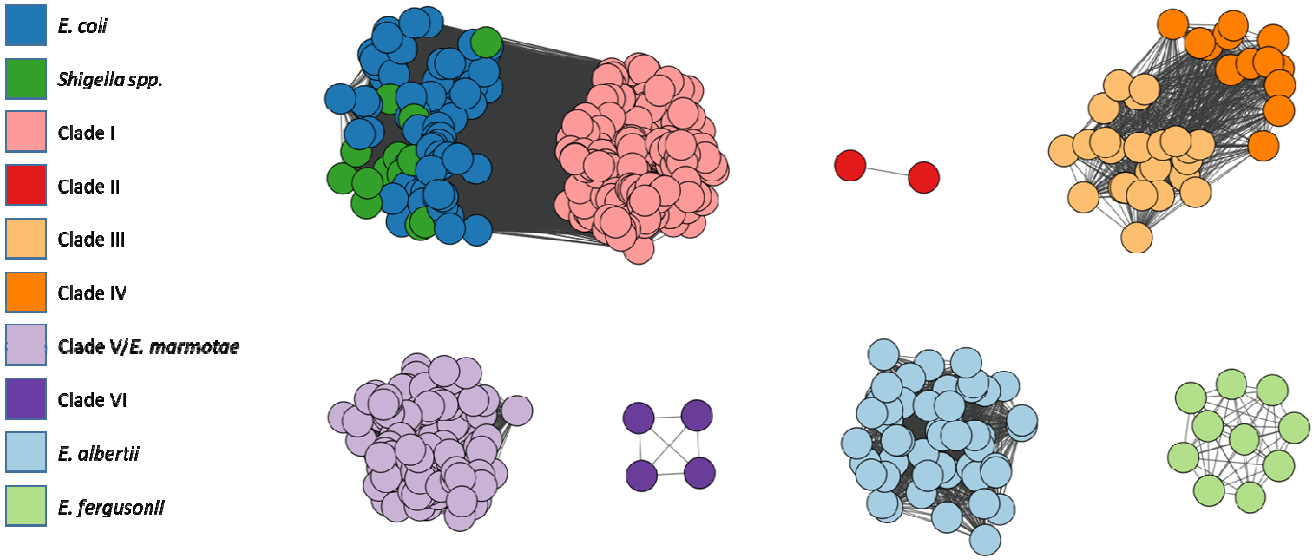

Fig 3. Network of 489 included *Escherichia* and *Shigella* strains, based on fastANI analysis. Edges represent ANI values over 95% (boundary for species delineation) between two strains.

Identification and description of *Escherichia ruysiae sp. nov.*

We discovered a cryptic clade IV strain in our collection, previously identified as *E. coli* in the COMBAT study, which investigated acquisition of ESBL-producing Enterobacteriaceae (ESBL-E) during international travel (Arcilla et al., 2017). We characterized this isolate, OPT1704^{T}, in detail.

The strain was isolated from a faecal sample provided immediately after return from a one-month journey to several countries in Asia. No ESBL-E were detected in a faecal sample provided directly before departure, suggesting the ESBL gene, and possibly strain OPT1704^{T}, were acquired during travel. The traveller reported diarrhoea during travel and did not report

152 antibiotic usage during travel. No ESBL-E were isolated in follow-up samples, suggesting loss of
153 the OPT1704^T strain or the ESBL gene within one month after return from travel.

154

155 Strain OPT1704^T formed circular, grey-white colonies on a COS sheep blood agar plate.
156 Individual cells were observed under a light microscope and were rod-shaped and
157 approximately 1 by 2 µm in size. The strain was shown to be Gram-negative, non-motile,
158 oxidase-negative and catalase-positive. The strain was capable to grow in the absence of
159 oxygen. On COS blood plates, it showed growth in the temperature range of 20-42 °C, and no
160 growth at 4 °C or at 50 °C or higher. The strain was also able to grow in NaCl concentrations
161 ranging from 0% to 6% in liquid lysogeny broth. MALDI-TOF (Bruker) and Vitek2 (BioMérieux)
162 systems both identified OPT1704^T as *E. coli* with high confidence scores (score>2 for MALDI-TOF
163 and "Excellent identification" for Vitek2).

