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

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

Introduction

Within the *Escherichia* genus, four species are recognized; *E. coli*, *E. fergusonii*[1], *E. albertii*[2] and most recently, *E. marmotae*[3]. *E. coli*, *E. fergusonii* and *E. albertii* have been associated with animal and human disease[4–6], while little is known about the clinical relevance and characteristics of the most recently described *Escherichia* species *E. marmotae*. Some *Escherichia* strains cannot be assigned to any of the four existing species. Based on genetic data, these strains cluster into six groups, which were termed '*Escherichia* cryptic clades'[7,8]. Recently, cryptic clade V was formally recognized as a separate species (*E. marmotae*)[3], leaving five cryptic clades that have not been recognized as species. As these have not yet received entries in taxonomic databases, isolates representing cryptic clades might be misidentified by tools that identify bacterial species based on whole-genome sequence (WGS). Also, the abundancy of these cryptic clade strains in public databases such as RefSeq (curated repository of reference sequences) or the Sequence Read Archive (SRA, repository of raw sequence data) is unknown.

Therefore, we have addressed two questions in this study:

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

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

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

Results

Assessing abundance of cryptic clade strains in RefSeq

We downloaded all 10,824 available *Escherichia* genomes from the RefSeq database and identified 92 cryptic clade genomes using ClermonTyper[9]. Since ClermonTyper is only able to detect strains belonging to known cryptic clades, we constructed a neighbor-joining phylogeny of the 10,824 RefSeq *Escherichia* strains using mash[10] and rapidNJ[11] to detect cryptic clade strains missed by ClermonTyper. This analysis revealed two additional cryptic clade strains, bringing the total of cryptic clade strains in RefSeq to 94, or 0.9% of all *Escherichia* genomes in RefSeq.

We then screened 79,911 Sequence Read Archive (SRA) *Escherichia* entries for presence of cryptic clade genomes. In short, we performed a rough assembly of the SRA data using SKESA and used fastANI to detect cryptic clade strains, using RefSeq cryptic clade genomes as references. 757 putative cryptic clade genomes were included in a neighborhood-joining phylogeny and were checked using ClermonTyper. This led to the discovery of 357 cryptic clade genomes in the SRA, or 0.45% of SRA *Escherichia* assemblies. We evaluated SKESA assemblies using Quast[12], and excluded 13 assemblies due to inadequate assembly qualities or aberrant assembly length. Since there are likely duplicates between SRA and RefSeq cryptic clade

genomes, we chose to continue with the 344 cryptic clade genomes from the SRA. An overview of the process is provided in figure 1.

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[13] and 13 *Shigella spp.* genomes, all from RefSeq. We constructed a maximum-likelihood phylogeny using IQ-Tree[14] from a SNP alignment produced by kSNP3[15] (fig. 2).

Based on species identification using Kraken2[16] all cryptic clade I, II, III, IV and VI strains were classified as *E. coli*. 82 out of 92 *E. marmotae* (formerly cryptic clade V) strains were identified correctly while the remaining *E. marmotae* strains were identified as *E. coli*.

Resistance and virulence genes in cryptic clade genomes

Genes encoding both subunits of shiga toxin were detected in 108/209 (52%) of cryptic clade I strains. In 26 cryptic clade I strains, *stx1a* and *stx1b* were detected together, in 75 strains *stx2a* and *stx2b* were detected together. In 7 clade I strains only one shiga toxin-encoding gene was detected. Extended-spectrum beta-lactamase (ESBL) genes were detected in 8.1% and 2.2% of

104 cryptic clade I and cryptic clade V/*E. marmotae* genomes, respectively. One clade I strain and
105 two *E. marmotae*/clade V strains harbored carbapenemase-encoding genes.

106

107 **Assignment of cryptic clades to existing or new species**

108 Based on a 95% average nucleotide identity (ANI)[17,18] 489 of the included *Escherichia* and
109 *Shigella* strains should be assigned to seven discrete species with no relationships between
110 these groups (table 1 and fig. 3). *E. coli* and cryptic clade I showed mean ANI values of 96.0%,
111 indicating *E. coli* and cryptic clade I should be assigned into a single species, possibly separated
112 into subspecies. Genomically, *Shigella spp*. belong to *E. coli* as a mean ANI is 97.6% with *E. coli*.
113 This is in line with earlier reports[19].

