

Developing an Antimicrobial Strategy for Sepsis in Malawi

-

Thesis submitted in accordance with the requirements of the Liverpool School of Tropical Medicine for the degree of Doctor in Philosophy by Joseph Michael Lewis

August 2019

Contents

Preface	9
1 Introduction	11
1.1 Chapter Overview	13
1.2 Sepsis in sub-Saharan Africa	13
1.3 ESBL-E in sub-Saharan Africa	13
1.4 Conclusions	13
1.5 Thesis overview	13
1.6 Appendix	13
1.7 References	13
2 Methods	15
2.1 Chapter Overview	17
2.2 Study site	17
2.3 Clinical Study	17
2.4 Diagnostic Laboratory Procedures	17
2.5 Molecular methods	17
2.6 Bioinformatics	17
2.7 Statistical Analysis	17
2.8 Study Team	17
2.9 Data Collection and Storage	17
2.10 Ethical Approval, Consent and Participant Remuneration	17
3 <i>Mycobacterium tuberculosis</i> BSI: an IPD meta analysis	19
4 Sepsis in Blantyre, Malawi	21
4.1 Chapter overview	21
4.2 Methods	21
4.3 Study population	21
4.4 Aetiology	21
4.5 Treatment	21
4.6 Outcome	21
5 Early response to resuscitation in sepsis	23
6 Gut mucosal carriage of ESBL-E in Blantyre, Malawi	25
7 Whole genome sequencing of ESBL <i>E. coli</i> carriage isolates	27
7.1 Chapter overview	27
7.2 Methods	28
7.3 Results	30
7.4 Appendix	35

References**37**

List of Tables

List of Figures

7.1	Kraken/bracken read assignment	31
7.2	N50 as a function of total assembly length for included assemblies	32
7.3	<i>E. coli</i> sequence type distribution	33
7.4	ML phylogenetic tree of study <i>E. coli</i>	34
7.5	ML phylogenetic tree of study <i>E. coli</i> in global context	36

