

1 **Genomic diversity of *Escherichia coli* isolates from non-human primates in the Gambia**

2 Ebenezer Foster-Nyarko^{1,2}, Nabil-Fareed Alikhan¹, Anuradha Ravi¹, Gaëtan Thilliez¹,

3 Nicholas Thomson¹, David Baker¹, Gemma Kay¹, Jennifer D. Cramer³, Justin O’Grady¹,

4 Martin Antonio^{2,4}, Mark J. Pallen^{1,5†}

5

6 ¹ Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, United Kingdom

7 ² Medical Research Council Unit the Gambia at the London School of Hygiene and Tropical
8 Medicine, Atlantic Boulevard Road, Fajara, The Gambia

9 ³ American Public University System, Charles Town, WV, USA

10 ⁴ Microbiology and Infection Unit, Warwick Medical School, University of Warwick,
11 Coventry, United Kingdom

12 ⁵ School of Veterinary Medicine, University of Surrey, Guildford, Surrey, United Kingdom.

13

14 [†]Correspondence: Professor Mark Pallen, Quadram Institute Bioscience, Norwich Research
15 Park, Norwich, Norfolk, United Kingdom

16 Email: Mark.Pallen@quadram.ac.uk

17

18 **Abstract**

19 Increasing contact between humans and non-human primates provides an opportunity for the
20 transfer of potential pathogens or antimicrobial resistance between host species. We have
21 investigated genomic diversity, and antimicrobial resistance in *Escherichia coli* isolates from
22 four species of non-human primate in the Gambia: *Papio papio* (n=22), *Chlorocebus sabaeus*
23 (n=14), *Piliocolobus badius* (n=6) and *Erythrocebus patas* (n=1). We performed Illumina
24 whole-genome sequencing on 101 isolates from 43 stools, followed by nanopore long-read
25 sequencing on eleven isolates. We identified 43 sequence types (STs) by the Achtman
26 scheme (ten of which are novel), spanning five of the eight known phylogroups of *E. coli*.
27 The majority of simian isolates belong to phylogroup B2—characterised by strains that cause
28 human extraintestinal infections—and encode factors associated with extraintestinal disease.
29 A subset of the B2 strains (ST73, ST681 and ST127) carry the *pks* genomic island, which
30 encodes colibactin, a genotoxin associated with colorectal cancer. We found little
31 antimicrobial resistance and only one example of multi-drug resistance among the simian
32 isolates. Hierarchical clustering showed that simian isolates from ST442 and ST349 are
33 closely related to isolates recovered from human clinical cases (differences in 50 and seven
34 alleles respectively), suggesting recent exchange between the two host species. Conversely,
35 simian isolates from ST73, ST681 and ST127 were distinct from human isolates, while five
36 simian isolates belong to unique core-genome ST complexes—indicating novel diversity
37 specific to the primate niche. Our results are of public health importance, considering the
38 increasing contact between humans and wild non-human primates.

39

40 **Keywords**

41 Non-human primates, *Escherichia coli*, phylogenomic diversity, Extraintestinal pathogenic *E.*
42 *coli*.

43 **Impact statement**

44 Little is known about the population structure, virulence potential and the burden of
 45 antimicrobial resistance among *Escherichia coli* from wild non-human primates, despite
 46 increased exposure to humans through the fragmentation of natural habitats. Previous studies,
 47 primarily involving captive animals, have highlighted the potential for bacterial exchange
 48 between non-human primates and humans living nearby, including strains associated with
 49 intestinal pathology. Using multiple-colony sampling and whole-genome sequencing, we
 50 investigated the strain distribution and population structure of *E. coli* from wild non-human
 51 primates from the Gambia. Our results indicate that these monkeys harbour strains that can
 52 cause extraintestinal infections in humans. We document the transmission of virulent *E. coli*
 53 strains between monkeys of the same species sharing a common habitat and evidence of
 54 recent interaction between strains from humans and wild non-human primates. Also, we
 55 present complete genome assemblies for five novel sequence types of *E. coli*.

57 **Author notes**

58 All supporting data, code and protocols have been provided within the article or through
 59 supplementary data files. Nine supplementary figures and six supplementary files are
 60 available with the online version of this article.

62 **Abbreviations**

63 ExPEC, Extraintestinal pathogenic *Escherichia coli*; ST, Sequence type; AMR,
 64 Antimicrobial resistance; MLST, Multi-locus sequence typing; VFDB, Virulence factors
 65 database; SNP, single nucleotide polymorphism; SPRI, Solid phase reversible
 66 immobilisation.

68 **Data summary**

69 The raw sequences and polished assemblies from this study are available in the National
70 Center for Biotechnology Information (NCBI) Short Read Archive, under the BioProject
71 accession number PRJNA604701. The full list and characteristics of these strains and other
72 reference strains used in the analyses are presented in Table 1 and Supplementary Files 1-4
73 (available with the online version of this article).

74

75 **Introduction**

76 *Escherichia coli* is a highly versatile species, capable of adapting to a wide range of
 77 ecological niches and colonising a diverse range of hosts (1, 2). In humans, *E. coli* colonises
 78 the gastrointestinal tract as a commensal, as well as causing intestinal and extraintestinal
 79 infection (2). *E. coli* is also capable of colonising the gut in non-human primates (3-5), where
 80 data from captive animals suggest that gut isolates are dominated by phylogroups B1 and A,
 81 which, in humans, encompass commensals as well as strains associated with intestinal
 82 pathology (6-9). *E. coli* strains encoding colibactin, or cytotoxic necrotising factor 1 have
 83 been isolated from healthy laboratory rhesus macaques (4, 10), while enteropathogenic *E.*
 84 *coli* strains can—in the laboratory—cause colitis in marmosets (11), rhesus macaques
 85 infected with simian immunodeficiency virus (12) and cotton-top tamarins (13).

86 There are two potential explanations for the co-occurrence of *E. coli* in humans and non-
 87 human primates. Some bacterial lineages may have been passed on through vertical
 88 transmission within the same host species for long periods, perhaps even arising from
 89 ancestral bacteria that colonised the guts of the most recent common ancestors of humans and
 90 non-human primate species (14-16). In such a scenario, isolates from non-human primates
 91 would be expected to be novel and distinct from the diversity seen in humans. However, there
 92 is also clearly potential for horizontal transfer of strains from one host species to another
 93 (17).

94 The exchange of bacteria between humans and human-habituated animals, particularly
 95 non-human primates, is of interest in light of the fragmentation of natural habitats globally
 96 (18-28). We have seen that wild non-human primates in the Gambia are frequently exposed
 97 to humans through tourism, deforestation and urbanisation. In Uganda, PCR-based studies
 98 have suggested transmission of *E. coli* between humans, non-human primates and livestock
 99 (26-28). Thus, wild non-human primates may constitute a reservoir for the zoonotic spread of

E. coli strains associated with virulence and antimicrobial resistance to humans. Alternatively, humans might provide a reservoir of strains with the potential for anthroponotic spread to animals—or transmission might occur in both directions (29).

We do not know how many different lineages can co-exist within the same non-human primate host. Such information may help us contextualise the potential risks associated with transmission of bacterial strains between humans and non-human primates. In humans, up to eleven serotypes could be sampled from picking eleven colonies from individual stool samples (30).

To address these issues, we have exploited whole-genome sequencing to explore the colonisation patterns, population structure and phylogenomic diversity of *E. coli* in wild non-human primates from rural and urban Gambia.

Methods

Study population and sample collection

In June 2017, wild non-human primates were sampled from six sampling sites in the Gambia: Abuko Nature Reserve (riparian forest), Bijilo Forest Park (coastal fenced woodland), Kartong village (mangrove swamp), Kiang West National park (dry-broad-leaf forest), Makasutu Cultural Forest (ecotourism woodland) and River Gambia National park (riparian forest) (Figure 1). We sampled all four of the diurnal non-human primate species indigenous to the Gambia. Monkeys in Abuko and Bijilo are frequently hand-fed by visiting tourists, despite prohibiting guidelines (31).

Troops of monkeys were observed and followed. We collected a single freshly passed formed stool specimen from 43 visibly healthy individuals (38 adults, 5 juveniles; 24 females, 11 males, 8 of undetermined sex), drawn from four species: *Erythrocebus patas* (patas monkey), *Papio papio* (Guinea baboon), *Chlorocebus sabaeus* (green monkey) and

Piliocolobus badius (Western colobus monkey). Stool samples were immediately placed into sterile falcon tubes, taking care to collect portions of stool material that had not touched the ground, then placed on dry ice and stored at 80°C within 6 h. The sample processing flow is summarised in Figure 2.

Microbiological processing

For the growth and isolation of *E. coli*, 0.1–0.2 g aliquots were taken from each stool sample into 1.5 ml microcentrifuge tubes under aseptic conditions. To each tube, 1 ml of physiological saline (0.85%) was added, and the saline-stool samples were vortexed for 2 min at 4200 rpm. The homogenised samples were taken through four ten-fold serial dilutions and a 100 µl aliquot from each dilution was spread on a plate of tryptone-bile-X-glucoronide agar using the cross-hatching method. Plates were incubated at 37°C for 18–24 h in air. Colony counts were performed for each serial dilution, counting translucent colonies with blue-green pigmentation and entire margins as *E. coli*. Up to five colonies from each sample were sub-cultured on MacConkey agar at 37°C for 18–24 h and then stored in 20% glycerol broth at -80°C.