164

165 **Whole-genome sequence analysis**

166 The whole-genome DNA sequence of strain OPT1704^T was determined using Illumina HiSeq and
167 Oxford Nanopore Technologies (ONT) sequencing platforms. The Illumina sequencing run
168 yielded a total of 6.3×10^6 paired-end reads, with a mean read length of 151 bp. Illumina reads
169 were downsampled using seqtk (version 1.3-r106, https://github.com/lh3/seqtk) to provide a
170 theoretical depth of coverage of 100X with the assumption that the OPT1704^T has a genome
171 size of approximately 5×10^6 bp. The ONT sequencing run yielded a total of 2.5×10^4 reads, with a
172 mean read length of 9078 bp before filtering. ONT reads were filtered on length and on read

173 identity using Illumina reads, leaving 1.5×10^4 reads with a mean length of 12580 bp. This
174 provided a theoretical depth of coverage of ~38X of ONT reads. Assembly using both Illumina
175 and Nanopore reads resulted in a complete genome, consisting of one circular chromosome
176 and one circular plasmid. GC content of the OPT1704^T genome was 50.6%.

177

178 Putative resistance and virulence genes were predicted from the draft genome using ABRicate
179 with the CARD (McArthur et al., 2013) and VFDB (Chen et al., 2005) databases. OPT1704^T
180 harbours 6 resistance genes that are typically plasmid-mediated in *E. coli*, associated with
181 reduced susceptibility to fluoroquinolones (*qnrS1*), aminoglycosides (*aph(6)-Id* & *aph(3'')-Ib*),
182 cephalosporins (*blaCTX-M-14*), trimethoprim (*dfrA14*) and sulphonamides (*sul2*). This is in line
183 with the reduced susceptibility to fluoroquinolones (norfloxacin, MIC: 2 mg/L and ciprofloxacin,
184 MIC: 0.5 mg/L by Vitek2), cephalosporins (cefuroxime, MIC: >32 mg/L and cefotaxime, MIC: 4
185 mg/L) and trimethoprim-sulfamethoxazole (MIC: >8 mg/L). Furthermore, several putative
186 virulence genes were predicted from the draft genome sequence associated with siderophore
187 function (*chuX*, *entS*, *fepABD*), fimbriae (*fimBCDGI*), a type II secretion system (*gspGHI*) and
188 capsular polysaccharide biogenesis (*kpsD*). The diarrhoeal symptoms of the traveller
189 contributing this strain would not be expected based on these predicted virulence genes.

190

191 Next, we calculated 16S rRNA sequence similarities, ANI values and digital DNA:DNA
192 hybridisation (dDDH) values between OPT1704^T and type strains of the four other *Escherichia*
193 species. This time, we used three separate tools to calculate ANI (fastANI, OrthoANIu and ANI

194 calculator from Enveomics) (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018;
195 Rodriguez-R & Konstantinidis, 2016; Yoon, Ha, Lim, Kwon, & Chun, 2017), since scalability was
196 no concern. 16S rRNA sequence similarities did not completely warrant assignment of
197 OPT1704^T to a novel species, but ANI analysis and dDDH did support assignment of OPT1704^T to
198 a novel species (table 2). This novel species was assigned *E. ruysiae sp. nov.* with OPT1704^T as
199 the proposed type strain. As our earlier analysis shows cryptic clade III clusters with cryptic
200 clade IV on the species level, we propose that *E. ruysiae sp. nov.* encompasses both cryptic
201 clade III and IV.

202

203 **Table 2.** Comparison of OPT1704^T whole-genome sequence with type strains of *E. albertii*, *E.*
204 *coli*, *E. fergusonii* and *E. marmotae.* In bold are the values that suggest assignment of OPT1704^T
205 to a novel species (<98.9% 16S rRNA sequence similarity, <95% ANI, <70% dDDH).

	E. ruysiae sp. nov. OPT1704^T				
	16S rRNA sequence similarity (%)	ANI (%, fastANI)	ANI (%, OrthoANIu)	ANI (%, ANI calculator Enveomics)	dDDH (%)
E. albertii NBRC107761^T	**98.5**	**90.0**	**90.0**	**89.2**	**39.8**
E. coli ATCC11775^T	**98.5**	**92.8**	**92.4**	**92.0**	**48.3**
E. fergusonii ATCC35469^T	99.4	**89.4**	**88.2**	**89.7**	**36.7**
E. marmotae DSM 28771^T	99.4	**92.2**	**92.2**	**91.4**	**47.1**

206

207 # Discussion

208 We propose an updated taxonomy of the *Escherichia* genus that includes seven *Escherichia*
209 species: 1) *E. coli,* cryptic clade I and *Shigella spp.*; 2) cryptic clade II; 3) *E. ruysiae sp. nov.*
210 (formerly cryptic clades III and IV); 4) *E. marmotae* (formerly cryptic clade V); 5) cryptic clade VI;
211 6) *E. fergusonii* and 7) *E. albertii*.