Preface

Placeholder

Chapter 1

Introduction

Placeholder

1.1 Chapter Overview

1.2 Sepsis in sub-Saharan Africa

1.2.1 Search strategy

1.2.2 Defining sepsis

1.2.3 Applicability of sepsis-3 definitions in sub-Saharan Africa

1.2.4 Sepsis epidemiology in sub-Saharan Africa

1.2.4.1 Incidence

1.2.4.2 Risk factors: the sepsis population in sub-Saharan Africa

1.2.4.3 Outcomes

1.2.5 Sepsis aetiology in sub-Saharan Africa

1.2.5.1 Bacterial zoonoses, Rickettsioses and arboviruses

1.2.5.2 HIV opportunistic infections: PCP, histoplasmosis and cryptococcal disease

1.2.6 Sepsis management

1.2.6.1 Early goal directed therapy

1.2.6.2 Evidence to guide antimicrobial therapy in sSA

1.2.6.3 Evidence to guide intravenous fluid therapy in sub-Saharan Africa

1.3 ESBL-E in sub-Saharan Africa

1.3.1 Search strategy

1.3.2 Introduction: definition and classification of ESBL-E

1.3.3 Global molecular epidemiology of ESBL-E: an overview

1.3.3.1 1980s-1990s: First identification of ESBL in nosocomial pathogens

1.3.3.2 1990s-2010s: Emergence and globalisation of CTX-M

1.3.3.3 Epidemiology of gut mucosal carriage of ESBL-E: the first step towards invasive infection

1.3.3.4 Molecular mechanisms underlying success of CTX-M: mobile genetic elements and high-risk clones

1.3.4 Epidemiology of ESBL-E in sub-Saharan Africa

1.3.4.1 Invasive ESBL-E infection

1.3.4.2 Gut mucosal carriage of ESBL-E in sub-Saharan Africa

Chapter 2

Methods

Placeholder

2.1 Chapter Overview

2.2 Study site

2.2.1 Malawi

2.2.2 Queen Elizabeth Central Hospital

2.2.3 Participating Laboratories

2.2.3.1 Malawi-Liverpool-Wellcome Clinical Research Programme

2.2.3.2 Malawi College of Medicine Tuberculosis Laboratory

2.2.3.3 Wellcome Trust Sanger Institute

2.3 Clinical Study

2.3.1 Entry Criteria

2.3.2 Study Visits and Patient Sampling

2.3.2.1 Enrollment assessment and first six hours

2.3.2.2 Subsequent visits

2.3.2.3 Blood, urine, and stool, sputum and CSF collection

2.3.2.4 Imaging: chest x-ray and ultrasound scanning

2.3.3 Outcomes and sample size calculations

2.4 Diagnostic Laboratory Procedures

2.4.1 Point of care diagnostics

2.4.2 Laboratory diagnostics

2.4.2.1 Haematology and biochemistry

2.4.2.2 Aerobic blood and CSF culture

2.4.2.3 Mycobacterial blood culture

2.4.2.4 Sputum Xpert

2.4.2.5 Urinary LAM

2.4.2.6 Selective stool culture for ESBL-E

2.4.2.7 Acute and convalescent serologies

2.5 Molecular methods

Chapter 3

Mycobacterium tuberculosis BSI: an IPD meta analysis

Chapter 4

Sepsis in Blantyre, Malawi

Placeholder

4.1 Chapter overview

4.2 Methods

4.3 Study population

4.4 Aetiology

4.5 Treatment

4.6 Outcome

Chapter 5

Early response to resuscitation in sepsis

Chapter 6

Gut mucosal carriage of ESBL-E in Blantyre, Malawi

Chapter 7

Whole genome sequencing of ESBL *E. coli* carriage isolates

7.1 Chapter overview

This chapter describes the use of whole-genome sequencing (WGS) of ESBL producing *E. coli* to understand the drivers of gut mucosal ESBL-E carriage. I will begin with a description of the genomic landscape of the isolates from this study: starting with simple descriptions of *E. coli* phylogroup and multilocus sequence type (MLST) I will place the isolates from this study in the context of the *E. coli* population, followed by higher-resolution contextualisation using phylogenetics to place isolates from this study in the context of a global *E. coli* collection. I will describe the genetic basis of antimicrobial resistance in these isolates and explore the extent to which AMR genes tend to cluster together beyond what would be expected by chance. Finally, I will attempt to use the resolution offered by WGS to attempt to answer two specific questions: firstly, what is the mechanism of rapid increase in ESBL-E carriage prevalence following hospital admission and antimicrobial exposure we see in this study? Secondly, what is the likely unit of ESBL-E transmission in this study? Are bacteria, or mobile genetic elements (MGE) implicated? And if, MGE, which: plasmids, transposons, integrons - or a combination?

These questions, phrased in this way, seem difficult or impossible to answer given the available WGS data, but by slightly reframing them they become tractable: first, what is the diversity of apparent hospital-acquired ESBL *E. coli* in comparison to apparent community-acquired isolates? Apparent hospital acquisitions could represent true acquisitions of, for example, a hospital-associated clone - but equally they could be an “unmasking” of minority variant *E. coli* in the microbiota, acquired in the community but not detected by culture because of low abundance, until enriched for by antimicrobial exposure. If the diversity of apparently hospital acquired isolates is contained within the diversity of community isolates, this would lend support to this latter hypothesis. The second question - what is the unit of transmission in this system - can be reframed by asking: what is the unit that is most conserved within patients, as compared to between patients? The questions then reduce to a dimensionality reduction problem: in order to address them both, it is necessary to classify either bacteria or MGE into mutually exclusive categories, in order to compare hospital to community isolates, and between-patient to within-patient. I describe the approach I have taken to this below.

7.2 Methods

7.2.1 Bioinformatic pipeline

The basic bioinformatic pipeline used is described in detail in Chapter 2, methods. Briefly, one *E. coli* colony from each patient sample was taken forward for DNA extraction and paired-end short-read whole genome sequencing using Illumina HiSeq X at the Wellcome Sanger Institute. Read quality control was undertaken with Kraken[1] v0.10.6 to assign reads to species and WSI QC pipeline which maps a random 100 Mbases from each sample to a reference and calculates depth of coverage, number of heterogeneous SNPs, GC content and insert size. Samples that contained > 80% non *E. coli*. reads were discarded and *de novo* assembly was undertaken with SPAdes[2] v3.11.0. Assembly statistics were calculated with QUAST[3] v4.6.0 and completeness and contamination of the assemblies assessed by checkM[4] v1.0.7. Contaminated assemblies (with checkM-defined contamination of > 25%) or poor assemblies (with less than 1Mb assembled length) were discarded. Annotation was carried out with Prokka[5] v1.5 with a genus specific database from RefSeq and the Roary v1.007 pan-genome pipeline[6] was used to identify a core genome. A core gene multiple sequence alignment was generated using mafft[7] v7.205, SNP-sites identified using SNP-sites[8] v2.4.1 and the resultant SNP alignment used to build a maximum likelihood phylogenetic tree using IQ-TREE[9] v1.6.3, using ascertainment bias correction to correct for the fact that the input pseudosequence contained only variable sites, and using the ModelFinder module used to find the best fitting nucleotide substitution model. This calculates the likelihood of a number of different models and chooses the model with the lowest (best fitting) Bayesian Information Criterion, a statistic which penalises model parameters. Reliability of inferred branch partitions was assessed with 1000 bootstrap replicates. Trees were visualised in the ggtree v1.14.4 package[10] in R.

Ariba[11] v2.12.1 was used to identify AMR-associated genes using the SRST2[12] database, to identify plasmid replicons using the PlasmidFinder database[13] and to perform *in silico* multi-locus sequence typing (MLST) using the database from <http://mlst.warwick.ac.uk/mlst/dbs/Ecoli> accessed via www.pubmlst.org. Because quinolone resistance often results from SNPs in the chromosome in the quinolone resistance determining regions (QRDRs) of the *gyrA*, *gyrB*, *parE* and *parC* genes - rather than acquisition of whole AMR-determining genes, as is the case with the other genes sought by Ariba - these genes were downloaded from the comprehensive antimicrobial resistance database (CARD, <https://card.mcmaster.ca/>) and Ariba used to call SNPs in them, with default settings. *E. coli* phylogrouping was performed with a quadruplex *in silico* PCR using the Clermont scheme[14] and isPcr v33x2 (<https://github.com/bowhan/kent/tree/master/src/isPcr>)

Rhierbaps package v1.1.0 in R[15] was used to cluster the core genome pseudosequence into sequence clusters (SCs). Two levels were used and these level 2 clusters used to test associations (see statistical analysis, below). To track putative mobile genetic elements ESBL-gene containing contigs were identified using BLASTn[16] v2.7.0 of all contigs against the SRST2 database and then contigs containing any given ESBL gene were grouped by the ESBL gene they contained (for example, all *bla_{ctxm15}* gene-containing clusters were grouped together), and each group clustered using cd-hit[17] v4.6 to produce mutually exclusive ESBL-gene-containing contig clusters for each identified ESBL gene. Henceforth, these clusters will be referred to as ESBL-clusters, for brevity. In order to attempt to determine the biological significance of the identified ESBL-clusters (i.e. what kind of MGE element they are likely to represent), basic statistics were plotted (number of samples contained within each cluster, length of longest contig in cluster in kbases, length distribution of all contigs in cluster relative to longest contig and distribution of sequence identity compared to the longest contig in the cluster). Presence of compound transposons, AMR determinants and plasmid replicons were identified by using BLAST with default settings of each ESBL-cluster representative sequence (as determined by cd-hit i.e one, the longest, for each ESBL-cluster) against the insertion sequence finder (ISfinder) database and the SRST2 database, taking the top hit (as determined by bitscore) for any given location, and visualising the results in gggenes v0.3.2. To assess lineage association, the ESBL-clusters were mapped back to the core genome SNP tree.

