Genomic analysis of extended-spectrum beta-lactamase (ESBL) producing *Escherichia coli* colonising adults in Blantyre, Malawi reveals previously undescribed diversity.

Joseph M. Lewis[1,2,3,4], Madalitso Mphasa[1], Rachel Banda[1], Mathew A. Beale[4], Jane Mallewa[5], Adam P. Roberts[2], Eva Heinz[2], Nicholas R. Thomson[4,6], Nicholas A Feasey[1,2]

1 Malawi-Liverpool Wellcome Clinical Research Programme

2 Liverpool School of Tropical Medicine

3 University of Liverpool

4 Wellcome Sanger Institute

5 College of Medicine, University of Malawi

6 London School of Hygiene and Tropical Medicine

Corresponding Author: Joseph M. Lewis

Department of Clinical Infection, Microbiology and Immunology

University of Liverpool

8 West Derby Street, Liverpool, L69 7BE

United Kingdom

jmlewis@liverpool.ac.uk

+44 (0)151 795 9687

Keywords: Whole-genome sequencing; Africa south of the Sahara; Drug resistance, microbial

Repositories: All data and code to replicate this analysis is available as the *blantyreESBL* v1.0.0 R package (https://doi.org/10.5281/zenodo.5554082) available at https://github.com/joelewis101/blantyreESBL. Reads from all isolates sequenced as part of this study have been deposited in the European Nucleotide Archive, and accession numbers (as well as accession numbers of publicly available genomes used in this analysis) are provided in the R package.

Abstract

Escherichia coli is a ubiquitous bacterial species, associated with drug resistant infections; hundreds of thousands of genomes are now available, but are biased towards high-income countries and clinical isolates. Data from sub-Saharan Africa (sSA) are underrepresented in global sequencing efforts and may represent a major source of genetic diversity with respect to transmissible antimicrobial resistance (AMR). We carried out a genomic investigation of extended-spectrum beta-lactamase (ESBL)-producing *E. coli* colonising adults in Blantyre, Malawi to assess the diversity and AMR determinants and to place these isolates in the context of globally available genomes. We carried out short-read whole-genome sequencing of 473 colonising ESBL *E. coli* isolated from stool and placed them in the context of a previous curated species wide collection of 10,146 isolates using the popPUNK clustering algorithm and by constructing a core gene phylogeny. The most frequently identified STs in Malawian isolates were the globally successful ST131 and ST410, and $bla_{\text{CTX-M}}$ were the dominant ESBL genes, mirroring global trends. However, 37% of Malawian isolates did not cluster with any isolates in the global collection, and the core gene phylogeny was consistent with local subclades including in ST410 and several phylogroup A lineages. Apparent undescribed diversity in Malawian *E. coli* could be due to local selection pressures or sampling biases in global *E. coli* collections. Taking a one health approach to further sampling of *E. coli* from Malawi and sSA, and principled incorporation into unbiased global collections is necessary to understand local, regional and global transmission of both *E. coli* and priority AMR genes.

Data Summary

All data and code to replicate this analysis is available as the *blantyreESBL* v1.0.0 R package (https://doi.org/10.5281/zenodo.5554082) available at https://github.com/joelewis101/blantyreESBL. Reads from all isolates sequenced as part of this study have been deposited in the European Nucleotide Archive, and accession numbers (as well as accession numbers of publicly available genomes used in this analysis) are provided in the R package.