212 Contrary to the present study and possibly because a smaller dataset was used, Luo et al.
213 (2011) observed a 'genetic continuum' between *E. coli* and cryptic clades. In line with findings
214 by Walk (2015) we did not find strains that share an ANI > 95% between the seven groups,
215 indicating discrete grouping(S. Walk, 2015). Although they should be assigned to a single
216 species based on ANI analysis, a clear separation between *E. coli/Shigella spp*. and clade I is
217 observed, possibly at the subspecies level. The same holds for cryptic clades III and IV. For the
218 other cryptic clades, no significant separation in population structure is visible, indicating these
219 clades are genetically more homogeneous.

220

221 Throughout our comprehensive scanning of RefSeq and SRA, we attempted to maximize the
222 diversity of *Escherichia* genomes captured. We built the phylogeny of >10,000 RefSeq genomes
223 which included all *Escherichia* cryptic clades described in the literature. As we based our SRA
224 search on the RefSeq cryptic clade genomes, we cannot exclude that we missed novel cryptic
225 clades present in the SRA but not in RefSeq. Additionally, we had to raise the threshold with
226 which we screened the SRA for cryptic clade I strains. Possibly, strains that are genetically
227 different could have been missed by choosing this strategy.

228

229 Based on genomic analyses of the *Escherichia* genus and characterization of strain OPT1704^T,
230 we proposed a novel species, *Escherichia ruysiae sp. nov*. Analysis of 16S rRNA genes of *E.*
231 *ruysiae* and type strains of other *Escherichia spp.* did not warrant assignment to a novel species,
232 while ANI and dDDH analysis of the same genomes did. The International Journal of Systematic
233 and Evolutionary Microbiology follows results of ANI or dDDH analysis instead of 16S rRNA
234 analysis when discrepancies are observed, meaning strain OPT1704^T should be assigned to a
235 novel species (Chun et al., 2018). Assignment of cryptic clades III and IV as a novel species *E.*
236 *ruysiae* means inclusion of the species in the NCBI taxonomy database, as happened with *E.*
237 *marmotae*/cryptic clade V before. As the NCBI taxonomy serves as the reference for many
238 genomic classification tools, assigning cryptic clades III and IV to a novel species is the most
239 effective way to stimulate correct classification of these strains.

240 Additionally, the assignment of *E. ruysiae sp. nov.* could be useful for recent efforts of NCBI to
241 retrospectively correct RefSeq entries using ANI, if the entries are assigned to the wrong species
242 (https://ncbiinsights.ncbi.nlm.nih.gov/2019/02/06/correct-existing-taxonomic-info-genbank-ani-
243 analysis/). However, this effort is also based on the NCBI taxonomy, which is why it is important
244 that cryptic clades that should be assigned as separate species are formally recognized as such.

245

246 All standard identification methods we used in this study – genomic and phenotypic –
247 concluded strain OPT1704^T is an *E. coli* strain. Genomic identification methods such as Kraken2
248 often use the RefSeq database in combination with the NCBI taxonomy to classify strains.

249 Cryptic clade strains are present in the RefSeq database, but are always labeled as either
250 indeterminate *Escherichia species* or as *E. coli*, as cryptic clades are missing from the NCBI
251 taxonomy. This causes Kraken2 to misidentify cryptic clade strains. Adding entries for cryptic
252 clades in the NCBI taxonomy (e.g. as formally recognized species) should solve this issue. It is
253 not unimaginable that the same issue holds for phenotypic characterization methods such as
254 the MALDI-TOF or Vitek2 systems, meaning some of the reference strains in those databases
255 could be wrongly labeled. Another possible explanation for the Vitek2 results is that it is
256 challenging to differentiate cryptic clades from *E. coli* based on biochemical properties. Walk et
257 al. (2009)(S. T. Walk et al., 2009) assessed 31 biochemical markers and attempted to
258 differentiate cryptic clades from *E. coli* based on the biochemical profiles, which was achieved
259 only partially. However, high-throughput metabolomic analyses of cryptic clade strains have not
260 been performed yet to the best of our knowledge, which means there could be differentiating
261 biochemical markers waiting to be discovered.