Genomic DNA extraction

A single colony from each subculture was picked into 1 ml Luria-Bertani broth and incubated overnight at 37°C. Broth cultures were spun at 3500rpm for 2 min and lysed using lysozyme, proteinase K, 10% SDS and RNase A in Tris EDTA buffer (pH 8.0). Suspensions were placed on a thermomixer with vigorous shaking at 1600 rpm, first at 37°C for 25 min and subsequently at 65°C for 15 min. DNA was extracted using solid-phase reversible immobilisation magnetic beads (Bectec Coulter Inc., Brea, CA, U.S.A.), precipitated with ethanol, eluted in Tris-Cl and evaluated for protein and RNA contamination using A_{260}/A_{280}

and A_{260}/A_{230} ratios on the NanoDrop 2000 Spectrophotometer (Fisher Scientific, Loughborough, UK). DNA concentrations were measured using the Qubit HS DNA assay (Invitrogen, MA, USA). DNA was stored at -20°C .

Illumina sequencing

Whole-genome sequencing was carried out on the Illumina NextSeq 500 platform (Illumina, San Diego, CA). We used a modified Nextera XT DNA protocol for the library preparation as follows. The genomic DNA was normalised to $0.5\text{ ng }\mu\text{l}^{-1}$ with 10 mM Tris-HCl. Next, $0.9\text{ }\mu\text{l}$ of Tagment DNA buffer (Illumina Catalogue No. 15027866) was mixed with $0.09\text{ }\mu\text{l}$ of Tagment DNA enzyme (Illumina Catalogue No. 15027865) and $2.01\text{ }\mu\text{l}$ of PCR-grade water in a master-mix. Next, $3\text{ }\mu\text{l}$ of the master-mix was added to a chilled 96-well plate. To this, $2\text{ }\mu\text{l}$ of normalised DNA (1 ng total) was added, pipette-mixed and the reaction heated to 55°C for 10 min on a PCR block. To each well, we added $11\text{ }\mu\text{l}$ of KAPA2G Robust PCR master-mix (Sigma Catalogue No. KK5005), comprising $4\text{ }\mu\text{l}$ KAPA2G buffer, $0.4\text{ }\mu\text{l}$ dNTPs, $0.08\text{ }\mu\text{l}$ polymerase and $6.52\text{ }\mu\text{l}$ PCR-grade water, contained in the kit per sample. Next, $2\text{ }\mu\text{l}$ each of P7 and P5 Nextera XT Index Kit v2 index primers (Illumina Catalogue numbers FC-131-2001 to 2004) were added to each well. Finally, the $5\text{ }\mu\text{l}$ of Tagmentation mix was added and mixed. The PCR was run as follows: 72°C for 3 min, 95°C for 1 min, 14 cycles of 95°C for 10 sec, 55°C for 20 sec and 72°C for 3 min. Following the PCR, the libraries were quantified using the Quant-iT dsDNA Assay Kit, high sensitivity kit (Catalogue No. 10164582) and run on a FLUOstar Optima plate reader. After quantification, libraries were pooled in equal quantities. The final pool was double-SPRI size-selected between 0.5 and $0.7\times$ bead volumes using KAPA Pure Beads (Roche Catalogue No. 07983298001). We then quantified the final pool on a Qubit 3.0 instrument (Invitrogen, MA, USA) and ran it on a high sensitivity D1000 ScreenTape (Agilent Catalogue No. 5067-5579) using the Agilent TapeStation 4200 to

calculate the final library pool molarity. The pooled library was run at a final concentration of 1.8 pM on an Illumina NextSeq500 instrument using a mid-output flow cell (NSQ® 500 Mid Output KT v2 300 cycles; Illumina Catalogue No. FC-404-2003) following the Illumina recommended denaturation and loading parameters, which included a 1% PhiX spike (PhiX Control v3; Illumina Catalogue FC-110-3001). The data was uploaded to BaseSpace (<http://www.basespace.illumina.com>) and then converted to FASTQ files.

Oxford nanopore sequencing

We used the rapid barcoding kit (Oxford Nanopore Catalogue No. SQK-RBK004) to prepare libraries according to the manufacturer's instructions. We used 400 ng DNA for library preparation and loaded 75 µl of the prepared library on an R9.4 MinION flow cell. The size of the DNA fragments was assessed using the Agilent 2200 TapeStation (Agilent Catalogue No. 5067-5579) before sequencing. The concentration of the final library pool was measured using the Qubit high-sensitivity DNA assay (Invitrogen, MA, USA).

Genome assembly and phylogenetic analysis

Sequences were analysed on the Cloud Infrastructure for Microbial Bioinformatics (32). Paired-end short-read sequences were concatenated, then quality-checked using FastQC v0.11.7 (33). Reads were assembled using Shovill (<https://github.com/tseemann/shovill>) and assemblies assessed using QUAST v 5.0.0, de6973bb (34). Draft bacterial genomes were annotated using Prokka v 1.13 (35). Multi-locus sequence types were called from assemblies according to the Achtman scheme using the mlst software (<https://github.com/tseemann/mlst>) to scan alleles in PubMLST (<https://pubmlst.org/>) (36). To identify and assign new STs, we used the ST search algorithm in EnteroBase, allowing for one allele mismatch (37). Snippy v4.3.2 (<https://github.com/tseemann/snippy>) was used for variant calling and core genome

alignment, including references genome sequences representing the major phylogroups of *E. coli* and *Escherichia fergusonii* as an outgroup (Supplementary File 1B). We used Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) to detect and remove recombinant regions of the core genome alignment (38). RAxML v 8.2.4 (39) was used for maximum-likelihood phylogenetic inference from this masked alignment based on a general time-reversible nucleotide substitution model with 1,000 bootstrap replicates. The phylogenetic tree was visualised using Mega v. 7.2 (40) and annotated using Adobe Illustrator v 23.0.3 (Adobe Inc., San Jose, California). Pair-wise single nucleotide polymorphism (SNP) distances between genomes were computed from the core-gene alignment using snp-dists v0.6 (<https://github.com/tseemann/snp-dists>).

Population structure and analysis of gene content

Merged short reads were uploaded to EnteroBase (41) where we used the Hierarchical Clustering (HierCC) algorithm to assign our genomes from non-human primates to HC1100 clusters, which in *E. coli* correspond roughly to the clonal complexes seen in seven-allele MLST. Core genome MLST (cgMLST) profiles based on the typing of 2, 512 core loci for *E. coli* facilitates single-linkage hierarchical clustering according to fixed core genome MLST (cgMLST) allelic distances, based on cgMLST allelic differences. Thus, cgST HierCC provides a robust approach to analyse population structures at multiple levels of resolution. The identification of closely-related genomes using HierCC has been shown to be 89% consistent between cgMLST and single-nucleotide polymorphisms (42). Neighbour-joining trees were reconstructed with Ninja—a hierarchical clustering algorithm for inferring phylogenies that is capable of scaling to inputs larger than 100,000 sequences (43). ARIBA v2.12.1 (44) was used to search short reads against the Virulence Factors Database (45) (VFDB-core) (virulence-associated genes), ResFinder (AMR) (46) and

PlasmidFinder (plasmid-associated genes) (47) databases (both ResFinder and PlasmidFinder databases downloaded 29 October 2018). Percentage identity of $\geq 90\%$ and coverage of $\geq 70\%$ of the respective gene length were taken as a positive result. Analyses were performed on assemblies using ABRicate v 0.8.7 (<https://github.com/tseemann/abricate>). A heat map of detected virulence- and AMR-associated genes was plotted on the phylogenetic tree using ggtree and phangorn in R studio v 3.5.1. We searched EnteroBase for all *E. coli* strains isolated from humans in the Gambia (n=128), downloaded the genomes and screened them for resistance genes using ABRicate v 0.9.8. Assembled genomes for isolates that clustered with our colibactin-encoding ST73, ST127 and ST681 isolates were downloaded and screened for the colibactin operon using ABRicate's VFDB database (accessed 28 July 2019). Assemblies reported to contain colibactin genes were aligned against the colibactin-encoding *Escherichia coli* IHE3034 reference genome (NCBI Accession: GCA_000025745.1) using minimap2 2.13-r850. BAM files were visualised in Artemis Release 17.0.1 (48) to confirm the presence of the *pks* genomic island which encodes the colibactin operon.

Hybrid assembly and analysis of plasmids and phages

Base-called FASTQ files were concatenated into a single file and demultiplexed into individual FASTQ files based on barcodes, using the qcat python command-line tool v 1.1.0 (<https://github.com/nanoporetech/qcat>). Hybrid assemblies of the Illumina and nanopore reads were created with Unicycler (49). The quality and completion of the hybrid assemblies were assessed with QUAST v 5.0.0, de6973bb and CheckM (34, 50). Hybrid assemblies were interrogated using ABRicate PlasmidFinder and annotated using Prokka (35). Plasmid sequences were visualised in Artemis using coordinates from ABRicate. Prophage identification was carried out using the phage search tool, PHASTER (51).