Introduction

Escherichia coli is a ubiquitous bacterium; a human gut commensal and common human pathogen. Beta-lactams (including third-generation cephalosporins, 3GC) are a widely used antimicrobial class worldwide for treatment of Gram-negative infections like *E. coli* but are largely rendered ineffective if the bacteria express extended-spectrum beta lactamase (ESBL) enzymes. ESBL-producing strains have disseminated globally, leaving carbapenems, in many cases as the only widely tolerated treatment option[1,2]. These agents are now equally under threat given the increasing spread of strains producing carbapenem-inactivating carbapenemase enzymes and *E. coli* producing extended-spectrum beta lactamase (ESBL) and carbapenemase enzymes have been identified as priority pathogens by the World Health Organisation[3]. Global genomic surveillance has provided insight into the mechanisms and epidemiology of the global spread of ESBL and carbapenemase producing *E. coli*, suggesting that capture of virulence and AMR determinants via horizontal gene transfer by so-called high risk clones results in fitness advantages and subsequent global dissemination[4]. This phenomenon is well described in *E. coli* sequence type (ST) 131[5], associated with the ESBL-encoding gene $bla_{\text{CTX-M-15}}$, but has also been recently described in other, carbapenemase-associated, *E. coli* lineages such as ST167[6] and ST410[7].

However, *E. coli* sequencing efforts are thus far biased towards collections from high-income settings[8]; the genomics of AMR in low- and middle- income settings like many of the countries of sub-Saharan Africa, where epidemiology of infection and antimicrobial pressures likely differ, are poorly described. Our study is set in Blantyre, Malawi, where carbapenem use is not routine, but the 3GC antimicrobial

ceftriaxone has been widely used in the hospital setting since its introduction to the Malawian national formulary in 2005[9]. Since that time, ESBL- producing *E. coli* have become an increasing problem in clinical practice and now represent 31% of invasive *E. coli* in Blantyre[10], whereas carbapenem resistance has so far only been sporadically described[11]. There is a significant unmet need for access to carbapenem antimicrobials to treat resistant infections, but the example of ceftriaxone shows that carbapenem resistance may be likely to disseminate rapidly if carbapenem use is increased. In this context, both robust stewardship protocols and ongoing genomic AMR surveillance are critical, as well as understanding the links between carriage and disease as the transmission routes of invasive infections are still poorly understood.

To that end, we present insights into the genomic diversity of ESBL-producing *E. coli* from a study of gut mucosal colonisation with ESBL Enterobacterales in Blantyre, Malawi, and describe the diversity and AMR determinants of ESBL *E. coli* including a comparison to large public datasets to understand the diversity of colonising E. coli in our setting and assess how representative mainly high-income country (HIC)-focused collections are of low- and middle-income- country (LMIC) settings.

Methods

The isolates analysed in this study were selectively cultured from stool and rectal swabs collected from adults in Blantyre, Malawi, as part of a study of longitudinal carriage of ESBL-producing Enterobacterales, as previously described[12]. Briefly, three groups of adults (≥ 16 years) were recruited: i) 225 adults with sepsis in the emergency department of Queen Elizabeth Central Hospital (QECH), Blantyre, Malawi; ii) 100 antimicrobial-unexposed adult inpatients; and iii) 100 community

dwelling adults with no antimicrobial exposure (except for long-term co-trimoxazole preventative therapy, CPT, or antituberculous chemotherapy) in the previous four weeks. Up to five stool samples (or rectal swab samples performed by trained study team members if participants were unable to provide stool) were collected over the course of six months and aerobically cultured overnight at 37ºC on ChromAGAR ESBL-selective chromogenic media (ChromAGAR, France) before being speciated with the API system (Biomeriuex, France).

A subsample of isolates identified as *E. coli* underwent DNA extraction and sequencing: one *E. coli* colony pick from the first 507 samples where *E. coli* was identified. DNA was extracted from overnight nutrient broth cultures using the Qiagen DNA mini kit (Qiagen, Germany) as per the manufacturer's instructions. DNA was sequenced at the Wellcome Sanger Institute on the Illumina HiSeq X10 instrument (Illumina Inc., United States). Species was confirmed with Kraken v0.10.6 and Braken v1.0[13]. We first reconstructed a core gene phylogeny for the study isolates: *de novo* assembly was undertaken with SPAdes v3.14.0[14] and the pipeline described by Page et al[15] and quality of the assemblies assessed with CheckM v1.1.2[16] and QUAST v5.0.2[17]. Assembly failures with a total assembled length of < 4Mb or assemblies with a CheckM-defined contamination of ≥ 10% were excluded from further analysis. Included assemblies had a median 92 (IQR 68-122) contigs and N50 of 180kbp (IQR 123-234kbp). Assemblies were annotated with Prokka v1.5 with a genus-specific database from RefSeq[18] and the Roary v1.007 pangenome pipeline[19] used to identify core genes with a BLAST threshold of 95% and paralogs not split. Genes present in ≥ 99% samples considered to be core. A pan-genome of 26,840 genes was identified of which 2,966 were core. The core genes were concatenated to a 1,388,742 base pseudosequence; 99,693 variable sites were