262

263 Cryptic clade I genomes in our collection harbored more virulence and resistance genes
264 compared to other clades. Notably, almost half of clade I genomes harbored the genes
265 necessary to produce either Shiga toxin 1 or 2. It might be that this is partly due to a sampling
266 bias – toxigenic clade I strains might be whole-genome sequenced more often – but genes
267 encoding Shiga toxin were not nearly as abundant in other cryptic clades. In fact, cryptic clade I
268 might represent an eighth non-O157 *stx*+ *Escherichia* lineage (Alikhan et al., 2019). Our analysis
269 definitely confirms the earlier suspicion of Walk (2015) that a relatively high percentage cryptic
270 clade I strains harbor shiga toxin (S. Walk, 2015). One positive point out of the perspective of

271 public health is that none of the *stx*+ clade I genomes harbored plasmid-mediated genes
272 conferring resistance to extended-spectrum beta-lactams, carbapenems or colistin.

273

274 Our study provides an improved and systematic taxonomy of the *Escherichia/Shigella* genus,
275 making effective use of public sequence databases. Based on this analysis, we describe a novel
276 *Escherichia* species, *E. ruysiae sp. nov*., on a phenotypic and genomic level. The genomic
277 analysis of *Escherichia* population structure and the description of *E. ruysiae* should aid the
278 correct identification of *Escherichia* species in the future.

279 Materials and Methods

280 **Assessing abundance of cryptic clade strains in RefSeq and SRA**

281 10,824 *Escherichia* genome assemblies were downloaded from the RefSeq database on July
282 17th, 2018 using ncbi-download (version 0.2.6, https://github.com/kblin/ncbi-genome-
283 download). The downloaded genomes were processed using ClermonTyper (version
284 1.3.0)(Beghain et al., 2018) to detect putative cryptic clade strains. To check for false negatives
285 in the ClermonTyper analysis, a neighbor-joining tree was constructed including all 10,824
286 *Escherichia* assemblies using mash (version 2.1) (Ondov et al., 2016), the square_mash script
287 from PySEER (version 1.2.0)(Lees, Galardini, Bentley, Weiser, & Corander, 2018) and rapidNJ
288 (version 2.3.2) (Simonsen et al., 2008).

289 79,911 *Escherichia* entries were assembled from the Sequence Read Archive (SRA) using SKESA
290 (version 2.3.0) (Souvorov, Agarwala, & Lipman, 2018), assembling with a single kmer size of 51

291 without using read pairing information and without removing adapter sequences. FastANI
292 (version 1.1)(Jain et al., 2018) was used to find genomes similar to any of the RefSeq cryptic
293 clade genomes among these 79,911 assemblies.

294 Thresholds were set at 95% for clades II to VI. All strains with an ANI above the threshold for a
295 particular clade were selected for further analyses. Similarity was expressed in %ANI above
296 certain thresholds. Since cryptic clade I strains are similar to *E. coli*, a higher threshold for clade
297 I was chosen as using a 95% ANI threshold resulted in >50,000 hits. The lowest ANI value
298 between any two RefSeq cryptic clade I strains was 97.6%, so we chose a threshold of 96.5% to
299 reduce the probability of false negatives as much as feasible.

300 All SRA strains that showed similarity to RefSeq cryptic clade strains were re-assembled using
301 SKESA with default settings. Finally, a mash distance-based phylogeny was produced of all
302 reassembled strains to confirm phylogenetic placement of strains.

303

304 **Genomic characterization of cryptic clade strains**

305 Assembly metrics of all SRA assemblies were analyzed using Quast (version 4.5)(Gurevich et al.,
306 2013). Additionally, all SRA assemblies were screened using ABRicate (version 0.8.10,
307 https://github.com/tseemann/abricate) for virulence genes (using the VFDB core database,
308 downloaded 31st October 2018)(Chen et al., 2005) or antimicrobial resistance (using the CARD
309 database, downloaded 31st October 2018) (McArthur et al., 2013). Kraken2 (version 2.0.7-beta)
310 was used with the 8GB MiniKraken database to perform species identification of cryptic clade

311 genomes. Genome-wide ANI was calculated using fastANI (version 1.1)(Jain et al., 2018) with
312 default settings.