250

251 **Antimicrobial susceptibility**

252 We determined the minimum inhibitory concentrations of amikacin, trimethoprim,
253 sulfamethoxazole, ciprofloxacin, cefotaxime and tetracycline for the isolates from non-human
254 primates using agar dilution (52). Two-fold serial dilutions of each antibiotic were performed
255 in molten Mueller-Hinton agar (Oxoid, Basingstoke, UK), from 32mg/L to 0.03 mg l⁻¹ (512
256 mg l⁻¹ to 0.03 mg l⁻¹ for sulfamethoxazole), using *E. coli* NCTC 10418 as control. MICs were
257 performed in duplicate and interpreted using breakpoint tables from the European Committee
258 on Antimicrobial Susceptibility Testing v. 9.0, 2019 (<http://www.eucast.org>).

259

260 **Results**

261 Twenty-four of 43 samples (56%) showed growth indicative of *E. coli*, yielding a total of 106
262 colonies. The isolates were designated by the primate species and the site from which they
263 were sampled as follows: *Chlorocebus sabaeus*, ‘Chlos’; *Papio papio*, ‘Pap’; *Piliocolobus*
264 *badius*, ‘Prob’; Abuko Nature Reserve, ‘AN’; Bijilo Forest Park, ‘BP’; Kartong village, ‘K’;
265 Kiang West National Park, ‘KW’; Makasutu Cultural Forest, ‘M’; and River Gambia
266 National Park, ‘RG’. After genome sequencing, five isolates (PapRG-04, (n=1); PapRG-03
267 (n=1); ChlosRG-12 (n=1); ChlosAN-13 (n=1); ProbAN-19 (n=1)) were excluded due to low
268 depth of coverage (<20x), leaving 101 genomes for subsequent analysis (Table 1).

269 We recovered 43 seven-allele sequence types (ten of them novel), spanning five of the
270 eight known phylogroups of *E. coli* and comprising 38 core-genome MLST complexes
271 (Figure 3). The majority of strains belonged to phylogroup B2 (42/101, 42%), which
272 encompasses strains that cause extraintestinal infections in humans (ExPEC strains) (6-8).
273 Strains from phylogroup B2 carried colonisation and fitness factors associated with
274 extraintestinal disease in humans (Figure 3). A subset of the B2 strains (13/42, 31%),

belonging to STs 73, 681 and 127, carried the *pks* genomic island, which encodes the DNA alkylating genotoxin, colibactin. Colibactin-encoding *E. coli* frequently cause colorectal cancer, urosepsis, bacteraemia and prostatitis, and are highly associated with other virulence factors such as siderophores and toxins (53-56).

Thirteen individuals were colonised by two or more STs and nine by two or more phylogroups (Supplementary File 1A). Five colony picks from a single Guinea baboon (PapRG-06) yielded five distinct STs, two of which are novel. Two green monkeys sampled from Bijilo (ChlosBP-24 and ChlosBP-25) shared an identical ST73 genotype, while two Guinea baboons from Abuko shared an ST226 strain—documenting transmission between monkeys of the same species. Among the monkey isolates, we found several STs associated with extraintestinal infections and/or AMR in humans: ST73, ST681, ST127, ST226, ST336, ST349 (57-62).

In seventeen monkeys, we observed a cloud of closely related genotypes (separated by 0-5 SNPs, Table 2A) from each strain, suggesting evolution within the host after acquisition of the strain. However, in two individuals, pair-wise SNP distances between genotypes from the same ST were substantial enough (25 SNPs and 77 SNPs) to suggest multiple acquisitions of each strain (Table 2B).

We identified the closest neighbours to all the recovered strains from our study (Table 3). Our results suggest, in some cases, recent interactions between humans or livestock and non-human primates. However, we also found a diversity of strains specific to the non-human primate niche. Hierarchical clustering analysis revealed that simian isolates from ST442 and ST349 (Achtman)—sequence types that are associated with virulence and AMR in humans (49, 55)—were closely related to human clinical isolates, with differences of 50 alleles and seven alleles in the core-genome MLST scheme respectively (Supplementary Figures 1-2). Similarly, we found evidence of recent interaction between simian ST939 isolates and strains

from livestock (Supplementary Figure 3). Conversely, simian ST73, ST127 and ST681 isolates were genetically distinct from human isolates from these sequence types (Supplementary Figures 4-6). The multi-drug resistant isolate PapAN-14-1 from ST349 was, however, closely related to an environmental isolate recovered from water (Supplementary Figure 7).

Five isolates were >1000 alleles away in the core-genome MLST scheme from anything in EnteroBase (Supplementary Figures 8 & 9). Four of these were assigned to novel sequence types in the seven-allele scheme (Achtman) (ST8550, ST8525, ST8532, ST8826), while one belonged to ST1873, which has only two other representatives in EnteroBase: one from a species of wild bird from Australia (*Sericornis frontalis*); the other from water. Besides, ST8550, ST8525, ST8532, ST8826 belonged to novel HierCC 1100 groups (cgST complexes), indicating that they were unrelated to any other publicly available *E. coli* genomes.

We observed few antimicrobial resistance genes in our study population, compared to what prevails in isolates from humans in the Gambia (Figure 4). Phenotypic resistance to single agents was confirmed in ten isolates: to trimethoprim in a single isolate, to sulfamethoxazole in four unrelated isolates and to tetracycline in four closely related isolates from a single animal. A single ST2076 (Achtman) isolate (PapAN-14-1) belonging to the ST349 lineage was resistant to trimethoprim, sulfamethoxazole and tetracycline. The associated resistance genes were harboured on an IncFIB plasmid.

Eighty percent (81/101) of the study isolates harboured one or more plasmids. We detected the following plasmid replicon types: IncF (various subtypes), IncB/K/O/Z, I1, IncX4, IncY, Col plasmids (various subtypes) and plasmids related to pO111 (rep B) (Supplementary File 2A). Long-read sequencing of six representative samples showed that the IncFIB plasmids encoded acquired antibiotic resistance, fimbrial adhesins and colicins

(Supplementary File 2B). Also, the IncFIC/FII, ColRNAI, Col156 and IncB/O/K/Z plasmids encoded fimbrial proteins and colicins. Besides, the IncX and Inc-I-Aplha encoded bundle forming pili *bfpB* and the heat-stable enterotoxin protein *StbB* respectively.

We generated complete genome sequences of five novel sequence types of *E. coli* (ST8525, ST8527, ST8532, ST8826, ST8827) within the seven-allele scheme (Achtman) (Supplementary File 3A) (63). Although none of these new genomes encoded AMR genes, one of them (PapRG-04-4) contained an IncFIB plasmid encoding fimbrial proteins, and a cryptic ColRNA plasmid. PHASTER identified thirteen intact prophages and four incomplete phage remnants (Supplementary File 3B). Two pairs of genomes from Guinea baboons from different parks shared common prophages: one pair carrying PHAGE_Enterog_933W, the other PHAGE-Enterog_lambda.

Discussion

We have described the population structure of *E. coli* in diurnal non-human primates living in rural and urban habitats from the Gambia. Although our sample size was relatively small, we have recovered isolates that span the diversity previously described in humans and have also identified ten new sequence types (five of them now with complete genome sequences). This finding is significant, considering the vast number of *E. coli* genomes that have been sequenced to date (9, 597 with MLST via sanger sequencing, and 127, 482 via WGS) (64).

Increasing contact between animal species facilitates the potential exchange of pathogens. Accumulating data shows that ExPEC strains are frequently isolated from diseased companion animals and livestock—highlighting the potential for zoonotic as well as anthroponotic transmission (65-70). In a previous study, green monkeys from Bijilo Park were found to carry lineages of *Staphylococcus aureus* thought to be acquired from humans (31). Our analyses suggest similar exchange of *E. coli* strains between humans and wild non-

human primates. However, non-human primates also harbour *E. coli* genotypes that are clinically important in humans, such as ST73, ST127 and ST681, yet are distinct from those circulating in humans—probably reflecting lineages that have existed in this niche for long periods.

We found that several monkeys were colonised with multiple STs, often encompassing two or more phylotypes. Although colonisation with multiple serotypes of *E. coli* is common in humans (30, 71) we were surprised to identify as many as five STs in a single baboon. Sampling multiple colonies from single individuals also revealed within-host diversity arising from microevolution. However, we also found evidence of acquisition in the same animal of multiple lineages of the same sequence type, although it is unclear whether this reflects a single transmission event involving more than one strain or serial transfers.

Antimicrobial resistance in wildlife is known to spread on plasmids through horizontal gene transfer (72). Given the challenge of resolving large plasmids using short-read sequences (73), we exploited long-read sequencing to document the contribution of plasmids to the genomic diversity that we observed in our study population. Consistent with previous reports (74), we found IncF plasmids which encoded antimicrobial resistance genes. Virulence-encoding plasmids, particularly colicin-encoding and the F incompatibility group ones, have long been associated with several pathotypes of *E. coli* (75). Consistent with this, we found plasmids that contributed to the dissemination of virulence factors such as the heat-stable enterotoxin protein *StbB*, colicins and fimbrial proteins.