7.2.2 Global *E. coli* collection

In order to place the isolates from this study in a global context, published *E. coli* assemblies were downloaded from the WSI servers. These included 149 ESBL-producing *E. coli* from a single centre study in Chachoengsao province, eastern Thailand[18]. In this study, human clinical isolates from standard care in Bhuddhasothorn hospital were selected on the basis of the ESBL phenotype, and environmental samples were collected as part of a cross sectional study and selectively cultured for ESBL-E in 2014-2015. I also downloaded assemblies of 362 enterotoxigenic *E. coli* (ETEC), selected for an ETEC genomic study from the Gothenburg University ETEC collection to represent a broad collection of ETEC isolated worldwide from 1980-2011[19]; 185 atypical enteropathogenic *E. coli* (aEPEC) sequenced for a study of aEPEC and selected from samples from the Global Enteric Multicentre Study (GEMS) in seven centres in Africa and Asia between 2007-2011[20]; and 94 *E. coli* from QECH in Blantyre, Malawi, a combination of invasive (bloodstream and CSF) and carriage isolates, selected for diversity in AMR phenotype from 1996-2014[21]. Details of the year, sample and country of isolation for these samples are given in the appendix to this chapter.

Phylogroup and MLST were determined for these context genomes as described above. AMR genes were identified with Ariba and the SRST2 database, as above, and context genomes were classified as ESBL if they contained any Bush-Jacoby group 2be ESBL gene .

7.2.3 Statistical analysis

In order to explore clustering of AMR genes, the Jaccard index was calculated for a given AMR-gene pair using the Jaccard v0.1.0 package in R. The Jaccard index, a measure of the similarity of two sets of data, is defined as *intersection over union*; in this context, for a given pair of AMR genes x and y , the Jaccard index $J(x, y)$ is the number of isolates that contain both gene x and y divided by the total number that contain either x or y :

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

By definition it lies between 0 (x and y never co-occur) and 1 (x and y always co-occur). Co-occurrence matrices using the Jaccard index were plotted using the heatmap v 1.0.12 package in R. The statistical significance of co-occurrence of genes was assessed by generating 2x2 contingency tables for a given gene pair and p values generated using a Fisher's test with Bonferroni correction; a p value of less than 0.05 was considered statistically significant. Co-occurrence networks of genes occurring commonly together (defined as Jaccard index > 0.5) at a rate greater than expected by chance ($p < 0.05$ following Bonferroni correction) and uncommonly occurring together (defined as Jaccard index < 0.1 and $p < 0.05$ following Bonferroni correction) were plotted using igraph v1.2.2 and ggraph v1.0.2 in R.

To explore hospital or community associations of any given *E. coli* clade, the location of isolation was first plotted against the phylogenetic tree; location of isolation was classified as hospital, community, or recent hospital discharge (defined as a date of isolation within 2 weeks of hospital discharge). This latter category was used because it is possible that a patient could acquire an ESBL-E clone in hospital but only be sampled once leaving hospital; using only hospital isolated and community isolated categories could therefore introduce bias. Hospital or community association of each sequence cluster was assessed using a Fisher's test of proportion of hospital associated samples (defined as sum of hospital isolated and recent hospital discharge) for the given sequence cluster as compared to proportion of hospital associated samples in the remainder of the samples, with a Bonferroni correction for multiple comparisons. $p < 0.05$ was again considered statistically significant.

To compare within-patient to between-patient conservation of bacteria (as represented by core genome alignment and sequence cluster) and ESBL-containing MGE (as represented by the ESBL-clusters) several approaches were taken. Firstly, I assessed whether either sequence cluster or ESBL-cluster were

conserved within an individual at all. I hypothesised that any within-patient correlation is likely to be a function of time: samples closer together in time may be more likely to be similar. To assess if this was the case for bacteria, pairwise core genome pseudosequence SNP distance was calculated using `snp-dists` v0.4 (<https://github.com/tseemann/snp-dists>) for all samples and plotted against the time difference (in days) between samples, within and between patients, and with a smoothed curve fitted using a general additive model with cubic splines. Because of significant overplotting, this was also plotted as a 2D density plot. Based on these plots, the within and between patient SNP distances were compared in two post-hoc defined groups binned by time distance between the samples (50 days or less vs. more than 50 days), and distributions compared with Kruskal-Wallis tests.

I then compared the within patient temporal clustering of ESBL-clusters and sequence clusters, by estimating the proportion of within-patient samples that contain the same ESBL-cluster or sequence cluster, as a function of time; essentially a temporal autocorrelation function. To estimate this, I considered pairwise comparison of all within-patient samples. For any given time between samples, t I defined a window of ± 5 days and estimated the probabilities as the number of all within-patient sample pairs in the window $[t - 5, t + 5]$ that contained the same sequence cluster or ESBL-cluster divided by the total number of all within-patient sample pairs within that time window. Exact binomial confidence intervals for these proportions were generated and probabilities plotted as a function of time. In order to estimate the probability of two samples containing the same sequence cluster or contig-cluster purely by chance, 1000 sample pairs were randomly drawn from all samples with replacement and the proportion of these samples that contained the same sequence cluster or ESBL-cluster calculated.