identified and extracted with snp-sites v2.4.1[20] and a maximum-likelihood phylogeny inferred from this alignment with IQ-TREE v1.6.3[21] with ascertainment bias correction. The ModelFinder module was used to select the best fitting nucleotide substitution model: the general time reversible model with FreeRate site heterogeneity with 5 parameters, which was fitted with 1000 ultrafast bootstrap replicates.

ARIBA v.2.12.1[22] was used on the reads to identify AMR-associated genes using the SRST2 curated version of the ARG-ANNOT database[23], and was used to call single nucleotide polymorphisms (SNPs) in the quinolone-resistance determining regions (QRDR) *gyrA, gyrB, parC* and *parE*, using the wild-type genes from the *Escherichia coli* K-12 substr. MG1655 (NC_000913.3) as reference. Any QRDR mutation conferring quinolone resistance in *E. coli* in the comprehensive antibiotic resistance database[24] (CARD) was considered to confer quinolone resistance. Beta lactamases were phenotypically classified according to https://ftp.ncbi.nlm.nih.gov/pathogen/betalactamases/Allele.tab. ARIBA was also used to determine *E. coli* multilocus sequence type (ST) as defined by the 7-gene Achtman scheme[25] hosted at pubMLST (https://pubmlst.org/), to identify plasmid replicons using the PlasmidFinder database[26], and to determine pathotype by identifying genes contained in the VirulenceFinder database[27]. Pathotype was assigned based on the criteria in Supplementary Table 1[28]. *E. coli* phylogroups were determined using the Clermont scheme and primers[29] with *in-silico* PCR on assemblies using isPcr v33 (https://github.com/bowhan/kent/tree/master/src/isPcr).

To place the isolates from this study in context of the wider *E. coli* population structure, we used a dataset from a previously described highly curated collection of

E. coli genomes. The collection is based on 10,146 *E. coli* genomes collected in Europe, the Americas, Asia, Africa and Oceana[8] and used several quality control steps to select 500 genomes representative of the largest 50 lineages, providing a curated set of genomes representing a balanced background dataset that we used to bring our samples into context. Here, we first used popPUNK v1.1.5[30] to cluster the assemblies from this study with all 10,146 isolates from the published collection, using the popPUNK database generated in the original publication allowing us to assign genomes from Malawi into the popPUNK groups defined in Horesh et al. We then used the 500 curated, representative assemblies[8] and the Malawi genomes generated in this study to infer a core gene phylogeny. To include clinical isolates from our setting as comparison to our colonisation isolates, we furthermore added 97 genomes from a previous study of *E. coli* at QECH, where archived samples were selected for sequencing to maximise temporal and antimicrobial susceptibility profile diversity[31]. QC, assembly, determination of ST and phylogroup of these samples proceeded as described above. Following QC, 5 genomes from this latter study were excluded, leaving 1,065 in the analysis. The Roary pan-genome pipeline identified 41,025 gene orthologs in this collection, of which 2,699 were core and formed a pseudosequence of 530,659 bases with 53,410 variable sites. These were extracted with snp-sites and used to infer the phylogeny, using IQ-TREE with GTR substitution model with FreeRate site heterogeneity with 5 parameters and ascertainment bias correction, and 1000 ultrafast bootstrap replicates.