114

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

119

120 **Table 1.** Mean ANI values between groups calculated using fastANI, for 489 included
121 *Escherichia* and *Shigella* strains. Values higher than the threshold for species delineation (95%)
122 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.**
Clade II		**99.0**	92.1	92.0	91.3	90.9	89.7	91.8	89.2	91.
Clade III			**98.8**	**96.6**	92.3	91.6	89.8	92.6	89.4	92.
Clade IV				**99.0**	92.2	91.7	89.9	92.8	89.5	92.
Clade V/*E. marmotae*					**99.3**	90.4	89.4	91.2	88.4	91.
Clade VI						**99.9**	89.6	94.4	90.1	94.
E. albertii							**98.6**	90.1	88.2	90.
E. coli								**97.7**	91.1	**97.**
E. fergusonii									**98.7**	91.
Shigella spp.										**98.**

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

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

Whole-genome sequence analysis

The whole-genome DNA sequence of strain OPT1704^T was determined using Illumina HiSeq and Oxford Nanopore Technologies (ONT) sequencing platforms. The Illumina sequencing run yielded a total of 6.3×10^6 paired-end reads, with a mean read length of 151 bp. Illumina reads were downsampled using seqtk (version 1.3-r106, https://github.com/lh3/seqtk) to provide a theoretical depth of coverage of 100X with the assumption that the OPT1704^T has a genome size of approximately 5×10^6 bp. The ONT sequencing run yielded a total of 2.5×10^4 reads, with a mean read length of 9078 bp before filtering. ONT reads were filtered on length and on read identity using Illumina reads, leaving 1.5×10^4 reads with a mean length of 12580 bp. This provided a theoretical depth of coverage of ~38X of ONT reads. Assembly using both Illumina

and Nanopore reads resulted in a complete genome, consisting of one circular chromosome and one circular plasmid. GC content of the OPT1704^{T} genome was 50.6%.

Putative resistance and virulence genes were predicted from the draft genome using ABRicate with the CARD[21] and VFDB[22] databases. OPT1704^{T} harbours 6 resistance genes that are typically plasmid-mediated in *E. coli*, associated with reduced susceptibility to fluoroquinolones (*qnrS1*), aminoglycosides (*aph(6)-Id* & *aph(3'')-Ib*), cephalosporins (*blaCTX-M-14*), trimethoprim (*dfrA14*) and sulphonamides (*sul2*). This is in line with the reduced susceptibility to fluoroquinolones (norfloxacin, MIC: 2 mg/L and ciprofloxacin, MIC: 0.5 mg/L by Vitek2), cephalosporins (cefuroxime, MIC: >32 mg/L and cefotaxime, MIC: 4 mg/L) and trimethoprim-sulfamethoxazole (MIC: >8 mg/L). Furthermore, several putative virulence genes were predicted from the draft genome sequence associated with siderophore function (*chuX, entS, fepABD*), fimbriae (*fimBCDGI*), a type II secretion system (*gspGHI*) and capsular polysaccharide biogenesis (*kpsD*). The diarrhoeal symptoms of the traveller contributing this strain would not be expected based on these predicted virulence genes.

Next, we calculated 16S rRNA sequence similarities, ANI values and digital DNA:DNA hybridisation (dDDH) values between OPT1704^{T} and type strains of the four other *Escherichia* species. This time, we used three separate tools to calculate ANI (fastANI, OrthoANIu and ANI calculator from Enveomics)[23–25], since scalability was no concern. 16S rRNA sequence similarities did not completely warrant assignment of OPT1704^{T} to a novel species, but ANI

177 analysis and dDDH did support assignment of OPT1704^T to a novel species (table 2). This novel
178 species was assigned *E. ruysiae sp. nov.* with OPT1704^T as the proposed type strain. As our
179 earlier analysis shows cryptic clade III clusters with cryptic clade IV on the species level, we
180 propose that *E. ruysiae sp. nov.* encompasses both cryptic clade III and IV.