313

314 **Phylogenetic analysis**

315 A total of 489 genomes (344 cryptic clade genomes and 145 *E. albertii, E. fergusonii, E. coli* and
316 *Shigella spp.*) were used to construct a maximum likelihood phylogeny. kSNP3 (version 3.1)
317 (Gardner et al., 2015) was used to extract SNPs from kmers present in all genomes. This core
318 SNP matrix from kmers was used as input for IQ-tree (version 1.6.6) (Nguyen et al., 2015),
319 which calculated a maximum-likelihood phylogeny under the GTR-GAMMA model, correcting
320 for ascertainment bias using 1000 ultrafast bootstraps (Hoang, Chernomor, von Haeseler, Minh,
321 & Vinh, 2017).

322

323 **Average Nucleotide Identity (ANI) analysis of 489 *Escherichia* and *Shigella* strains**

324 ANI analysis was performed using fastANI (version 1.1) which efficiently and accurately calculates
325 ANI for a large genome collection. 489 *Escherichia* and *Shigella* strains were used as query and
326 reference, providing a total of 238,632 comparisons (without self-comparisons). Mean ANI
327 values were calculated between all strains from compared groups. Subsequently, fastANI
328 output was written to a full matrix using the square_mash script of PySEER (Lees et al., 2018).
329 Values lower than 95% were removed in R (version 3.5.2) and a Cytoscape network was created
330 using the graph_from_adjacency_matrix function from the igraph package (version 1.2.4)
331 (Csardi & Nepusz, 2006), the jsonlite package (version 1.6) (Ooms, 2014) and the toCytoscape

332 function from https://github.com/idekerlab/cy-rest-R. The network was plotted in Cytoscape
333 (version 3.7.1) (Shannon et al., 2003) and layout was provided through a weighted Prefuse
334 Force Directed Layout algorithm distributed with Cytoscape.

335

336 **Phenotypic characterization of *Escherichia ruysiae sp. nov.***

337 *Escherichia ruysiae sp. nov.* strain OPT1704^T strain was grown on COS sheep blood agar plates
338 (BD) at 37 °C unless stated otherwise. Anaerobic growth was assessed by growing OPT1704^T in
339 a sealed container with a BD Anaerobic GasPak within. A hanging drop preparation was used to
340 detect motility. The VITEK 2 system using a GN ID card (BioMérieux) and the MALDI-TOF system
341 (Bruker) were used with default settings for phenotypic identification of OPT1704^T. The Vitek
342 system was also used to obtain antimicrobial susceptibility profiles for multiple antibiotic
343 classes. EUCAST v9.0 clinical breakpoints were used to classify MICs into S/I/R categories.

344

345 **Genomic characterization of *Escherichia ruysiae sp. nov.***

346 DNA for Illumina whole-genome sequencing was extracted using the Qiagen Blood and Tissue
347 kit according to the manufacturer's instructions. The sequencing library was prepared using the
348 Kapa HTP Library Preparation kit and subsequently sequenced on an Illumina HiSeq 4000
349 platform. Adapter sequences were removed using fastp (Zhou, Chen, Chen, & Gu, 2018).

350 DNA for Oxford Nanopore Technologies whole-genome sequencing was extracted using the
351 Qiagen MagAttract HMW DNA kit according to the manufacturer's instructions. The sequencing

352 library was prepared using the Nanopore rapid barcoding kit (SQK-RBK004) and sequenced in a
353 multiplexed run. Basecalling was performed using MinKNOW in fast basecalling mode and qcat
354 (version 1.0.7, https://github.com/nanoporetech/qcat) was used to demultiplex reads. Reads
355 were filtered using Filtlong (version 0.2.0, https://github.com/rrwick/Filtlong) where 90% of
356 best reads were retained, judged on read length and read identity based on Illumina
357 sequencing data. Filtered Nanopore reads and subsampled Illumina reads were used to perform
358 hybrid assembly using Unicycler (version 0.4.6, Wick 2017).