To better describe the phylogeny of two dominant STs in our dataset, ST410 and ST167, in greater resolution, we inferred phylogenies for these STs by mapping to a ST-specific reference genome, with global context genomes from Enterobase.[32] We identified all ST140 and all ST167 genomes listed in Enterobase on 1st March 2021,

and downloaded those which we could link to publicly available Illumina short reads and metadata (year and country of isolation). We performed QC with fastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and multiqc v1.8[33] , trimmed raw reads with Trimmomatic v0.39[34], removing adapter sequences and leading or trailing bases with phred score < 4, bases with a mean score < 20 (over a sliding window of 4 bases), and any reads with length below 36 following removal of low quality bases. We mapped the reads to ST specific references from the curated FDA-ARGOS database[35] (GenBank accession CP023870.1 for ST167 and CP023870.1 for ST410) using the snippy v4.6.0[36] pipeline with default settings; mapped assemblies with mean mapped depth below 20x were excluded. Following QC 281 ST167 and 511 ST410 genomes were included. Areas of putative recombination were then masked following their prediction with Gubbins v 3.0.0[37], variable sites extracted (39,802 sites in a 4,711,093 base alignment for ST410 and 42,526 sites in 4,897,877 bases for ST167) using snp-sites and a phylogeny reconstructed with IQ-TREE as above. Presence of AMR genes and plasmid replicons were inferred as above.

All statistical analyses were carried out in R v4.1.1 (R Foundation for Statistical Computing, Vienna, Austria) and trees were visualized using the *ggtree* v2.2.4[38] package. Summary statistics, where presented, are medians and interquartile ranges or proportions unless otherwise stated. The clinical study which provided the isolates for this analysis was approved by the Liverpool School of Tropical Medicine Research Ethics Committee (16-062) and University of Malawi College of Medicine Research Ethics Committee (COMREC P.11/16/2063). Reads from all isolates sequenced in this study have been deposited in the European Nucleotide Archive: accession numbers and associated metadata are provided in Supplementary Data.

All data and code to replicate this analysis are available as the *blantyreESBL* v1.0.0[39] R package available at https://joelewis101.github.io/blantyreESBL/. Reads from all isolates sequenced as part of this study have been deposited in the European Nucleotide Archive, and accession numbers (as well as accession numbers of publicly available genomes used in this analysis) are provided in the R package.

Results

Population structure

Following quality control, 473 *E. coli* genomes sequenced for this study were included in the analysis, 440 from participants enrolled in hospital, and 33 from community members, with a median 2 (IQR 1-5) samples per participant. A full description of study participants and temporal trends has previously been made[12].The most common phylogroup was A (43%), followed by phylogroup B2 (20%), F (11%), B1 (9%), C (9%) and D (5%) with two samples typing as so-called cryptic clades (Clade I or II) and six untyped by the Clermont scheme (Figure 1A). The phylogroup distribution differed between the Malawian isolates and the global collection (Supplementary Figure 1) with a lower proportion of phylogroup A isolates and higher proportion of phylogroup E isolates in the global collection. Fifty-seven recognised STs were identified, with a median 2 (IQR 1-9, range 1-64) samples per ST (Figure 1B) and as expected were largely monophyletic and mapped well to the core gene tree topology (Supplementary Figure 2). The three most frequent STs accounted for 32% of isolates: ST131 was most commonly identified (64/473 [14%] of isolates) followed by ST410 (45/473 [10%]) and ST167 (38/473 [8%]).

We next placed the Malawian carriage isolates in context of the wider species diversity by using a representative collection of key lineages in a curated dataset[8]. Using popPUNK to assign the Malawian isolates to the clusters defined by Horesh et al, the 473 isolates from this study were assigned to 109 clusters of median size 1 (IQR 1-3). The distribution of clusters differed between the isolates from this study and the curated global collection (Figure 2); the biggest 50 popPUNK clusters in the global collection contained 76% of global isolates but only 140/473 (30%) of isolates from this study, and 175/473 (37%) of isolates from this study formed new clusters that were not present in the global collection. The largest two popPUNK clusters in this study were commonly represented in the global collection: lineage 2 (n = 53 isolates from this study, all ST131), a global phylogroup B2, ST131-associated lineage) and lineage 40_708 (n = 44 isolates form this study, all ST410), a phylogroup C, ST410 associated lineage. However, other large clusters in this study had very few representatives in the global collection: the third and fourth largest clusters in this study were lineage 684 (n = 29 in this study, all phylogroup A ST44) and 451 (n = 27 all phylogroup A ST636) had only one isolate each in the global collection, for example (Table 1).