181

182 **Table 2.** Comparison of OPT1704^T whole-genome sequence with type strains of *E. albertii*, *E.*
183 *coli*, *E. fergusonii* and *E. marmotae*. In bold are the values that suggest assignment of OPT1704^T
184 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**

185

Discussion

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

Contrary to the present study and possibly because a smaller dataset was used, Luo et al. (2011) observed a 'genetic continuum' between *E. coli* and cryptic clades. In line with findings by Walk (2015) we did not find strains that share an ANI > 95% between the seven groups, indicating discrete grouping[26]. Although they should be assigned to a single species based on ANI analysis, a clear separation between *E. coli/Shigella spp.* and clade I is observed, possibly at the subspecies level. The same holds for cryptic clades III and IV. For the other cryptic clades, no significant separation in population structure is visible, indicating these clades are genetically more homogeneous. The species clusters we find are also found by the Genome Taxonomy Database (GTDB), where cryptic clade II is *Escherichia sp. 001660175*, cryptic clade VI is *Escherichia sp. 002965065* and *E. ruysiae sp. nov.* is *Escherichia sp. 000208585*[27].

Throughout our comprehensive scanning of RefSeq and SRA, we attempted to maximize the diversity of *Escherichia* genomes captured. We built the phylogeny of >10,000 RefSeq genomes which included all *Escherichia* cryptic clades described in the literature. As we based our SRA search on the RefSeq cryptic clade genomes, we cannot exclude that we missed novel cryptic clades present in the SRA but not in RefSeq. Additionally, we had to raise the threshold with

207 which we screened the SRA for cryptic clade I strains. Possibly, strains that are genetically
208 different could have been missed by choosing this strategy.

209

210 Based on genomic analyses of the *Escherichia* genus and characterization of strain OPT1704^T,
211 we proposed a novel species, *Escherichia ruysiae sp. nov*. Analysis of 16S rRNA genes of *E.*
212 *ruysiae* and type strains of other *Escherichia spp.* did not warrant assignment to a novel species,
213 while ANI and dDDH analysis of the same genomes did. The International Journal of Systematic
214 and Evolutionary Microbiology follows results of ANI or dDDH analysis instead of 16S rRNA
215 analysis when discrepancies are observed, meaning strain OPT1704^T should be assigned to a
216 novel species[17]. Assignment of cryptic clades III and IV as a novel species *E. ruysiae* means
217 inclusion of the species in the NCBI taxonomy database, as happened with *E. marmotae*/cryptic
218 clade V before. As the NCBI taxonomy serves as the reference for many genomic classification
219 tools, assigning cryptic clades III and IV to a novel species is the most effective way to stimulate
220 correct classification of these strains.

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

226

All standard identification methods we used in this study – genomic and phenotypic – concluded strain OPT1704^T is an *E. coli* strain. Genomic identification methods such as Kraken2 often use the RefSeq database in combination with the NCBI taxonomy to classify strains. Cryptic clade strains are present in the RefSeq database, but are always labeled as either indeterminate *Escherichia species* or as *E. coli*, as cryptic clades are missing from the NCBI taxonomy. This causes Kraken2 to misidentify cryptic clade strains. Adding entries for cryptic clades in the NCBI taxonomy (e.g. as formally recognized species) should solve this issue. It is not unimaginable that the same issue holds for phenotypic characterization methods such as the MALDI-TOF or Vitek2 systems, meaning some of the reference strains in those databases could be wrongly labeled. Another possible explanation for the Vitek2 results is that it is challenging to differentiate cryptic clades from *E. coli* based on biochemical properties. Walk et al. (2009)[8] assessed 31 biochemical markers and attempted to differentiate cryptic clades from *E. coli* based on the biochemical profiles, which was achieved only partially. However, high-throughput metabolomic analyses of cryptic clade strains have not been performed yet to the best of our knowledge, which means there could be differentiating biochemical markers waiting to be discovered.

Cryptic clade I genomes in our collection harbored more virulence and resistance genes compared to other clades. Notably, almost half of clade I genomes harbored the genes necessary to produce either Shiga toxin 1 or 2. It might be that this is partly due to a sampling bias – toxigenic clade I strains might be whole-genome sequenced more often – but genes encoding Shiga toxin were not nearly as abundant in other cryptic clades. In fact, cryptic clade I

might represent an eighth non-O157 *stx*+ *Escherichia* lineage[28]. Our analysis definitely confirms the earlier suspicion of Walk (2015) that a relatively high percentage cryptic clade I strains harbor shiga toxin[26]. One positive point out of the perspective of public health is that none of the *stx*+ clade I genomes harbored plasmid-mediated genes conferring resistance to extended-spectrum beta-lactams, carbapenems or colistin.