Finally, to inform the question as to what the likely unit of transmission in this system is, I assessed what was most conserved within patients, in pairwise sample comparison: bacteria (as represented by core gene sequence cluster), ESBL-containing MGE (as represented by ESBL-cluster), or both. Simple proportions in all-against-all pairwise comparison - stratified by whether between-patient or within-patient - were calculated: the proportion of samples that contain the same core gene sequence cluster only, the proportion of samples contain the same ESBL-cluster only, and the proportion that contain both sequence cluster and ESBL-cluster. Proportions were compared between within and between-patient strat in these three groups using Fisher's exact test, with $p < 0.05$ considered statistically significant.

7.3 Results

7.3.1 Samples and quality control

In total, 520 *E. coli* underwent DNA extraction and were shipped from Malawi to WSI; these represented all sequential isolates at the time of final DNA extraction, which occurred in two batches in February 2018 and October 2018. Kracken/Bracken read assignment of these samples is shown in Figure 7.1. The majority of samples have $> 90\%$ of reads assigned to *E. coli*; a minority have $< 90\%$ of reads assigned to *E. coli* but a very closely related species such as *Shigella*, and as such are likely to be pure *E. coli* culture with read misclassification. However, 12 samples have $> 80\%$ reads assigned to a non-*E. coli* species such as *Klebsiella pneumoniae*. These samples were assumed to represent upstream species misidentification or, perhaps more likely, selection of the wrong sample from the freezer archive for culture and DNA extraction, given that for any sample ID there are often several bacterial species identified and cryopreserved. These samples were excluded from further analysis.

Of the remaining 508 samples, there were a median (IQR) of 2339594 (2112842.5-2533930.5) reads, with a median (IQR) depth of coverage (obtained by mapping a random 100Mbases to a reference *E. coli* genome, *Escherichia coli* strain K-12 substrain MG1655, NCBI reference NC_000913.3) of 58 (51-66). One sample had an order of magnitude lower number of reads (291556) with depth of coverage 0; this was assumed to represent sequencing failure and it was excluded from further analysis.

The output from quast and checkM are shown in Figure 7.2, where N50 (the minimum contig length