We next reconstructed a contextual core-gene phylogeny using our 473 new Malawi genomes, the 500 representative assemblies from the global collection, and 97 genomes from a previous study at QECH representing clinical isolates from the same setting as our carriage collection. (Figure 3). Malawian isolates clustered with global isolates throughout the tree, suggesting global transmission of *E. coli* strains. This was the pattern seen for ST131 (Figure 3D), the globally successful ExPEC lineage which was the most commonly identified ST in this collection, and where Malawian isolates were closely related to global isolates.

However, in contrast to ST131, the tree topology for the second and third most frequently identified STs in the Malawian collection (ST410, Figure 3B and ST167, Figure 3C) was consistent with a paradigm of locally circulating subclades: in the case of ST410, clonal expansion of a Malawian subclade closely related to global ST410 isolates (Figure 3B) and, in the case of ST167, multiple related phylogroup A Malawi-associated lineages (including ST167, ST617, ST44, ST656 and ST9847, Figure 4C). In fact, some of these lineages (ST167, ST44, ST617) were clustered by popPUNK with isolates from the global collection (Figures 3B-D), but these were not included in the 500 representative assemblies in the core gene tree because they fell outside the largest 50 lineages. Other lineages (ST656 and ST9847) formed novel popPUNK clusters that were not described in the global collection.

To explore the genomic epidemiology of ST410 and ST167 further we reconstructed global ST 410 and 167 phylogenies by mapping each ST dataset to ST-specific reference genomes, incorporating 281 ST167 and 511 ST410 genomes from Enterobase (Supplementary Figures 3 and 4); subtrees for the section containing the Malawian isolates are shown in Figure 4. These too were consistent with local Malawian subclades. The Malawian ST410 (except for a single isolate) fell within the globally distributed carbapenem-associated B4/H24RxC lineage, were monophyletic, and did not cluster with global isolates, but descended from a single common ancestor in the B4/H24RxC lineage with good (>95%) bootstrap support (Figure 4A) Malawian ST167 isolates were comprised of three lineages, one of which was monophyletic with good (>95%) bootstrap support and did not cluster with any global isolates (Figure 4B); the other two clustered closely with isolates from Asia, Europe, and the Americas.

Resistance and virulence determinants

The identified AMR determinants in the isolates sequenced for this study are shown in Figure 5. Only one ST2083 isolate contained a carbapenemase encoding gene: *bla*$_{\text{NDM-5}}$ carried on a 46.2kbp Inc-X3 plasmid. This separated by only 11 pairwise SNPs from the *bla*$_{\text{NDM-5}}$ associated plasmid, pNDM-MGR194 that we have described previously[11]. The remainder (n = 472) carried at least one ESBL-encoding gene, most commonly *bla*$_{\text{CTX-M-15}}$ which was present in 319/473 (67%) of isolates. All other identified ESBL-encoding genes were members of the *bla*$_{\text{CTX-M}}$ family except for *bla*$_{\text{SHV-12}}$ which was identified in 26/473 isolates across 6 STs, but particularly associated with ST656; all 17 ST656 isolates in the collection carried *bla*$_{\text{SHV-12}}$.