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

Materials and Methods

Assessing abundance of cryptic clade strains in RefSeq and SRA

10,824 *Escherichia* genome assemblies were downloaded from the RefSeq database on July 17th, 2018 using ncbi-download (version 0.2.6, https://github.com/kblin/ncbi-genome-download). The downloaded genomes were processed using ClermonTyper (version 1.3.0)[9] to detect putative cryptic clade strains. To check for false negatives in the ClermonTyper analysis, a neighbor-joining tree was constructed including all 10,824 *Escherichia* assemblies using mash (version 2.1)[10], the square_mash script from PySEER (version 1.2.0)[29] and rapidNJ (version 2.3.2)[11].

79,911 *Escherichia* entries were assembled from the Sequence Read Archive (SRA) using SKESA (version 2.3.0)[30], assembling with a single kmer size of 51 without using read pairing information and without removing adapter sequences. FastANI (version 1.1)[23] was used to find genomes similar to any of the RefSeq cryptic clade genomes among these 79,911 assemblies.

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

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

Genomic characterization of cryptic clade strains

Assembly metrics of all SRA assemblies were analyzed using Quast (version 4.5)[12]. Additionally, all SRA assemblies were screened using ABRicate (version 0.8.10, https://github.com/tseemann/abricate) for virulence genes (using the VFDB core database, downloaded 31st October 2018)[22] or antimicrobial resistance (using the CARD database, downloaded 31st October 2018)[21]. Kraken2 (version 2.0.7-beta) was used with the 8GB

MiniKraken database to perform species identification of cryptic clade genomes. Genome-wide ANI was calculated using fastANI (version 1.1)[23] with default settings.

Phylogenetic analysis

A total of 489 genomes (344 cryptic clade genomes and 145 *E. albertii, E. fergusonii, E. coli* and *Shigella spp.*) were used to construct a maximum likelihood phylogeny. kSNP3 (version 3.1)[15] was used to extract SNPs from kmers present in all genomes. This core SNP matrix from kmers was used as input for IQ-tree (version 1.6.6)[14], which calculated a maximum-likelihood phylogeny under the GTR-GAMMA model, correcting for ascertainment bias using 1000 ultrafast bootstraps[31].

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

ANI analysis was performed using fastANI (version 1.1) which efficiently and accurately calculates ANI for a large genome collection. 489 *Escherichia* and *Shigella* strains were used as query and reference, providing a total of 238,632 comparisons (without self-comparisons). Mean ANI values were calculated between all strains from compared groups. Subsequently, fastANI output was written to a full matrix using the square_mash script of PySEER[29]. Values lower than 95% were removed in R (version 3.5.2) and a Cytoscape network was created using the graph_from_adjacency_matrix function from the igraph package (version 1.2.4)[32], the jsonlite package (version 1.6)[33] and the toCytoscape function from https://github.com/idekerlab/cy-

rest-R. The network was plotted in Cytoscape (version 3.7.1)[34] and layout was provided through a weighted Prefuse Force Directed Layout algorithm distributed with Cytoscape.

Phenotypic characterization of *Escherichia ruysiae sp. nov.*

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

Genomic characterization of *Escherichia ruysiae sp. nov.*

DNA for Illumina whole-genome sequencing was extracted using the Qiagen Blood and Tissue kit according to the manufacturer's instructions. The sequencing library was prepared using the Kapa HTP Library Preparation kit and subsequently sequenced on an Illumina HiSeq 4000 platform. Adapter sequences were removed using fastp[35].

DNA for Oxford Nanopore Technologies whole-genome sequencing was extracted using the Qiagen MagAttract HMW DNA kit according to the manufacturer's instructions. The sequencing library was prepared using the Nanopore rapid barcoding kit (SQK-RBK004) and sequenced in a

multiplexed run. Basecalling was performed using MinKNOW in fast basecalling mode and qcat (version 1.0.7, https://github.com/nanoporetech/qcat) was used to demultiplex reads. Reads were filtered using Filtlong (version 0.2.0, https://github.com/rrwick/Filtlong) where 90% of best reads were retained, judged on read length and read identity based on Illumina sequencing data. Filtered Nanopore reads and subsampled Illumina reads were used to perform hybrid assembly using Unicycler (version 0.4.6, Wick 2017).