This study could have been enhanced by sampling human populations living near those of our non-human primates; however, we compensated for this limitation by leveraging the wealth of genomes in publicly available databases. Besides, we did not sample nocturnal monkeys due to logistic challenges; however, these have more limited contact with humans than the diurnal species. Despite these limitations, however, this study provides insight into

the diversity and colonisation patterns of *E. coli* among non-human primates in the Gambia, highlighting the impact of human continued encroachment on natural habitats and revealing important phylogenomic relationships between strains from humans and non-human primates.

References

1. Sousa CP. The versatile strategies of *Escherichia coli* pathotypes: a mini review. Journal of Venomous Animals and Toxins including Tropical Diseases. 2006;12:363-73.
2. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. Nature Reviews Microbiology. 2004;2(2):123-40.
3. Carvalho VM, Irino K, Onuma D, Castro AFPd. Random amplification of polymorphic DNA reveals clonal relationships among enteropathogenic *Escherichia coli* isolated from non-human primates and humans. Brazilian Journal of Medical and Biological Research. 2007;40:237-41.
4. Martin HR, Taylor NS, Buckley EM, Marini RP, Patterson MM, Fox JG. Characterization of cytotoxic necrotizing factor 1-producing *Escherichia coli* strains from faeces of healthy macaques. Journal of Medical Microbiology. 2009;58(10):1354-8.
5. enaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nature Reviews Microbiology. 2010;8(3):207-17.
6. 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. Microbial Genomics. 2018;4(7):e000192.

- 400 7. Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur
401 E. A specific genetic background is required for acquisition and expression of
402 virulence factors in *Escherichia coli*. Molecular Biology and Evolution.
403 2004;21(6):1085-94.
- 404 8. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, et al. The link
405 between phylogeny and virulence in *Escherichia coli* extraintestinal infection.
406 Infection and Immunity. 1999;67(2):546-53.
- 407 9. Clayton JB, Danzeisen JL, Trent AM, Murphy T, Johnson TJ. Longitudinal
408 characterization of *Escherichia coli* in healthy captive non-human primates. Frontiers
409 in Veterinary Science. 2014;1:24.
- 410 10. Feng Y, Mannion A, Madden CM, Swennes AG, Townes C, Byrd C, et al. Cytotoxic
411 *Escherichia coli* strains encoding colibactin and cytotoxic necrotizing factor (CNF)
412 colonize laboratory macaques. Gut Pathogens. 2017;9:71.
- 413 11. Thomson JA, Scheffler JJ. Hemorrhagic typhlocolitis associated with attaching and
414 effacing *Escherichia coli* in common marmosets. Laboratory Animal Science.
415 1996;46(3):275-9.
- 416 12. Mansfield KG, Lin KC, Newman J, Schauer D, MacKey J, Lackner AA, et al.
417 Identification of enteropathogenic *Escherichia coli* in simian immunodeficiency
418 virus-infected infant and adult rhesus macaques. Journal of Clinical Microbiology.
419 2001;39(3):971-6.
- 420 13. Mansfield KG, Lin KC, Xia D, Newman JV, Schauer DB, MacKey J, et al.
421 Enteropathogenic *Escherichia coli* and ulcerative colitis in cotton-top tamarins
422 (*Saguinus oedipus*). The Journal of Infectious Diseases. 2001;184(6):803-7.

- 423 14. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds:
424 evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*.
425 2008;6(10):776-88.
- 426 15. Muegge BD, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, et al.
427 Diet drives convergence in gut microbiome functions across mammalian phylogeny
428 and within humans. *Science (New York, NY)*. 2011;332(6032):970.
- 429 16. Moeller AH, Caro-Quintero A, Mjunga D, Georgiev AV, Lonsdorf EV, Muller MN,
430 et al. Cospeciation of gut microbiota with hominids. *Science (New York, NY)*.
431 2016;353(6297):380-2.
- 432 17. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza
433 Y, et al. Meta-analyses of studies of the human microbiota. *Genome Research*.
434 2013;23(10):1704-14.
- 435 18. Iovine RdO, Dejuste C, Miranda F, Filoni C, Bueno MG, de Carvalho VM. Isolation
436 of *Escherichia coli* and *Salmonella* spp. from free-ranging wild animals. *Brazilian*
437 *Journal of Microbiology*. 2015;46(4):1257-63.
- 438 19. Benavides JA, Godreuil S, Bodenham R, Ratiarison S, Devos C, Petretto M-O, et al.
439 No evidence for transmission of antibiotic-resistant *Escherichia coli* strains from
440 humans to wild western lowland gorillas in Lopé National Park, Gabon. *Applied and*
441 *Environmental Microbiology*. 2012;78(12):4281-7.
- 442 20. Ryan SJ, Walsh PD. Consequences of non-intervention for infectious disease in
443 African great apes. *PLOS one*. 2011;6(12):e29030-e.
- 444 21. Bublitz DC, Wright PC, Rasambainarivo FT, Arrigo-Nelson SJ, Bodager JR,
445 Gillespie TR. Pathogenic enterobacteria in lemurs associated with anthropogenic
446 disturbance. *American Journal of Primatology*. 2015;77(3):330-7.

- 447 22. Calvignac-Spencer S, Leendertz SA, Gillespie TR, Leendertz FH. Wild great apes as
448 sentinels and sources of infectious disease. *Clinical Microbiology and Infection*.
449 2012;18(6):521-7.
- 450 23. Daszak P, Cunningham AA, Hyatt AD. Anthropogenic environmental change and the
451 emergence of infectious diseases in wildlife. *Acta Tropica*. 2001;78(2):103-16.
- 452 24. Weiss D, Wallace RM, Rwego IB, Gillespie TR, Chapman CA, Singer RS, et al.
453 Antibiotic-resistant *Escherichia coli* and class 1 integrons in humans, domestic
454 animals, and wild primates in rural Uganda. *Applied and Environmental*
455 *Microbiology*. 2018;84(21):e01632-18.
- 456 25. Dobson A, Foufopoulos J. Emerging infectious pathogens of wildlife. *Philosophical*
457 *Transactions of the Royal Society of London*. 2001;356(1411):1001-12.
- 458 26. Goldberg TL, Gillespie TR, Rwego IB, Estoff EL, Chapman CA. Forest
459 fragmentation as cause of bacterial transmission among non-human primates, humans,
460 and livestock, Uganda. *Emerging Infectious Diseases*. 2008;14(9):1375-82.
- 461 27. Goldberg TL, Gillespie TR, Rwego IB, Wheeler E, Estoff EL, Chapman CA. Patterns
462 of gastrointestinal bacterial exchange between chimpanzees and humans involved in
463 research and tourism in western Uganda. *Biological Conservation*. 2007;135(4):511-
464 7.
- 465 28. Rwego IB, Isabirye-Basuta G, Gillespie TR, Goldberg TL. Gastrointestinal bacterial
466 transmission among humans, mountain gorillas, and livestock in Bwindi Impenetrable
467 National Park, Uganda. *Conservation Biology*. 2008;22(6):1600-7.
- 468 29. Rwego IB, Isabirye-Basuta G, Gillespie TR, Goldberg TL. Gastrointestinal bacterial
469 transmission among humans, mountain gorillas, and livestock in Bwindi Impenetrable
470 National Park, Uganda. *Conservation Biology*. 2008;22(6):1600-7.

- 471 30. Lidin-Janson G, Kaijser B, Lincoln K, Olling S, Wedel H. The homogeneity of the
472 faecal coliform flora of normal school-girls, characterized by serological and
473 biochemical properties. *Medical Microbiology and Immunology*. 1978;164(4):247-53.
- 474 31. Senghore M, Bayliss SC, Kwambana-Adams BA, Foster-Nyarko E, Manneh J, Dione
475 M, et al. Transmission of *Staphylococcus aureus* from humans to green monkeys in
476 the Gambia as revealed by whole-genome sequencing. *Applied and Environmental*
477 *Microbiology*. 2016;82(19):5910-7.
- 478 32. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, et al.
479 CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource
480 for the medical microbiology community. *Microbial Genomics*. 2016;2(9):e000008.
- 481 33. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality
482 control. *F1000Res*. 2018;7:1338.
- 483 34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
484 genome assemblies. *Bioinformatics (Oxford, England)*. 2013;29(8):1072-5.
- 485 35. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford,*
486 *England)*. 2014;30(14):2068-9.
- 487 36. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at
488 the population level. *BMC Bioinformatics*. 2010;11(1):595.
- 489 37. Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, et al. Multilocus
490 sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLOS*
491 *Pathogens*. 2012;8(6):e1002776.
- 492 38. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid
493 phylogenetic analysis of large samples of recombinant bacterial whole genome
494 sequences using Gubbins. *Nucleic Acids Research*. 2015;43(3):e15.

495 39. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
496 large phylogenies. *Bioinformatics* (Oxford, England). 2014;30(9):1312-3.

497
498 40. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular
499 Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*.
500 2013;30(12):2725-9.

501 41. Zhou Z, Alikhan N-F, Mohamed K, Achtman M. The user's guide to comparative
502 genomics with Enterobase. Three case studies: micro-clades within *Salmonella*
503 *enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core
504 genomic diversity of all *Escherichia*. *BioRxiv*. 2019:613554.

505 42. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Göker M, Riedel T, et al. Global
506 genomic population structure of *Clostridioides difficile*. *BioRxiv*. 2019:727230.