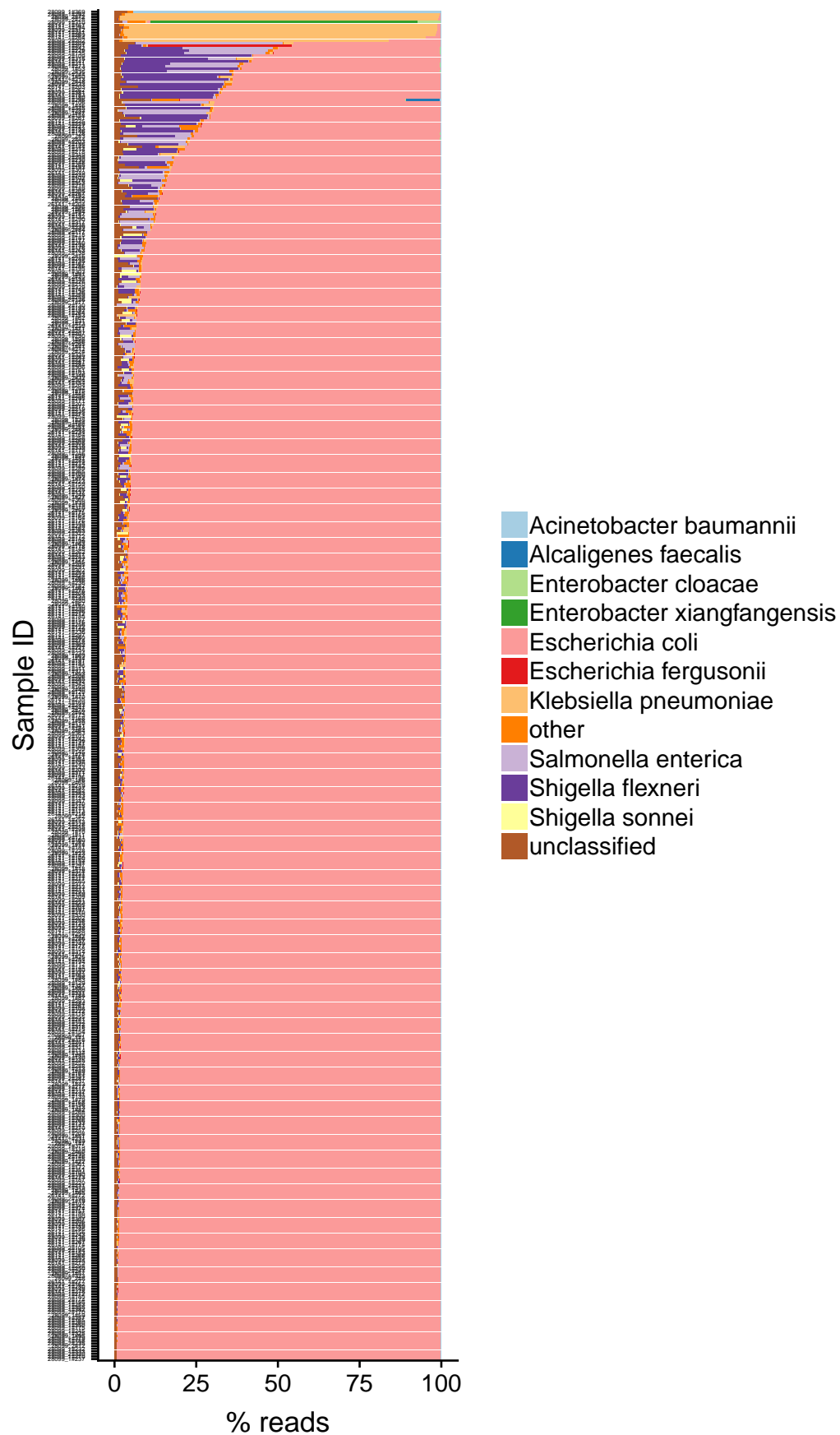


Figure 7.1: Read assignment by Kracken and Bracken of all samples

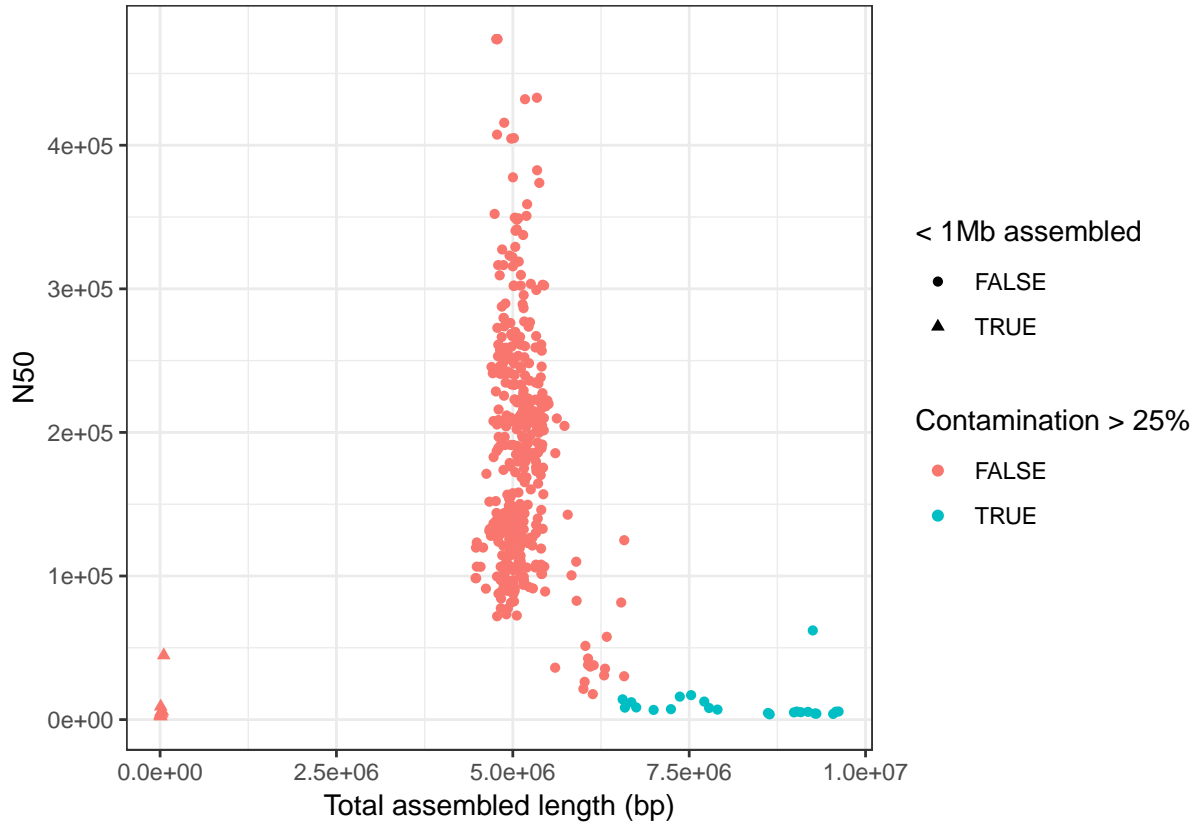


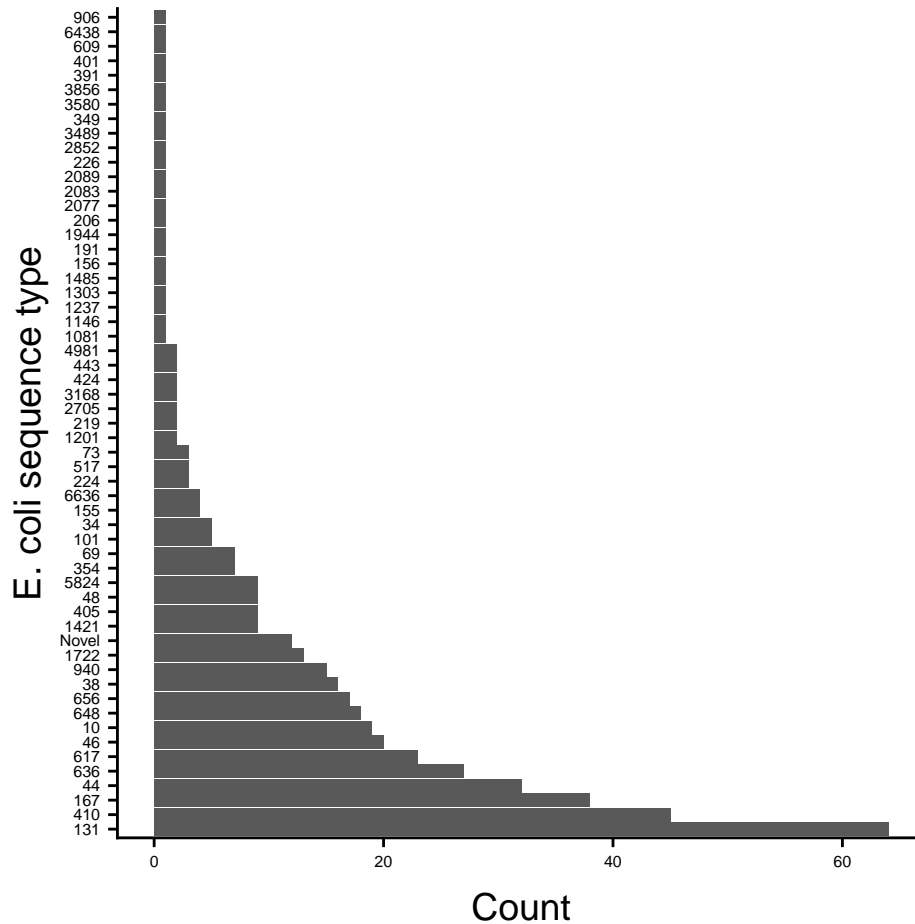
Figure 7.2: N50 as a function of total assembled length. Failed assemblies with less than 1Mb assembled shown as triangles. Contaminated assemblies with checkM-defined contamination above 25% shown in blue.