ABRicate was used as described above to predict resistance and virulence genes. 16S rRNA gene sequence was extracted using barrnap (version 0.9, https://github.com/tseemann/barrnap). Digital DNA:DNA hybridization scores were calculated using the DSMZ webtool[36] and ANI was calculated using OrthoANIu[25], Enveomics ANI calculator[24] and fastANI[23].

Acknowledgments

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

Data availability

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

Appendixes

The COMBAT consortium (in alphabetical order)

Maris S. Arcilla[1], Martin C.J. Bootsma[2], Perry J. van Genderen[3], Abraham Goorhuis[4], Martin Grobusch[4], Jarne M. van Hattem[5], Menno D. de Jong[5], Damian C. Melles[1], Nicky Molhoek[6], Astrid M.L. Oude Lashof[7], John Penders[7], Constance Schultsz[5, 8], Ellen E. Stobberingh[5], Henri A. Verbrugh[1]

[1]Erasmus University Medical Centre, Department of Medical Microbiology and Infectious Diseases, Rotterdam,

[2]Utrecht University, Department, Faculty of Science, Utrecht,

[3]Havenziekenhuis - Institute for Tropical Diseases, Department of Internal Medicine, Rotterdam,

[4]Academic Medical Centre, Center of Tropical Medicine and Travel Medicine, Amsterdam,

[5]Academic Medical Centre, Department of Medical Microbiology, Amsterdam,

[6]Havenziekenhuis - Institute for Tropical Diseases, Travel Clinic, Rotterdam,

[7]Maastricht University Medical Centre, School for Nutrition, Toxicology and Metabolism and School for Public Health and Primary Care, Department of Medical Microbiology, Department of Medical Microbiology, Maastricht,

[8]Academic Medical Centre, Department of Global Health-Amsterdam Institute for Global Health and the Development, Amsterdam,

All in the Netherlands

References

1. Farmer JJ, Fanning GR, Davis BR, et al. Escherichia fergusonii and Enterobacter taylorae, two new species of Enterobacteriaceae isolated from clinical specimens. *J Clin Microbiol*. 1985;21(1):77-81.

2. Huys G, Cnockaert M, Janda JM, Swings J. Escherichia albertii sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *Int J Syst Evol Microbiol*. 2003;53(3):807-810.

3. Liu S, Jin D, Lan R, et al. Escherichia marmotae sp. nov., isolated from faeces of Marmota himalayana. *Int J Syst Evol Microbiol*. 2015;65(7):2130-2134.

4. Russo TA, Johnson JR. Medical and economic impact of extraintestinal infections due to Escherichia coli: Focus on an increasingly important endemic problem. *Microbes Infect*. 2003;5(5):449-456. doi:10.1016/S1286-4579(03)00049-2

5. Savini V, Catavitello C, Talia M, et al. Multidrug-Resistant Escherichia fergusonii: a Case of Acute Cystitis. *J Clin Microbiol*. 2008;46(4):1551-1552. doi:10.1128/JCM.01210-07

6. Ooka T, Seto K, Kawano K, et al. Clinical Significance of Escherichia albertii. *Emerg Infect Dis J*. 2012;18(3):488. doi:10.3201/eid1803.111401

7. Gangiredla J, Mammel MK, Barnaba TJ, et al. Draft Genome Sequences of Escherichia albertii, Escherichia fergusonii, and Strains Belonging to Six Cryptic Lineages of Escherichia spp. *Microbiol Resour Announc*. 2018;6(18). doi:10.1128/genomeA.00271-18

8. Walk ST, Alm EW, Gordon DM, et al. Cryptic lineages of the genus Escherichia. *Appl Environ Microbiol*. 2009;75(20):6534-6544. doi:10.1128/AEM.01262-09

9. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genomics*. 2018:1-8. doi:10.1099/mgen.0.000192

10. Ondov BD, Treangen TJ, Melsted P, et al. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17(1):1-14. doi:10.1186/s13059-016-0997-x

11. Simonsen M, Mailund T, Pedersen CNS. Rapid Neighbour-Joining. In: Crandall KA, Lagergren J, eds. *Algorithms in Bioinformatics*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008:113-122.

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

13. Ochman H, Selander RK. Standard reference strains of Escherichia coli from natural populations. *J Bacteriol*. 1984;157(2):690-693.

14. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268-274. doi:10.1093/molbev/msu300

15. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*. 2015;31(17):2877-2878. doi:10.1093/bioinformatics/btv271

16. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15:R46. doi:10.1186/gb-2014-15-3-r46

17. Chun J, Oren A, Ventosa A, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol*. 2018;68(1):461-466. doi:10.1099/ijsem.0.002516

18. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*. 2007;57(1):81-91.

19. Pettengill EA, Pettengill JB, Binet R. 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. *Front Microbiol*. 2016;6:1573. doi:10.3389/fmicb.2015.01573

20. Arcilla MS, van Hattem JM, Haverkate MR, et al. Import and spread of extended-spectrum beta-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. *Lancet Infect Dis*. 2017;17(1):78-85. doi:10.1016/S1473-3099(16)30319-X

21. McArthur AG, Waglechner N, Nizam F, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother*. 2013;57(7):3348-3357. doi:10.1128/AAC.00419-13

22. Chen L, Yang J, Yu J, et al. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res*. 2005;33(DATABASE ISS.):325-328. doi:10.1093/nar/gki008

23. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114. doi:10.1038/s41467-018-07641-9

24. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr*. 2016;4:e1900v1. doi:10.7287/peerj.preprints.1900v1

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

26. Walk S. The “Cryptic” Escherichia. *EcoSal Plus*. 2015. https://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.ESP-0002-2015.

27. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy. *bioRxiv*. 2019. doi:10.1101/771964

28. Alikhan N-F, Bachmann NL, Zakour NL Ben, et al. Multiple evolutionary trajectories for non-O157 Shiga toxigenic Escherichia coli. *bioRxiv*. 2019. doi:10.1101/549998

29. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*. 2018;34(24):4310-4312. doi:10.1093/bioinformatics/bty539

30. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol*. 2018;19(1):153. doi:10.1186/s13059-018-1540-z

31. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol*. 2017;35(2):518-522. doi:10.1093/molbev/msx281

32. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal*. 2006;Complex Sy:1695.

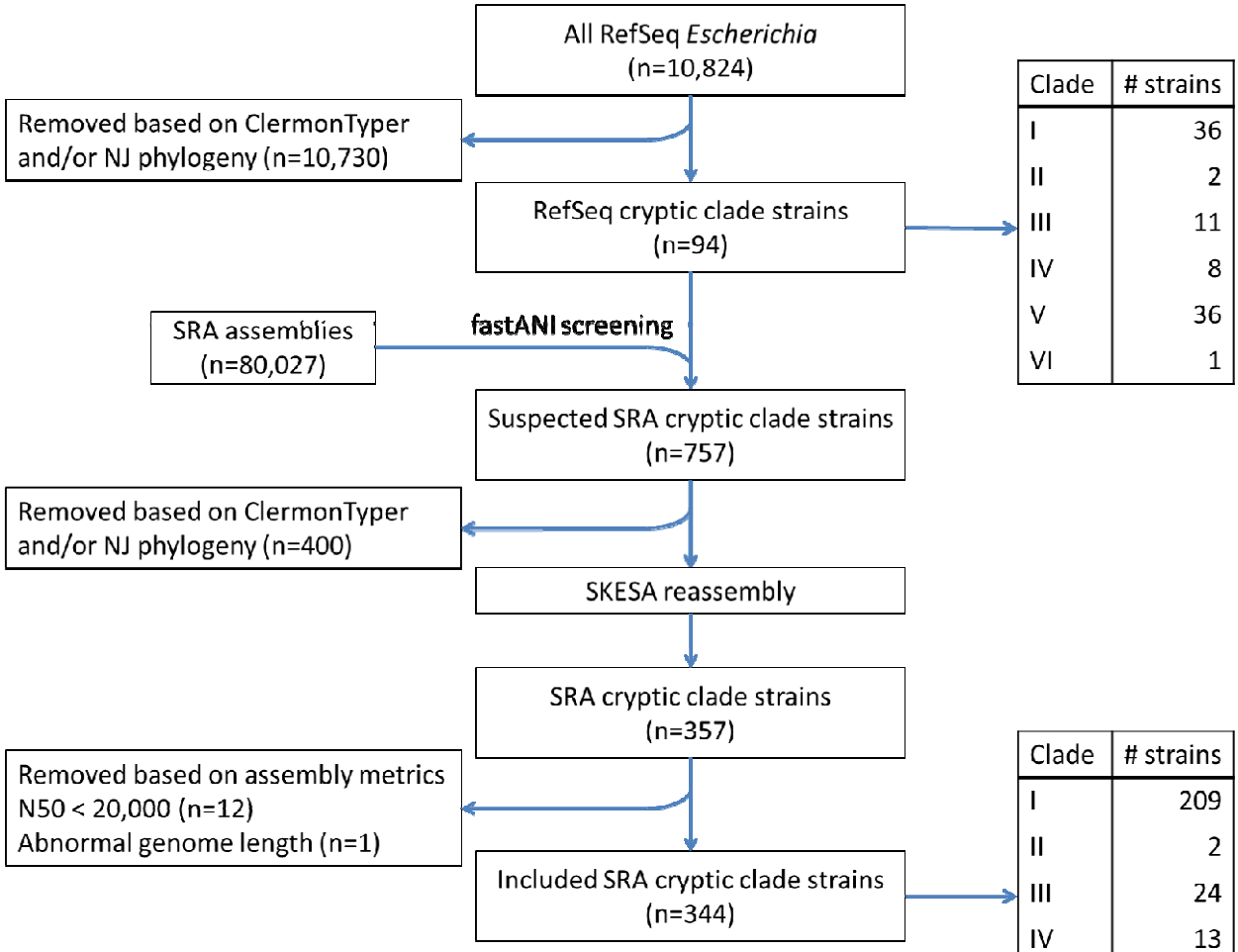
33. Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. *ArXiv E-Prints*. 2014:arXiv:1403.2805.

34. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

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

36. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*. 2013;14(1):60. doi:10.1186/1471-2105-14-60

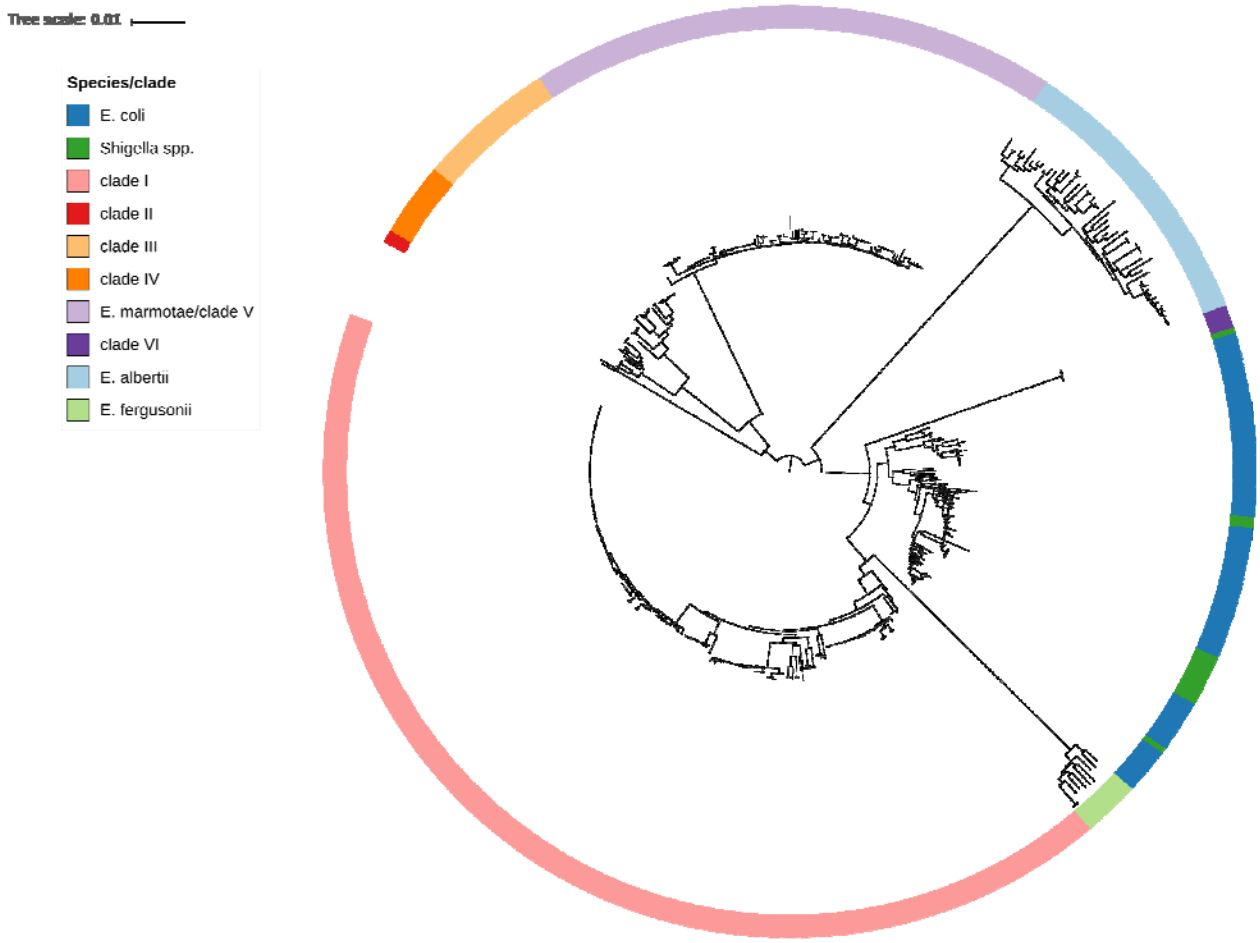
37. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47(W1):W256-W259. doi:10.1093/nar/gkz239