359 ABRicate was used as described above to predict resistance and virulence genes. 16S rRNA
360 gene sequence was extracted using barrnap (version 0.9,
361 https://github.com/tseemann/barrnap). Digital DNA:DNA hybridization scores were calculated
362 using the DSMZ webtool (Meier-Kolthoff, Auch, Klenk, & Göker, 2013) and ANI was calculated
363 using OrthoANIu (Yoon et al., 2017), Enveomics ANI calculator (Rodriguez-R & Konstantinidis,
364 2016) and fastANI (Jain et al., 2018).

365

366 Acknowledgments

367 The COMBAT study was funded by Netherlands Organization for Health, Research and
368 Development (ZonMw; 50-51700-98-120) and EU-H2020 programme (COMPARE, 643476). The
369 authors would like to thank Rob Weijts and Patricia Brinke for their help in phenotypic
370 characterization of type strain OPT1704^{T} of *Escherichia ruysiae sp. nov*.

371

372 # Data availability

373 All supplementary data is available via FigShare. This includes a Readme
374 (https://doi.org/10.6084/m9.figshare.9824633.v1), a list of accession numbers for the 10,824
375 *Escherichia* genomes downloaded from NCBI (https://doi.org/10.6084/m9.figshare.9824234), a
376 list of accession numbers for the 79,911 *Escherichia* entries assembled from the SRA using
377 SKESA (https://doi.org/10.6084/m9.figshare.9824261), the re-assemblies of the 344 selected
378 cryptic clade strains (https://doi.org/10.6084/m9.figshare.9824270) and a summary of the
379 bioinformatic commands used (https://doi.org/10.6084/m9.figshare.9777746). Illumina and
380 ONT fastq files are available on ENA under accession numbers ERR3518913 and ERR3518914,
381 respectively. Illumina and ONT sequencing data, as well as Unicycler assembly of strain
382 OPT1704^{T} are available on ENA under the project number PRJEB34275. Pure cultures of strain
383 OPT1704^{T} are available through the Netherlands Culture Collection of Bacteria of the
384 Westerdijk Institute under number NCCB100732
385 (http://www.westerdijkinstitute.nl/Collections/DefaultInfo.aspx?Page=Bacteria), the National
386 Collection of Type Cultures of Public Health England under number NCTC14359
387 (https://www.phe-culturecollections.org.uk/collections/nctc.aspx) or through the
388 corresponding author.

389

390 # Appendixes

391 A full list of members of the COMBAT consortium is supplied in the file
392 "Members_COMBAT_consortium.docx".

References

Alikhan, N.-F., Bachmann, N. L., Zakour, N. L. Ben, Petty, N. K., Stanton-Cook, M., Gawthorne, J. A., … Beatson, S. A. (2019). Multiple evolutionary trajectories for non-O157 Shiga toxigenic Escherichia coli. *BioRxiv*. https://doi.org/10.1101/549998

Arcilla, M. S., van Hattem, J. M., Haverkate, M. R., Bootsma, M. C. J., van Genderen, P. J. J., Goorhuis, A., … Penders, J. (2017). Import and spread of extended-spectrum beta-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): A prospective, multicentre cohort study. *The Lancet Infectious Diseases*, *17*(1), 78–85. https://doi.org/10.1016/S1473-3099(16)30319-X

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., & Clermont, O. (2018). ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microbial Genomics*, 1–8. https://doi.org/10.1099/mgen.0.000192

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Research*, *33*(DATABASE ISS.), 325–328. https://doi.org/10.1093/nar/gki008

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., … Trujillo, M. E. (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, *68*(1), 461–466. https://doi.org/10.1099/ijsem.0.002516

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Sy*, 1695.

Farmer, J. J., Fanning, G. R., Davis, B. R., O\textquoterightHara, C. M., Riddle, C., Hickman-Brenner, F. W., … Brenner, D. J. (1985). Escherichia fergusonii and Enterobacter taylorae, two new species of Enterobacteriaceae isolated from clinical specimens. *Journal of Clinical Microbiology*, *21*(1), 77–81.