507 43. Wheeler TJ. Large-scale neighbor-joining with NINJA. In: *Proceedings of the 9th*
508 *Workshop on Algorithms in Bioinformatics*. 2009; 5724:375-389.

509 44. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA:
510 rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial*
511 *Genomics*. 2017;3(10):e000131.

512 45. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic
513 platform with an interactive web interface. *Nucleic Acids Research*.
514 2019;47(D1):D687-D692.

515 46. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al.
516 Identification of acquired antimicrobial resistance genes. *The Journal of*
517 *Antimicrobial Chemotherapy*. 2012;67(11):2640-4.

518

519 47. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, et al.
520 In silico detection and typing of plasmids using PlasmidFinder and plasmid
521 multilocus sequence typing. *Antimicrobial Agents Chemother.* 2014;58(7):3895-903.

522 48. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated
523 platform for visualization and analysis of high-throughput sequence-based
524 experimental data. *Bioinformatics (Oxford, England).* 2012;28(4):464-9.

525 49. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
526 assemblies from short and long sequencing reads. *PLOS Computational Biology.*
527 2017;13(6):e1005595.

528 50. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing
529 the quality of microbial genomes recovered from isolates, single cells, and
530 metagenomes. *Genome Research.* 2015;25(7):1043-55.

531 51. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better,
532 faster version of the PHAST phage search tool. *Nucleic Acids Research.*
533 2016;44(W1):W16-21.

534 52. Wiegand I, Hilpert K, Hancock RE. Agar and broth dilution methods to determine the
535 minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature*
536 *Protocols.* 2008;3(2):163-75.

537 53. Krieger JN, Dobrindt U, Riley DE, Oswald E. Acute *Escherichia coli* prostatitis in
538 previously healthy young men: bacterial virulence factors, antimicrobial resistance, and
539 clinical outcomes. *Urology.* 2011;77(6):1420-5.

540 54. Fais T, Delmas J, Barnich N, Bonnet R, Dalmasso G. Colibactin: More than a new
541 bacterial toxin. *Toxins (Basel).* 2018;10(4):151.

- 542 55. Johnson JR, Johnston B, Kuskowski MA, Nougayrede JP, Oswald E. Molecular
543 epidemiology and phylogenetic distribution of the *Escherichia coli* pks genomic
544 island. Journal of Clinical Microbiology. 2008;46(12):3906-11.
- 545 56. Micenkova L, Benova A, Frankovicova L, Bosak J, Vrba M, Sevcikova A, et al.
546 Human *Escherichia coli* isolates from hemocultures: Septicemia linked to urogenital
547 tract infections is caused by isolates harboring more virulence genes than bacteraemia
548 linked to other conditions. International Journal of Medical Microbiology.
549 2017;307(3):182-9.
- 550 57. Ho W-S, Gan H-M, Yap K-P, Balan G, Yeo CC, Thong K-L. Genome sequence of
551 multidrug-resistant *Escherichia coli* EC302/04, isolated from a human tracheal
552 aspirate. Journal of Bacteriology. 2012;194(23):6691-2.
- 553 58. Manges AR, Johnson JR. Reservoirs of extraintestinal pathogenic *Escherichia coli*.
554 Microbiology Spectrum. 2015;3(5): UTI-0006-2012.
- 555 59. Kamjumphol W, Wongboot W, Suebwongsa N, Kluabwang P, Chantaroj S, Okada K.
556 Draft genome sequence of a colistin-resistant *Escherichia coli* ST226: A clinical
557 strain harbouring an mcr-1 variant. Journal of Global Antimicrobial Resistance.
558 2019;16:168-9.
- 559 60. Markovska R, Stoeva T, Boyanova L, Stankova P, Schneider I, Keuleyan E, et al.
560 Multicentre investigation of carbapenemase-producing *Klebsiella pneumoniae* and
561 *Escherichia coli* in Bulgarian hospitals – Interregional spread of ST11 NDM-1-
562 producing *K. pneumoniae*. Infection, Genetics and Evolution. 2019;69:61-7.
- 563 61. Salinas L, Cárdenas P, Johnson TJ, Vasco K, Graham J, Trueba G. Diverse
564 Commensal *Escherichia coli* clones and plasmids disseminate antimicrobial
565 resistance genes in domestic animals and children in a semirural community in
566 Ecuador. mSphere. 2019;4(3):e00316-19.

- 567 62. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global
568 Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. Clinical Microbiology
569 Reviews. 2019;32(3):e00135-18.
- 570 63. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in
571 *Escherichia coli*: an evolutionary perspective. Molecular Microbiology.
572 2006;60(5):1136-51.
- 573 64. Zhou Z, Alikhan N-F, Mohamed K, Achtman M. The user's guide to comparative
574 genomics with Enterobase. Three case studies: micro-clades within *Salmonella*
575 *enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core
576 genomic diversity of all *Escherichia*. BioRxiv. 2019:613554.
- 577 65. Zogg AL, Zurfluh K, Schmitt S, Nuesch-Inderbinen M, Stephan R. Antimicrobial
578 resistance, multilocus sequence types and virulence profiles of ESBL producing and
579 non-ESBL producing uropathogenic *Escherichia coli* isolated from cats and dogs in
580 Switzerland. Veterinary Microbiology. 2018;216:79-84.
- 581 66. Johnson JR, Clabots C, Kuskowski MA. Multiple-host sharing, long-term persistence,
582 and virulence of *Escherichia coli* clones from human and animal household members.
583 Journal of Clinical Microbiology. 2008;46(12):4078-82.
- 584 67. Johnson JR, Owens K, Gajewski A, Clabots C. *Escherichia coli* colonization patterns
585 among human household members and pets, with attention to acute urinary tract
586 infection. Journal of Infectious Disease. 2008;197(2):218-24.
- 588 68. Johnson JR, Miller S, Johnston B, Clabots C, Debroy C. Sharing of *Escherichia coli*
589 sequence type ST131 and other multidrug-resistant and urovirulent *E. coli* strains
590 among dogs and cats within a household. Journal of Clinical Microbiology.
591 2009;47(11):3721-5.

69. Achtman M, Heuzenroeder M, Kusecek B, Ochman H, Caugant D, Selander RK, et al. Clonal analysis of *Escherichia coli* O2:K1 isolated from diseased humans and animals. *Infection and Immunity*. 1986;51(1):268-76.
70. Ewers C, Grobbel M, Stamm I, Kopp PA, Diehl I, Semmler T, et al. Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum- β -lactamase-producing *Escherichia coli* among companion animals. *Journal of Antimicrobial Chemotherapy*. 2010;65(4):651-60.
71. Schlager TA, Hendley JO, Bell AL, Whittam TS. Clonal diversity of *Escherichia coli* colonizing stools and urinary tracts of young girls. *Infection and Immunity*. 2002;70(3):1225-9.
72. Vittecoq M, Godreuil S, Prugnotte F, Durand P, Brazier L, Renaud N, et al. Antimicrobial resistance in wildlife. *Journal of Applied Ecology*. 2016;53(2):519-29.
73. A redondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial Genomics*. 2017;3(10):e000128.
74. Ca Carattoli A. Resistance plasmid families in *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy*. 2009;53(6):2227.
75. Johnson TJ, Nolan LK. Pathogenomics of the virulence plasmids of *Escherichia coli*. *Microbiology and Molecular Biology Reviews*. 2009;73(4):750-74.

Data bibliography

1. Foster-Nyarko, E. et al, NCBI BioProject PRJNA604701 (2020).
2. Forde, B. M., Ben Zakour, N. L., Stanton-Cook, M., Phan, M. D., Totsika, M. et al., 17 representative *E. coli* reference isolates (2014). NCBI accession numbers are provided in Table 1B.

3. Nougayrede J.P, Homburg S, Taieb F., Boury M., Brzuszkiewicz E., et al., *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells (2006). NCBI accession: GCA_000025745.1.

Funding information

MP, EFN, NT, AR, GT, JO and GK were supported by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent projects 44414000A and 4408000A. NFA and DB were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/ CCG1860/1). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We want to thank Dr Andrew Page and Dr Thanh Le-Viet for their thoughtful advice on the long-read analysis. We also thank Dr Mark Webber for proofreading the manuscript and giving constructive feedback.

Author contributions

Conceptualization, MA, MP; data curation, MP, NFA; formal analysis, EFN, analytical support, GT; funding, MP and MA; sample collection, JDC; laboratory experiments, EFN, DB; supervision, AR, NFA, GK, JO, MP, MA; manuscript preparation – original draft, EFN; review and editing, NT, AR, JO, NFA, MP; review of final manuscript, all authors.

Conflicts of interest

The authors have no conflicts of interest to declare.