upon which at least half assembled bases are contained) is plotted as a function of total assembled length. The expected *E. coli* genome length is around 4.6Mb and most samples cluster close to this at a total assembled length of ~ 5Mb. However it is clear that some assemblies have failed, with low N50 and low assembled length. It is also apparent that some samples seem to be contaminated, as indicated by low N50 and much longer than expected total assembled length. Defining assembly failure as < 1Mb assembled length (triangles in the plot, $n = 9$) and contamination as checkM-defined contamination of > 25% (blue points in the plot, $n = 24$) and excluding both groups results in 33 further samples being excluded from further analysis.

In total, therefore, 46/520 (11%) of samples which were submitted for sequencing were excluded from downstream analysis. The remaining 474 samples represent 69% (474/686) of the cultured *E. coli* in this study; 354 are from patients with sepsis, 86 are from hospitalised inpatients and 33 are from community members, with a median of 2 (range 1-5) samples per participant. N50, total assembled length and number of assembled contigs are shown in the appendix to this chapter.

7.3.2 Phylogroup, MLST and core genome phylogeny of study isolates

The commonset *E. coli* phylogroup was phylogroup A: 204/473 (43%) samples belonged to phylogroup A, followed by phylogroup B2 (96/473 [20%]), F (53/473 [11%]), B1 (43/473 [9%]) and C (43/473 [9%]) and D (26/473 [5%]). Two samples were Clade I or II (so called cryptic clades) and 6/473 (1%) were unknown phylogroup using the Clermont PCR scheme. In the MLST analysis, 56 recognised

Figure 7.3: *E. coli* sequence type distribution

sequence types (STs) were identified, and 12 samples were novel STs; however over half (249/473 [53%]) of samples were represented by the top seven most frequent STs (Figure 7.3). ST131 was the most commonly isolated sequence type (64/473 [14%] of isolates) followed by ST410 (45/473 [10%] of isolates) and ST167 (38/473 [8%] of isolates).

The Roary pan-genome pipeline identified a core genome in the study isolates of 2966 genes, with a pan-genome of 26840 genes. The resultant core gene pseudosequence of length 1388742 bases contained 99693 variable sites, which were used to infer the maximum likelihood phylogenetic tree. The IQTREE ModelFinder module determined that a general time reversible (GTR) model with FreeRate site heterogeneity with 5 parameters provided the best fit to the data. The inferred tree is shown in Figure 7.3 along with isolate phylogroup and sequence types; in general, as expected, sequence types were largely monophyletic and phylogroups tended to cluster together.

7.3.3 Study isolates in a global context

The global collection of *E. coli* comprised 1273 samples, including the 473 from this study. 753/1253 (60%) were from Africa, 335/1253 (27%) from Asia and 167 (13%) from South America. The majority of samples, 1026/1253 (82%), were from stool, with 106/1253 (8%) truly invasive samples from blood or CSF and 63/1253 (5%) possibly invasive samples from urine, pus, or sputum. 65/1253 (5%) of samples were environmental, all from Thailand. 670/1253 (53%) of samples contained at least one ESBL-encoding gene. The majority of isolates with ESBL gene (622/670 [92%]) came from this study or