Co-occurring determinants of resistance to aminoglycosides (99% [472/473] of isolates), trimethoprim (97% [459/473]), sulphonamides (99% [468/473]) and quinolones were very common (86% [407/473]); whereas genes conferring resistance to chloramphenicol less so (52% [248/473]). QRDR mutations were the most frequently identified quinolone resistance determinants (in *gyrA* in 74% [351/473] and *parC* in 63% [296/473] isolates) but plasmid-mediated quinolone resistance determinants were also found (*qnrS* [12%, 58/473], *qnrB* [1% 6/473] and *qep* [10%, 47/473]). The *gyrA* mutants were S83L (n=351), D87N (n= 294) and *parC* mutants S80I (n=296) and S84G (n=6). There were some lineage associations apparent on mapping the AMR determinants back to the tree (Supplementary Figure 5). Identified plasmid replicons are predominantly IncF (Figure 1C) which is the plasmid type usually found associated with *bla*$_{\text{CTX-M-15}}$ [40] itself particularly associated with ST131 strains.[41] Most (461/473 [97%]) strains lacked the genes to classify them

as any pathotype: 2/473 were identified as aEPEC/EPEC and 10/473 as EAEC (Supplementary Figure 1).

The resistance determinants of the global and Malawian ST410 and 167 isolates are shown mapped to the ST-specific trees in Supplementary Figures 3 and 4 and Figure 5. As expected in these carbapenemase-associated lineages, carbapenem resistance determinants were common in the global isolates – but they were absent in the Malawian isolates. All Malawian ST410 carried the *bla*$_{\text{CTX-M-15}}$ and *bla*$_{\text{CMY-94}}$ genes present in the rest of the global B4/H24RxC lineage but lacked the characteristic *bla*$_{\text{OXA-181}}$ or *bla*$_{\text{NDM-5}}$ carbapenemases and IncX3 plasmid replicon which was present in most of the carbapenemase associated isolates in the lineage.

Discussion

We present here a genomic analysis of the diversity of ESBL-producing *E. coli* in Blantyre, Malawi, both enhancing both the local and global understanding of *E. coli* genomic epidemiology by adding data from an under-sampled location. ESBL-producing *E. coli* colonising adults in Malawi represents the diversity of the species with all major phylogroups, and 57 STs represented. Some global trends in ST are broadly reflected in Blantyre; ST131, for example, the most frequently isolated pathogenic ST in many settings worldwide[42], was the most commonly isolated ST, followed by the globally emergent AMR-associated ST410.[6,7] In Blantyre, as worldwide, *bla*$_{\text{CTX-M}}$ are the dominant ESBL family, especially *bla*$_{\text{CTX-M-15}}$. In our isolates, which were selected based on ESBL production, only one isolate carried a carbapenemases. As observed commonly in collections of ESBL-producing bacteria dependent on mobile elements, aminoglycoside, trimethoprim and sulphonamide resistance determinants were near-universal, and ciprofloxacin and chloramphenicol

resistance determinants common. In terms of plasmids the co-occurrence of *bla*$_{\text{CTX-M-15}}$ and IncF plasmids in our isolates reflects predominant global observations.[40,43]

This AMR-determinant distribution may be influenced by local antimicrobial pressures: carbapenem antimicrobials are at best sporadically available in QECH, but co-trimoxazole as CPT is widely used in this high HIV-prevalence setting as lifelong prophylaxis against infection in people living with HIV, as per World Health Organisation guidelines[44]. The high prevalence of genes conferring resistance to co-trimoxazole in this collection raises the possibility that use of CPT could be creating selection pressure for other AMR-determinants in Blantyre and it may be that consideration of a more nuanced approach to the deployment of CPT is needed in an era of increasing Gram-negative resistance within high HIV-prevalence settings.

ESBL producing Gram-negative infections are an increasing clinical problem in Malawi, and there is a significant unmet need for access to carbapenem antimicrobials to treat them. Carbapenemase producing Enterobacterales including *E. coli* (*bla*$_{\text{NDM-5}}$ in *E. coli* ST636) have recently been described in other regions of Malawi[45] and the presence of carbapenemases in this collection despite minimal local carbapenem use highlights the need for ongoing surveillance as these last-line antimicrobials are introduced. In ST410, the Malawian isolates emerge from a globally distributed *bla*$_{\text{NDM-5}}$/*bla*$_{\text{OXA-181}}$ carbapenemase-associated lineage but lacked any carbapenemase genes; it may be that in the absence of the selection pressure of carbapenems themselves any fitness cost associated with acquisition and maintenance of carbapenemase harbouring mobile genetic elements is too great for them to persist. Availability of carbapenems is predicted to change.