643

644 **Ethical statement**

645 No human nor animal experimentation is reported.

646

Table 1: Study isolates

Name	Source	Individual sampling number	Colony-pick	Sampling site	ST
PapRG-03-1	<i>Papio papio</i>	3	1	River Gambia national park	336
PapRG-03-2	<i>Papio papio</i>	3	2	River Gambia national park	336
PapRG-03-3	<i>Papio papio</i>	3	3	River Gambia national park	336
PapRG-03-4	<i>Papio papio</i>	3	4	River Gambia national park	336
PapRG-03-5	<i>Papio papio</i>	3	5	River Gambia national park	336
PapRG-04-1	<i>Papio papio</i>	4	1	River Gambia national park	1665
PapRG-04-2	<i>Papio papio</i>	4	2	River Gambia national park	1204
PapRG-04-4	<i>Papio papio</i>	4	3	Makasutu cultural forest	8826
PapRG-04-5	<i>Papio papio</i>	4	4	Makasutu cultural forest	1204
PapRG-05-2	<i>Papio papio</i>	5	1	Makasutu cultural forest	1431
PapRG-05-3	<i>Papio papio</i>	5	2	Makasutu cultural forest	99
PapRG-05-4	<i>Papio papio</i>	5	3	Makasutu cultural forest	6316
PapRG-05-5	<i>Papio papio</i>	5	4	Makasutu cultural forest	1431
PapRG-06-1	<i>Papio papio</i>	6	1	Makasutu cultural forest	4080
PapRG-06-2	<i>Papio papio</i>	6	2	Makasutu cultural forest	2521
PapRG-06-3	<i>Papio papio</i>	6	3	Makasutu cultural forest	8827
PapRG-06-4	<i>Papio papio</i>	6	4	Makasutu cultural forest	1204
PapRG-06-5	<i>Papio papio</i>	6	5	River Gambia national park	8525
ProbRG-07-1	<i>Piliocolobus badius</i>	7	1	River Gambia national park	73
ProbRG-07-2	<i>Piliocolobus badius</i>	7	2	River Gambia national park	73
ProbRG-07-3	<i>Piliocolobus badius</i>	7	3	River Gambia national park	73
ProbRG-07-4	<i>Piliocolobus badius</i>	7	4	River Gambia national park	73
ProbRG-07-5	<i>Piliocolobus badius</i>	7	5	River Gambia national park	73
ChlosRG-12-1	<i>Chlorocebus sabaesus</i>	12	1	River Gambia national park	8824
ChlosRG-12-2	<i>Chlorocebus sabaesus</i>	12	2	River Gambia national park	196
ChlosRG-12-3	<i>Chlorocebus sabaesus</i>	12	3	River Gambia national park	196
ChlosRG-12-5	<i>Chlorocebus sabaesus</i>	12	4	River Gambia national park	40
ChlosAN-13-1	<i>Chlorocebus sabaesus</i>	13	1	Abuko Nature Reserve	8526
ChlosAN-13-2	<i>Chlorocebus sabaesus</i>	13	2	Abuko Nature Reserve	8550
ChlosAN-13-4	<i>Chlorocebus sabaesus</i>	13	3	Abuko Nature Reserve	1973
ChlosAN-13-5	<i>Chlorocebus sabaesus</i>	13	4	Abuko Nature Reserve	1973
PapAN-14-1	<i>Papio papio</i>	14	1	Abuko Nature Reserve	2076
PapAN-14-2	<i>Papio papio</i>	14	2	Abuko Nature Reserve	939
PapAN-14-3	<i>Papio papio</i>	14	3	Abuko Nature Reserve	226
PapAN-14-4	<i>Papio papio</i>	14	4	Abuko Nature Reserve	226
PapAN-14-5	<i>Papio papio</i>	14	5	Abuko Nature Reserve	226
PapAN-15-1	<i>Papio papio</i>	15	1	Abuko Nature Reserve	226
PapAN-15-2	<i>Papio papio</i>	15	2	Abuko Nature Reserve	5073
PapAN-15-3	<i>Papio papio</i>	15	3	Abuko Nature Reserve	226
PapAN-15-4	<i>Papio papio</i>	15	4	Abuko Nature Reserve	126
PapAN-15-5	<i>Papio papio</i>	15	5	Abuko Nature Reserve	8823
ChlosAN-17-1	<i>Chlorocebus sabaesus</i>	17	1	Abuko Nature Reserve	681
ChlosAN-17-2	<i>Chlorocebus sabaesus</i>	17	2	Abuko Nature Reserve	362
ChlosAN-17-3	<i>Chlorocebus sabaesus</i>	17	3	Abuko Nature Reserve	681
ChlosAN-17-4	<i>Chlorocebus sabaesus</i>	17	4	Abuko Nature Reserve	681
ChlosAN-18-1	<i>Chlorocebus sabaesus</i>	18	1	Abuko Nature Reserve	681
ChlosAN-18-2	<i>Chlorocebus sabaesus</i>	18	2	Abuko Nature Reserve	681
ChlosAN-18-3	<i>Chlorocebus sabaesus</i>	18	3	Abuko Nature Reserve	681
ChlosAN-18-4	<i>Chlorocebus sabaesus</i>	18	4	Abuko Nature Reserve	681
ChlosAN-18-5	<i>Chlorocebus sabaesus</i>	18	5	Abuko Nature Reserve	349

ProbAN-19-2	<i>Piliocolobus badius</i>	19	1	Abuko Nature Reserve	8825
ChlosBP-21-1	<i>Chlorocebus sabaesus</i>	21	1	Bijilo forest park	677
ChlosBP-21-2	<i>Chlorocebus sabaesus</i>	21	2	Bijilo forest park	677
ChlosBP-21-3	<i>Chlorocebus sabaesus</i>	21	3	Bijilo forest park	677
ChlosBP-21-4	<i>Chlorocebus sabaesus</i>	21	4	Bijilo forest park	677
ChlosBP-21-5	<i>Chlorocebus sabaesus</i>	21	5	Bijilo forest park	677
ChlosBP-23-1	<i>Chlorocebus sabaesus</i>	23	2	Bijilo forest park	8527
ChlosBP-23-2	<i>Chlorocebus sabaesus</i>	23	3	Bijilo forest park	8527
ChlosBP-23-3	<i>Chlorocebus sabaesus</i>	23	4	Bijilo forest park	3306
ChlosBP-24-1	<i>Chlorocebus sabaesus</i>	24	1	Bijilo forest park	73
ChlosBP-24-2	<i>Chlorocebus sabaesus</i>	24	2	Bijilo forest park	73
ChlosBP-24-3	<i>Chlorocebus sabaesus</i>	24	3	Bijilo forest park	73
ChlosBP-24-4	<i>Chlorocebus sabaesus</i>	24	4	Bijilo forest park	73
ChlosBP-24-5	<i>Chlorocebus sabaesus</i>	24	5	Bijilo forest park	73
ChlosBP-25-1	<i>Chlorocebus sabaesus</i>	25	1	Bijilo forest park	73
ChlosBP-25-2	<i>Chlorocebus sabaesus</i>	25	2	Bijilo forest park	73
ChlosBP-25-3	<i>Chlorocebus sabaesus</i>	25	3	Bijilo forest park	73
ChlosBP-25-4	<i>Chlorocebus sabaesus</i>	25	4	Bijilo forest park	73
ChlosBP-25-5	<i>Chlorocebus sabaesus</i>	25	5	Bijilo forest park	73
ChlosM-29-1	<i>Chlorocebus sabaesus</i>	29	1	Makasutu cultural forest	1873
ChlosM-29-2	<i>Chlorocebus sabaesus</i>	29	2	Makasutu cultural forest	1873
PapM-31-1	<i>Papio papio</i>	31	1	Makasutu cultural forest	2800
PapM-31-2	<i>Papio papio</i>	31	2	Makasutu cultural forest	135
PapM-31-3	<i>Papio papio</i>	31	3	Makasutu cultural forest	5780
PapM-31-4	<i>Papio papio</i>	31	4	Makasutu cultural forest	1727
PapM-31-5	<i>Papio papio</i>	31	5	Makasutu cultural forest	5780
PapM-32-1	<i>Papio papio</i>	32	2	Makasutu cultural forest	8532
PapM-32-2	<i>Papio papio</i>	32	3	Makasutu cultural forest	212
PapM-32-3	<i>Papio papio</i>	32	4	Makasutu cultural forest	212
PapM-32-4	<i>Papio papio</i>	32	5	Makasutu cultural forest	212
PapM-32-5	<i>Papio papio</i>	32	6	Makasutu cultural forest	212
PapM-33-1	<i>Papio papio</i>	33	1	Makasutu cultural forest	8533
PapM-33-2	<i>Papio papio</i>	33	2	Makasutu cultural forest	8533
PapM-33-3	<i>Papio papio</i>	33	3	Makasutu cultural forest	8533
PapM-33-4	<i>Papio papio</i>	33	4	Makasutu cultural forest	38
PapM-33-5	<i>Papio papio</i>	33	5	Makasutu cultural forest	8533
PapM-34-1	<i>Papio papio</i>	34	1	Makasutu cultural forest	676
PapM-34-2	<i>Papio papio</i>	34	2	Makasutu cultural forest	676
PapM-34-3	<i>Papio papio</i>	34	3	Makasutu cultural forest	676
PapM-34-4	<i>Papio papio</i>	34	4	Makasutu cultural forest	676
PapM-36-1	<i>Papio papio</i>	36	1	Makasutu cultural forest	8535
PapM-36-2	<i>Papio papio</i>	36	2	Makasutu cultural forest	8535
PapKW-44-1	<i>Papio papio</i>	44	1	Kiang West national park	442
PapKW-44-2	<i>Papio papio</i>	44	2	Kiang West national park	442
PapKW-44-3	<i>Papio papio</i>	44	3	Kiang West national park	442
PapKW-44-4	<i>Papio papio</i>	44	4	Kiang West national park	442
ProbK-45-1	<i>Piliocolobus badius</i>	45	1	Kartong village	127
ProbK-45-2	<i>Piliocolobus badius</i>	45	2	Kartong village	127
ProbK-45-3	<i>Piliocolobus badius</i>	45	3	Kartong village	127
ProbK-45-4	<i>Piliocolobus badius</i>	45	4	Kartong village	127
ProbK-45-5	<i>Piliocolobus badius</i>	45	5	Kartong village	127