Gangiredla, J., Mammel, M. K., Barnaba, T. J., Tartera, C., Gebru, S. T., Patel, I. R., … Lacher, D. W. (2018). Draft Genome Sequences of Escherichia albertii, Escherichia fergusonii, and Strains Belonging to Six Cryptic Lineages of Escherichia spp. *Microbiology Resource Announcements*, *6*(18). https://doi.org/10.1128/genomeA.00271-18

Gardner, S. N., Slezak, T., & Hall, B. G. (2015). kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*, *31*(17), 2877–2878. https://doi.org/10.1093/bioinformatics/btv271

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology, 57*(1), 81–91.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2017). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Huys, G., Cnockaert, M., Janda, J. M., & Swings, J. (2003). Escherichia albertii sp. Nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *International Journal of Systematic and Evolutionary Microbiology*, *53*(3), 807–810.

437 Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI
438 analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*,
439 *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9
440 Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., & Corander, J. (2018). pyseer: A comprehensive
441 tool for microbial pangenome-wide association studies. *Bioinformatics*, *34*(24), 4310–4312.
442 https://doi.org/10.1093/bioinformatics/bty539
443 Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments.
444 *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239
445 Liu, S., Jin, D., Lan, R., Wang, Y., Meng, Q., Dai, H., … Xu, J. (2015). Escherichia marmotae sp. Nov.,
446 isolated from faeces of Marmota himalayana. *International Journal of Systematic and*
447 *Evolutionary Microbiology*, *65*(7), 2130–2134.
448 McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., … Wright, G. D. (2013). The
449 comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, *57*(7),
450 3348–3357. https://doi.org/10.1128/AAC.00419-13
451 Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species
452 delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*,
453 *14*(1), 60. https://doi.org/10.1186/1471-2105-14-60
454 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective
455 stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and*
456 *Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300
457 Ochman, H., & Selander, R. K. (1984). Standard reference strains of Escherichia coli from natural
458 populations. *Journal of Bacteriology*, *157*(2), 690–693.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0997-x

Ooka, T., Seto, K., Kawano, K., Kobayashi, H., Etoh, Y., Ichihara, S., … Hayashi, T. (2012). Clinical Significance of Escherichia albertii. *Emerging Infectious Disease Journal*, *18*(3), 488. https://doi.org/10.3201/eid1803.111401

Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *ArXiv E-Prints*, arXiv:1403.2805.

Pettengill, E. A., Pettengill, J. B., & Binet, R. (2016). Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Frontiers in Microbiology*, *6*, 1573. https://doi.org/10.3389/fmicb.2015.01573

Rodriguez-R, L. M., & Konstantinidis, K. T. (2016). The enveomics collection: A toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*, *4*, e1900v1. https://doi.org/10.7287/peerj.preprints.1900v1

Russo, T. A., & Johnson, J. R. (2003). Medical and economic impact of extraintestinal infections due to Escherichia coli: Focus on an increasingly important endemic problem. *Microbes and Infection*, *5*(5), 449–456. https://doi.org/10.1016/S1286-4579(03)00049-2

Savini, V., Catavitello, C., Talia, M., Manna, A., Pompetti, F., Favaro, M., … D\textquoterightAntonio, D. (2008). Multidrug-Resistant Escherichia fergusonii: A Case of Acute Cystitis. *Journal of Clinical Microbiology*, *46*(4), 1551–1552. https://doi.org/10.1128/JCM.01210-07

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Simonsen, M., Mailund, T., & Pedersen, C. N. S. (2008). Rapid Neighbour-Joining. In K. A. Crandall & J. Lagergren (Eds.), *Algorithms in Bioinformatics* (pp. 113–122). Berlin, Heidelberg: Springer Berlin Heidelberg.

Souvorov, A., Agarwala, R., & Lipman, D. J. (2018). SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biology*, *19*(1), 153. https://doi.org/10.1186/s13059-018-1540-z

Walk, S. (2015). The "Cryptic" Escherichia. *EcoSal Plus*. Retrieved from https://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.ESP-0002-2015

Walk, S. T., Alm, E. W., Gordon, D. M., Ram, J. L., Toranzos, G. A., Tiedje, J. M., & Whittam, T. S. (2009). Cryptic lineages of the genus Escherichia. *Applied and Environmental Microbiology*, *75*(20), 6534–6544. https://doi.org/10.1128/AEM.01262-09

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Yoon, S.-H., Ha, S.-M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek*, *110*(10), 1281–1286. https://doi.org/10.1007/s10482-017-0844-4

Zhou, Y., Chen, Y., Chen, S., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560