Chloramphenicol resistance determinants were common, but absent in 48% of samples; chloramphenicol has historically been a first-line treatment for severe febrile illness in Malawi[46] but development of resistance has curtailed its use in favour of ceftriaxone[47]. These results show that chloramphenicol could still have a role to play as a reserve agent in the treatment of ESBL-*E. coli*, but this would require quality assured diagnostic microbiology to support it.

By placing the Malawian isolates in a global context we found unsuspected diversity in the Malawian collection not captured in the curated species-wide collection. There was a difference in relative prevalence of lineages as defined by popPUNK between Malawian and global samples, with many Malawian isolates forming clusters distinct from global isolates, and both the core gene and map-to reference phylogenies (using different context genomes) were consistent with local Malawian subclades. It is plausible that differing selection pressures (e.g. antimicrobial use, water, sanitation and hygiene infrastructure) or unique niches in this low-income setting could result in local success of different lineages to high income settings.

However, these findings also likely reflect our sampling strategy; this study selected for ESBL-producing carriage isolates whereas the species-wide collection included studies with a variety of inclusion criteria. Despite this, under sampling of *E. coli* from sSA is likely to introduce the most bias; for example only 246/10,146 genomes in the global *E. coli* database were from the African continent[8]. Scaling up of genomic surveillance in sSA with sampling of human, environmental and animal isolates can redress this imbalance, improve understanding of the global transmission of AMR *E. coli*, and should be an international priority for future studies. Efforts to expand global collections (such as the global collection used here) in an unbiased manner as

further genomes from low and middle income countries are sequenced will also be key. Other limitations include that, in this study, some participants provided multiple samples so it is possible that this introduced bias into the collection. Further, our study was predominantly based at a single urban hospital, over a time period of around only 2 years so may not be generalisable.

In conclusion, we present an analysis of 473 genomes of ESBL-producing *E. coli* colonising adults from Blantyre, Malawi, significantly expanding genomic surveillance of AMR in this low-income setting. We find that the diversity of *E. coli* in Blantyre broadly reflects global diversity, but with a suggestion of local subclades that highlights a need for further targeted sequencing of isolates from Malawi and sSA to understand local, regional, and global *E. coli* transmission. Carbapenem resistance is present in Malawi[11,45], and it likely that increased carbapenem use (driven by ESBL resistance) will select for it. There is a critical need for robust stewardship strategies plus ongoing surveillance, as these agents are introduced.

Author contributions

Conceptualisation: JL, NT, NAF. Methodology: JL, NT, NAF, MAB, EH, JM, APR. Investigation: JL, MM, RB. Formal analysis: JL, NT, NAF, EH, MAB. Writing – original draft preparation; JL. Writing – review and editing: JL, MM, RB, MB, JM, EH, APR, NT, NAF. Supervision: NAF, NT.

Funding information

This work was supported by the Wellcome Trust [Clinical PhD fellowship 109105z/15/a to JL and 206545/Z/17/Z, the core grant to the Malawi-Liverpool-