Table 2A: Within-host single nucleotide polymorphism diversity between multiple genomes of the same ST recovered from the same monkey

Sample ID	STs (colonies per ST)	Pair-wise SNP distances between multiple colonies of the same ST	Comment(s)
PapRG-03	336 (n=5)	0-2	
PapRG-04	1204 (n=2)	4	
PapRG-05	1431 (n=2)	0	
ProbRG-07	73 (n=5)	0-1	
ChlosRG-12	196 (n=2)	25	
PapAN-14	226 (n=3)	1	
PapAN-15	226 (n=2)	1	
ChlosAN-17	681 (n=3)	0-3	
ChlosAN-18	681 (n=4)	0	
ChlosBP-21	677 (n=4)	5	
ChlosBP-23	8527 (n=2)	0	
ChlosBP-24	73 (n=5)	0-5	
ChlosBP-25	73 (n=5)	0-79	Please see Table 2B
PapM-32	212 (n=4)	0	
PapM-33	8533 (n=4)	0-4	
PapM-34	676 (n=4)	0-1	
PapM-36	8535 (n=2)	0-1	
PapKW-44	442 (n=4)	1-2	
ProbK-45	127 (n=5)	0-4	

In individuals where multiple colonies yielded the same genotype (n=19), five had entirely identical genotypes, while we observed a cloud of closely related genetic variants (0-5 SNPs, Table 1) in twelve individuals. However, in two monkeys (highlighted with red boxes), pair-wise SNP comparisons suggested multiple infection events (See Table 2B).

Table 2B: Within-host diversity in green monkey 25 (ChlosBP-25)

Sample ID	Clone designation
ChlosBP-25	
ChlosBP-25-1	1
ChlosBP-25-2	2
ChlosBP-25-3	2
ChlosBP-25-4	2
ChlosBP-25-5	3
Pair-wise SNP distances between clones	
	Clone 1 Clone 2 Clone 3
Clone 1	0 12 79
Clone 2	12 0 67
Clone 3	79 67 0

Table 3: Genomic relationship between study isolates and publicly available *E. coli* genomes

7-allele ST	HC100 subgroups	Non-human primate host	Closest neighbours' source	Neighbours' country of isolation	Allelic distance
349	-	<i>Chlorocebus sabaeus</i> 18	Human (bloodstream infection)	Canada	7
2076	-	<i>Papio papio</i> 14	Environment (water)	Unknown	25
939	-	<i>Papio papio</i> 14	Livestock	US	40
442	-	<i>Papio papio</i> 44	Human	China	50
2800	-	<i>Papio papio</i> 31	Unknown	Vietnam	59
1973	-	<i>Chlorocebus sabaeus</i> 13	Unknown	Unknown	64
8533	-	<i>Papio papio</i> 33	Environment (water)	Unknown	69
6316	-	<i>Papio papio</i> 05	Human	Kenya	97
1727	-	<i>Papio papio</i> 34	Human	Kenya	98
676	-	<i>Papio papio</i> 34	Human (bloodstream infection)	UK	98
8823	-	<i>Papio papio</i> 15	Rodent (guinea pig)	Kenya	101
1431	-	<i>Papio papio</i> 05	Human	US	109
5073	-	<i>Papio papio</i> 15	Human	US	112
226	73641	<i>Papio papio</i> 14	Human	Tanzania	112
8827	-	<i>Papio papio</i> 06	Human	Unknown	122
1204	83197	<i>Papio papio</i> 04	Livestock	Japan	127
1204	83197	<i>Papio papio</i> 04	Livestock	Japan	130
677	-	<i>Chlorocebus sabaeus</i> 21	Human	US	132
40	-	<i>Chlorocebus sabaeus</i> 12	Human	UK	137
1204	83164	<i>Papio papio</i> 06	Livestock	Japan	173
99	-	<i>Papio papio</i> 05	Human	UK	180
362	-	<i>Chlorocebus sabaeus</i> 17	Food	Kenya	180
8825	-	<i>Piliocolobus badius</i> 19	Human	France	189
336	-	<i>Papio papio</i> 03	Poultry	Kenya	189
73	-	<i>Chlorocebus sabaeus</i> 24	Human	Sweden	189
196	-	<i>Chlorocebus sabaeus</i> 12	Human	Sweden	197
2521	-	<i>Papio papio</i> 06	Livestock	US	201
127		<i>Piliocolobus badius</i> 45	Companion animal	US	229
681		<i>ChlosAN</i> 17	Human	Norway	251
38	-	<i>Papio papio</i> 33	human	UK	265
135	-	<i>Papio papio</i> 31	Poultry	US	281
8824	-	<i>Chlorocebus sabaeus</i> 12	Environmental*	US	296
226	100039	<i>Papio papio</i> 14	Human	Sri Lanka	318
8527	-	<i>Chlorocebus sabaeus</i> 23	Human	Kenya	323
8535	-	<i>Papio papio</i> 36	Environmental (soil)	US	368
1665	-	<i>Papio papio</i> 04	Livestock	UK	371
4080	-	<i>Papio papio</i> 06	Human	Denmark	507
8526	-	<i>Chlorocebus sabaeus</i> 13	Livestock	US	708
8532	-	<i>Papio papio</i> 32	Non-human primate	Gambia (PapM-31-3)	1102
8826	-	<i>Papio papio</i> 04	Livestock	Mozambique	1255
8525	-	<i>Papio papio</i> 06	Livestock/companion animal	Switzerland	1659
1873	-	<i>Chlorocebus sabaeus</i> 29	Environment	US	1685
8550	-	<i>Chlorocebus sabaeus</i> 13	Unknown	Unknown	2006

*Source details unknown.

Isolates from humans were recovered from stools, except where indicated otherwise.

Figure legends

Figure 1. Study sites and distribution of study subjects.

Figure 2. Study sample-processing flow diagram.

Figure 3. A plot showing the maximum likelihood phylogeny of the study isolates overlaid with the prevalence of potential virulence genes among the study isolates. The tree was reconstructed based on non-repetitive core SNPs calculated against the *E. coli* K-12 reference strain (NCBI accession: NC_000913.3), using RAxML with 1000 bootstrap replicates. *E. coli* MG1655 was used as the reference and *E. fergusonii* as the outroot species. Recombinant regions were removed using Gubbins (Reference 38). The tip labels indicate the sample IDs, with the respective in silico Achtman sequence types (STs) and HC1100 (cgST complexes) are indicated next to the tip labels. Both the sample IDs and the STs (Achtman) are colour-coded to indicate the various phylogroups as indicated. Novel STs (Achtman) are indicated by an asterisk (*). *Escherichia fergusonii* and the *E. coli* reference genomes representing the major *E. coli* phylogroups are in black. Primate species are indicated in the strain names as follows: *Chlorocebus sabaeus*, ‘Chlos’; *Papio papio*, ‘Pap’; *Piliocolobus badius*, ‘Prob’. The sampling sites are indicated as follows: BP, Bijilo forest park; KW, Kiang-West National park; RG, River Gambia National Park; M, Makasutu Cultural forest; AN, Abuko Nature reserve; K, Kartong village. Co-colonising seven-allele (Achtman) sequence types (STs) in single individuals are shown by the prefix of the strain names depicting the colony as 1, 2 up to 5. We do not show multiple colonies of the same Achtman ST recovered from a single individual. In such cases, only one representative is shown. Virulence genes are grouped according to their function, with genes encoding the colibactin genotoxin highlighted with a red box. The full names of virulence factors are provided in Supplementary file 5.

Figure 4: A bar graph comparing the prevalence of antimicrobial resistance genotypes in *E. coli* isolated from humans in the Gambia (n=128) as found in EnteroBase (Reference 41) to that found among the study isolates (n=101). The antimicrobial resistance genes detected were as follows: Aminoglycoside: *aph(6)*-Id, *ant aac(3)*-IIa, *ant(3'')*-Ia, *aph(3'')*-Ib, *aadA1*, *aadA2*; Beta-lactamase: *blaOXA-1*, *blaTEM-1B*, *blaTEM-1B*, *blaTEM-1C*, *blaSHV-1*; Trimethoprim: *dfrA*; Sulphonamide: *sul1*, *sul2*; Tetracycline: *tet(A)*, *tet(B)*, *tet(34)*, *tet(D)*; Macrolide, *mph(A)*; Chloramphenicol, *catA1*. Screening of resistance genes was carried out using ARIBA ResFinder (Reference 44) and confirmed by ABRicate (<https://github.com/tseemann/abrigate>). A percentage identity of $\geq 90\%$ and coverage of $\geq 70\%$ of the respective gene length were taken as a positive result.