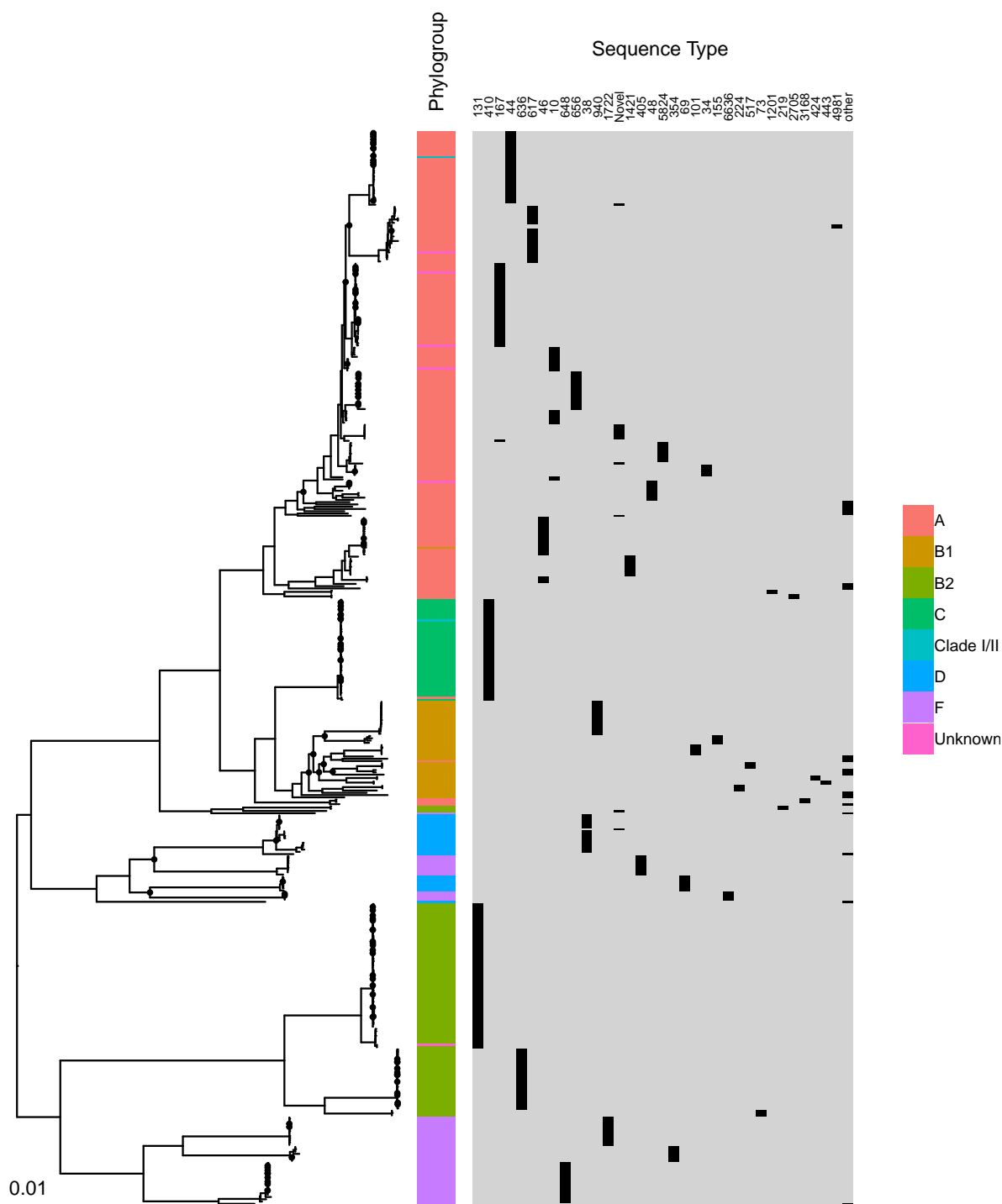


Figure 7.4: Maximum likelihood phylogenetic tree of included study *E. coli* isolates showing phylogroups and sequence types. Bootstrap support of less than 90% is indicated by a black circle at a given node. Scale bar indicates 0.01 SNPs/site.

the Thai ESBL study. Phylogroup A was the commonest phylogroup in the global collection (482/1273 [38%]), followed by B1 (333/1273 [26%]) and B2 (191/1273 [15%]); phylogroup C was uncommon in the global collection (74/1273 [6%]) but the majority of the phylogroup C samples came from this study (43/74 [58%]). All of these 43 phylogroup C isolates belonged to a single ST, ST410; this ST was not seen at all in the previous Malawian study of largely invasive isolates, despite being the second-commonest ST in this study. ST131 was again the commonest ST in the global collection.

The Roary pan-genome pipeline identified 2872 core genes in a pan genome of 44840 genes; this large pan-genome is consistent with the open *E. coli* pan genome that will continue to increase in size as isolates are added. The core gene alignment contained 604817 bases with 77194 variable sites, which were used to infer the maximum likelihood phylogenetic tree, using same nucleotide substitution model as previously.

The inferred tree is shown in Figure 7.5). Isolates from this study are distributed throughout the tree, and there is widespread mixing of isolates from diverse geographic regions. Though invasive isolates are spread throughout the tree, there is a tendency for them to cluster together, particularly in phylogroup B2, a phylogroup which has a recognised association with ExPEC (needs ref). The Malawian ST410 isolates clustered tightly together, though are most closely related to clinical ESBL-producing ST410 isolates from Thailand. By comparison, ST131 isolates from this study were distributed amongst ST131 isolates from other studies, both in Malawi and elsewhere.