Clade	# strains
I	36
II	2
III	11
IV	8
V	36
VI	1

Clade	# strains
I	209
II	2
III	24
IV	13
V	92
VI	4

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.



485

486 **Figure 2.** Midpoint-rooted maximum-likelihood phylogeny of 489 *Escherichia* and *Shigella*

487 strains based on a core SNP alignment of 3445 positions. Nodes with bootstrap values under 95

488 were collapsed. Visualized in iTOL [37].

489

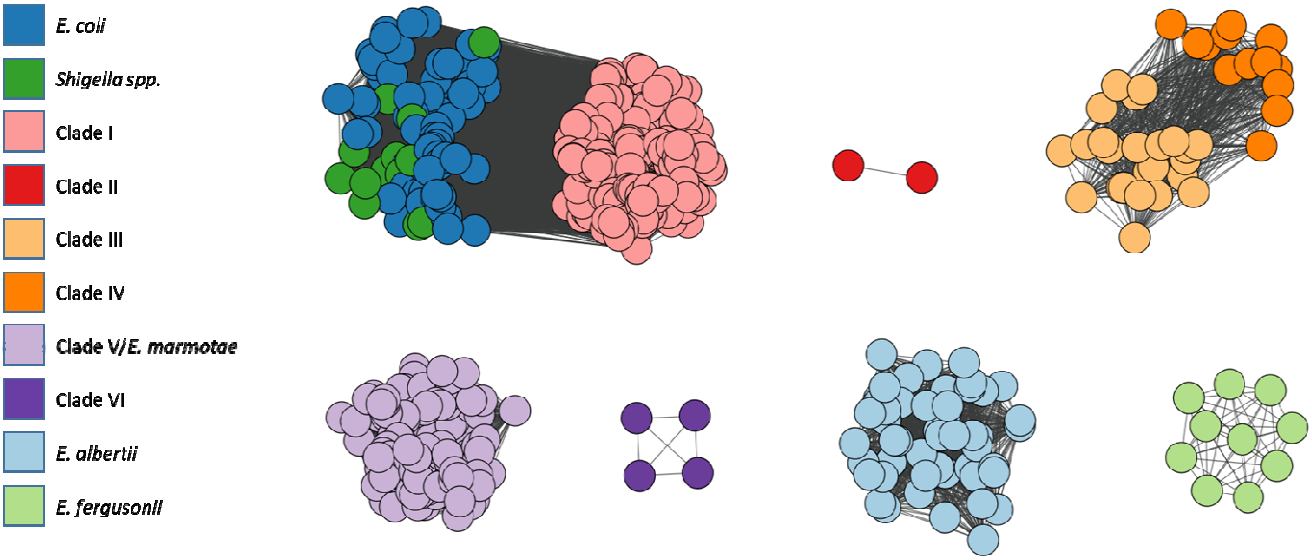


Fig 3. Network of 489 included *Escherichia* and *Shigella* strains, based on fastANI analysis. Edges represent ANI values over 95% (boundary for species delineation) between two strains.