Supplementary Figure 1. A Ninja neighbour-joining tree showing the phylogenetic relationship between Achtman ST442 strains from this study and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex). The locations of the isolates are displayed, with the genome count displayed in parenthesis. Branch lengths display the allelic distances separating genomes. Gambian strains are highlighted in red. The sub-tree (B) shows the closest relatives to the study strains, with the allelic distance separating them displayed with the arrow. Dotted lines represent long branches which have been shortened.

Supplementary Figure 2. A Ninja neighbour-joining tree showing the phylogenetic relationship between the ST349 (Achtman) strain from this study and all other publicly available genomes within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, with genome counts displayed in parenthesis. Gambian strains are

highlighted in red. The study ST349 strain is separated from a clinical ST349 strain by only seven alleles (<7 SNPs), as depicted in the subtree (B). Long branches are shortened (indicated by dashes).

Supplementary Figure 3. A phylogenetic neighbour-joining tree reconstructed with the study ST939 (Achtman) strain and all publicly available genomes that fell within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, with red highlights around the nodes indicating the Gambian strains. The allelic distance between the study strain and its nearest relative, a bovine ST939 strain, has been given, depicted by the arrow. Dotted lines indicate shortened long branches.

Supplementary Figure 4. A Ninja neighbour-joining tree reconstructed with Achtman ST73 colibactin⁺ strains from this study and all other publicly available ST73 (Achtman) strains that fell within the same HC1100 cluster (cgST complex) in EnteroBase (Reference 41). The sources of the isolates are displayed, with Gambian strains highlighted in red. The Gambian non-human primate strains are on separate long branches, although nested within clades populated by human strains from other countries, suggestive of probably an ancient transmission between the two hosts. The branch lengths for the Gambian strains are displayed. Dotted lines represent long branches which have been shortened.

Supplementary Figure 5. A Ninja neighbour-joining tree showing the phylogenetic relationship between ST127 strains from this study and other publicly available strains that occur within the same HC1100 cluster (cgST complex). The sources of the isolates are displayed in the legends, with Gambian strains highlighted in red. Branch lengths display the allelic distances separating genomes. The sub-tree (B) shows the closest relatives to the study

strains, with the allelic distances separating them displayed with the arrow. Dotted lines represent long branches which have been shortened. Dotted lines represent long branches which have been shortened.

Supplementary Figure 6. A Ninja neighbour-joining tree showing the phylogenetic relationship between ST681 strains from this study and other publicly available strains that fell within the same HC1100 cluster (cgST complex). The study strains fell into two separate HC100 clusters, which are depicted in the two subtrees (B and C). The closest neighbours to both HC100 clusters are displayed, with the branch labels indicating the allelic distances between strains. The locations of the isolates are displayed for each tree, with Gambian strains highlighted in red. Dotted lines represent long branches which have been shortened.

Supplementary Figure 7. A phylogenetic tree showing the phylogenetic relationship between ST2076 strain (an MDR strain) and all other publicly available genomes that fell within the same HC1100 cluster (cgST complex). The legend shows the locations of the isolates, Gambian strains are highlighted in red. The subtree (B) shows the allelic distance between the study strain and its nearest relative, an ST2076 isolate recovered from water. Dotted lines indicate shortened long branches.

Supplementary Figure 8. A Ninja phylogenetic tree showing the closest neighbours of simian ST1873 strain—an environmental (soil) isolate belonging to ST83, separated from the study strain by 1659 alleles. The legends of both the main tree and the subtree show the locations of the isolates Gambian strains are highlighted in red. In the subtree (B), the closest neighbour to the simian ST1873 strain is also highlighted in red. Dotted lines are used to indicate shortened long branches.

Supplementary Figure 9. Ninja phylogenetic trees showing the closest neighbours to simian isolates belonging to novel sequence types (Achtman) ST8550 (A), ST8532 (B) and ST8525 (C), ST8826 (D). The allelic distances between these study isolates and their closest neighbours are >1100 alleles, and the closest neighbours belong to seven-allele STs which share less than five out of the seven MLST loci. Each genome (ST8550, ST8532, ST8525) belongs to a unique cgST complex (novel groups at HierCC 1100), indicative of novel diversity within the non-human primate niche.

Supplementary files

Supplementary File 1. A. Characteristics of the study population, displaying the primate species, their age and gender, and the *E. coli* sequence types (Achtman MLST STs) and phylotypes recovered from individual samples. Novel STs are designated by an asterisk (*).
B. Reference strains that were included in this study.

Supplementary File 2. A. Predicted plasmids from short-read sequences, using ARIBA PlasmidFinder (Reference 44).
B. A table indicating the virulence and (or) resistance genes located on representative plasmids that were sequenced by Oxford nanopore technology. The size of each plasmid and the functions of the respective genes encoded thereon are also indicated.

Supplementary File 3. A. A summary of the sequencing statistics of the novel sequence types derived from this study. **B.** Prophage types detected from long-read sequences using PHASTER (reference 51).

Supplementary File 4. A summary of the sequencing statistics of the study isolates.

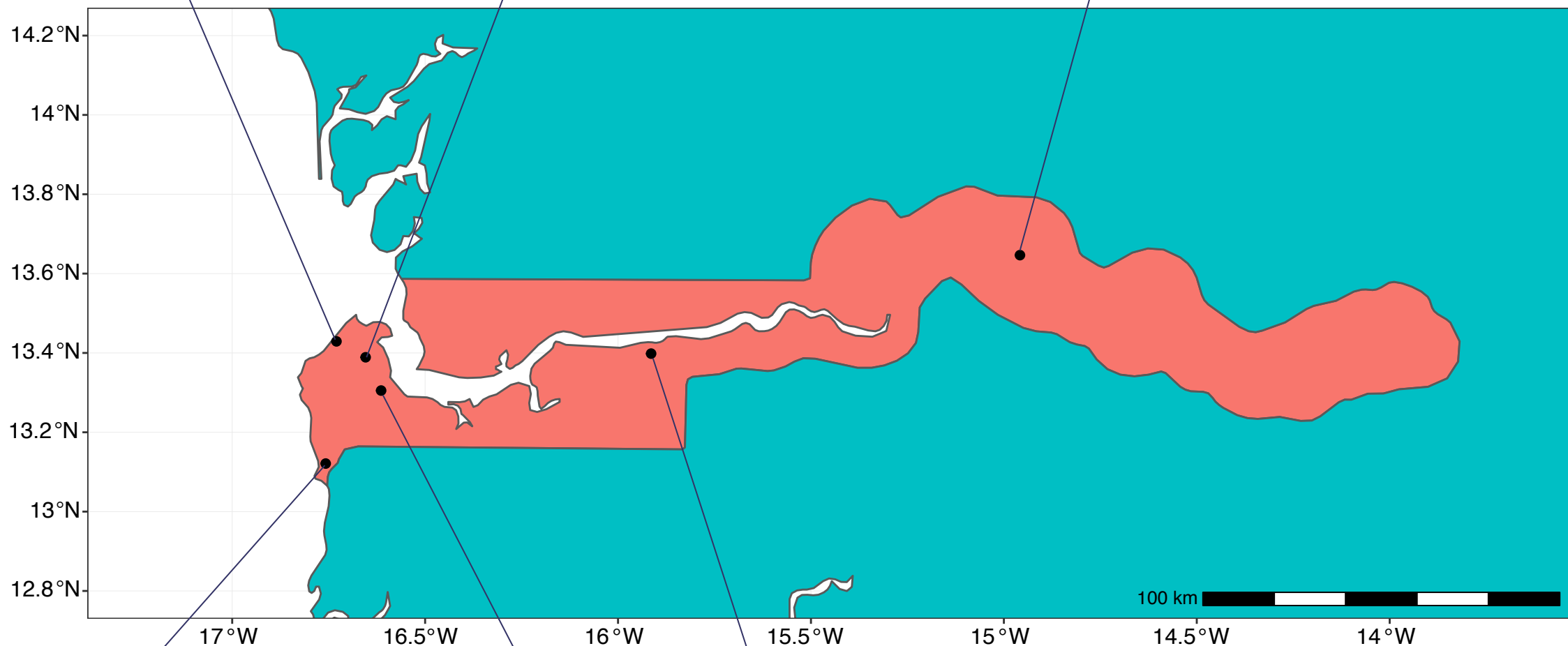
Supplementary File 5. List of virulence factors detected using ARIBA VFDB (Reference 44).

Supplementary File 6. Pair-wise single nucleotide polymorphism distances calculated from the core genome alignment using snp-dists v0.6 (<https://github.com/tseemann/snp-dists>).

Bijilo National Park
Chlorocebus sabaesus (5)

Abuko Nature Park
Papio papio (2)
Chlorocebus sabaesus (5)
Piliocolobus badius (1)

River Gambia National Park
Papio papio (6)
Piliocolobus badius (4)
Erythrocebus patas (1)
Chlorocebus sabaesus (1)



Kartong village
Piliocolobus badius (1)

Makasutu Cultural Forest
Chlorocebus sabaesus (3)
Papio papio (7)

Kiang West National Park
Papio papio (7)