7.3.4 Antimicrobial resistance determinants

7.3.5 Testing associations: Sequence clusters and ESBL-clusters

7.4 Appendix

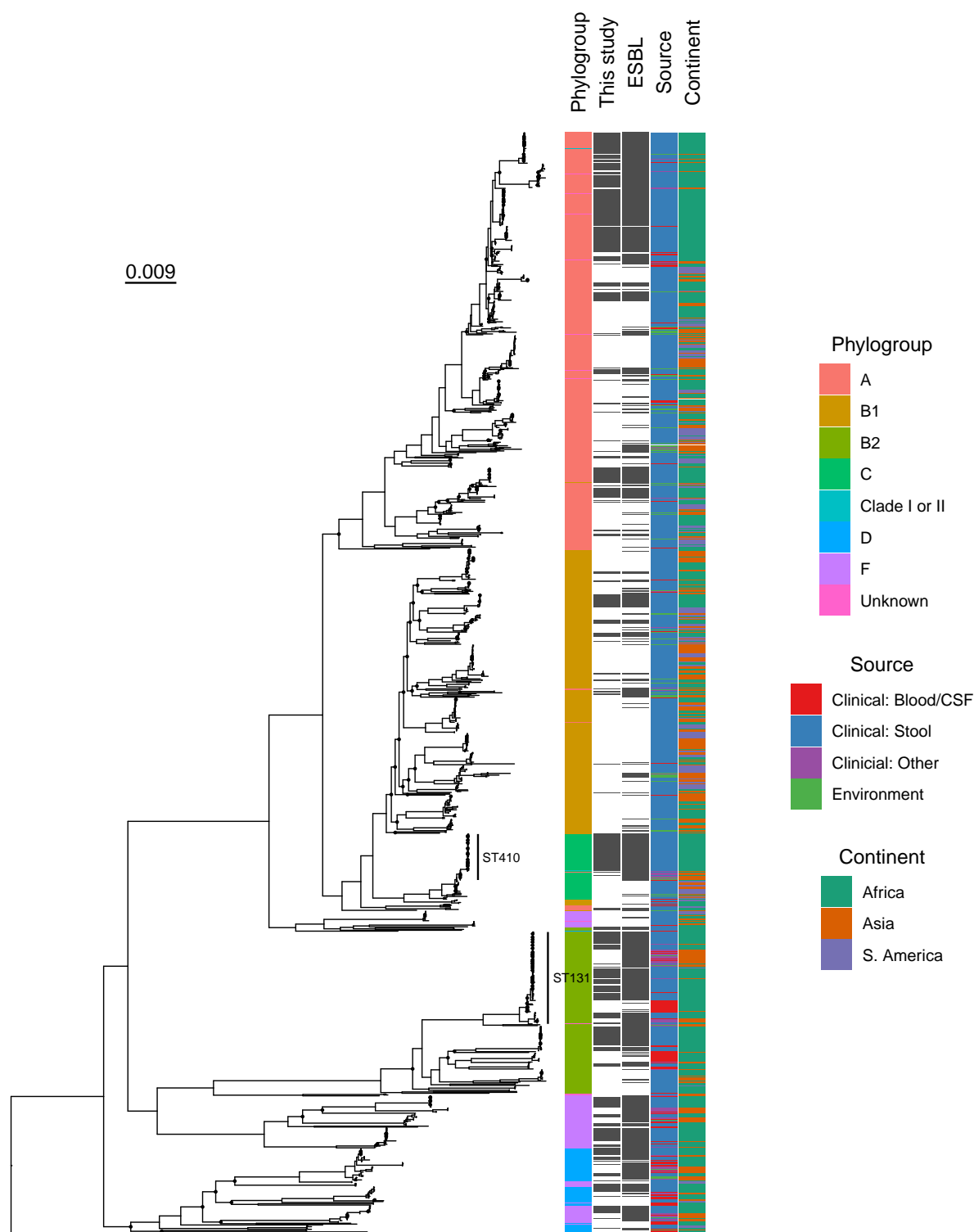


Figure 7.5: Midpoint rooted maximum likelihood phylogenetic tree of included study *E. coli* isolates along with global context isolates, showing phylogroups, source sample type and continent of isolation (coloured bars). Dark grey bars indicate isolates from this study or isolates with ESBL gene presence, as labelled (this study or ESBL, respectively). Two most frequently isolated STs (131 and 410) labelled. Bootstrap support of less than 90% is indicated by a black circle at a given node. Scale bar indicates 0.009 SNPs/site.

References

- 1 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014;**15**:R46. doi:10.1186/gb-2014-15-3-r46
- 2 Bankevich A, Nurk S, Antipov D *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 2012;**19**:455–77. doi:10.1089/cmb.2012.0021
- 3 Gurevich A, Saveliev V, Vyahhi N *et al.* QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**:1072–5. doi:10.1093/bioinformatics/btt086
- 4 Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* 2015;**25**:1043–55. doi:10.1101/gr.186072.114
- 5 Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9. doi:10.1093/bioinformatics/btu153
- 6 Page AJ, Cummins CA, Hunt M *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;**31**:3691–3. doi:10.1093/bioinformatics/btv421
- 7 Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 2013;**30**:772–80. doi:10.1093/molbev/mst010
- 8 Page AJ, Taylor B, Delaney AJ *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* 2016;**2**. doi:10.1099/mgen.0.000056
- 9 Nguyen L-T, Schmidt HA, Haeseler A von *et al.* IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 2015;**32**:268–74. doi:10.1093/molbev/msu300
- 10 Yu G, Smith DK, Zhu H *et al.* Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 2017;**8**:28–36. doi:10.1111/2041-210X.12628
- 11 Hunt M, Mather AE, Sánchez-Busó L *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics* 2017;**3**:e000131. doi:10.1099/mgen.0.000131
- 12 Inouye M, Dashnow H, Raven L-A *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine* 2014;**6**:90. doi:10.1186/s13073-014-0090-6
- 13 Carattoli A, Zankari E, García-Fernández A *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy* 2014;**58**:3895–903. doi:10.1128/AAC.02412-14
- 14 Clermont O, Christenson JK, Denamur E *et al.* The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environmental Microbiology Reports* 2013;**5**:58–65. doi:10.1111/1758-2229.12019

- 15 Cheng L, Connor TR, Siren J *et al.* Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Molecular Biology and Evolution* 2013;**30**:1224–8. doi:10.1093/molbev/mst028
- 16 Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *Journal of Molecular Biology* 1990;**215**:403–10. doi:10.1016/S0022-2836(05)80360-2
- 17 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. doi:10.1093/bioinformatics/btl158
- 18 Runcharoen C, Raven KE, Reuter S *et al.* Whole genome sequencing of ESBL-producing *Escherichia coli* isolated from patients, farm waste and canals in Thailand. *Genome Medicine* 2017;**9**:81. doi:10.1186/s13073-017-0471-8
- 19 Mentzer A von, Connor TR, Wieler LH *et al.* Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nature Genetics* 2014;**46**:1321–6. doi:10.1038/ng.3145
- 20 Ingle DJ, Tauschek M, Edwards DJ *et al.* Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants. *Nature Microbiology* 2016;**1**:15010. doi:10.1038/nmicrobiol.2015.10
- 21 Musicha P, Feasey NA, Cain AK *et al.* Genomic landscape of extended-spectrum beta-lactamase resistance in *Escherichia coli* from an urban African setting. *J Antimicrob Chemother* 2017;**72**:1602–9. doi:10.1093/jac/dkx058