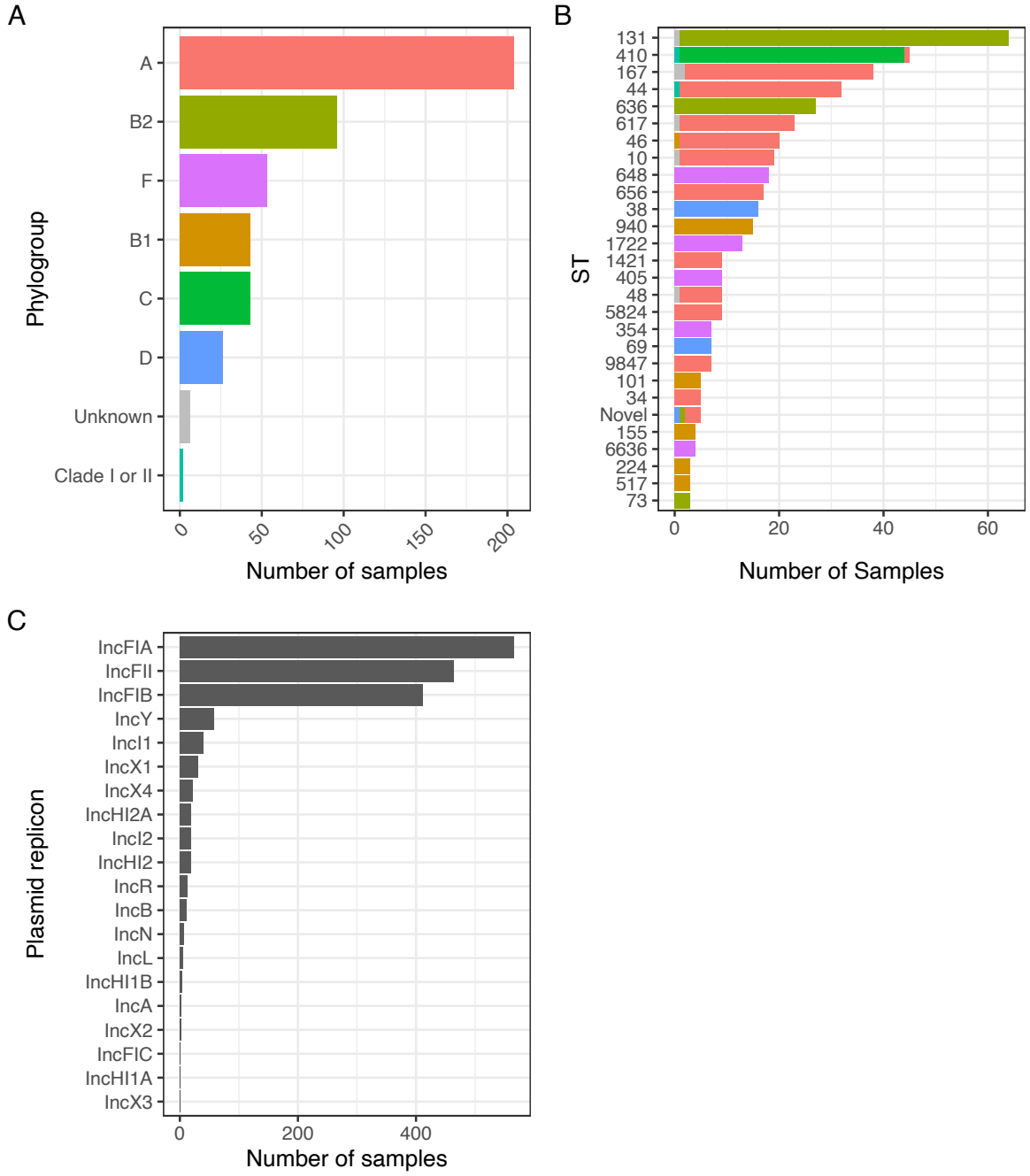
Wellcome Programme]. MAB and NRT are supported by Wellcome funding to the Sanger Institute (#206194).

Acknowledgements

Many thanks to the study team: Lucy Keyala, Tusekile Phiri, Grace Mwaminawa, Witness Mtambo, Gladys Namacha, Monica Matola; to the MLW laboratory teams, particularly Brigitte Denis; and to the MLW data team, particularly Lumbani Makhaza and Clemens Masesa. The authors acknowledge the sequencing team at the Wellcome Sanger Institute, and Christoph Puethe and the Pathogen Informatics team for computational support.

Conflicts of Interest

The authors have no conflicts of interest to declare.



410

Figure 1: ST (A) and phylogroup (B) distribution of included isolates and (C) identified Inc-type plasmid replicons.

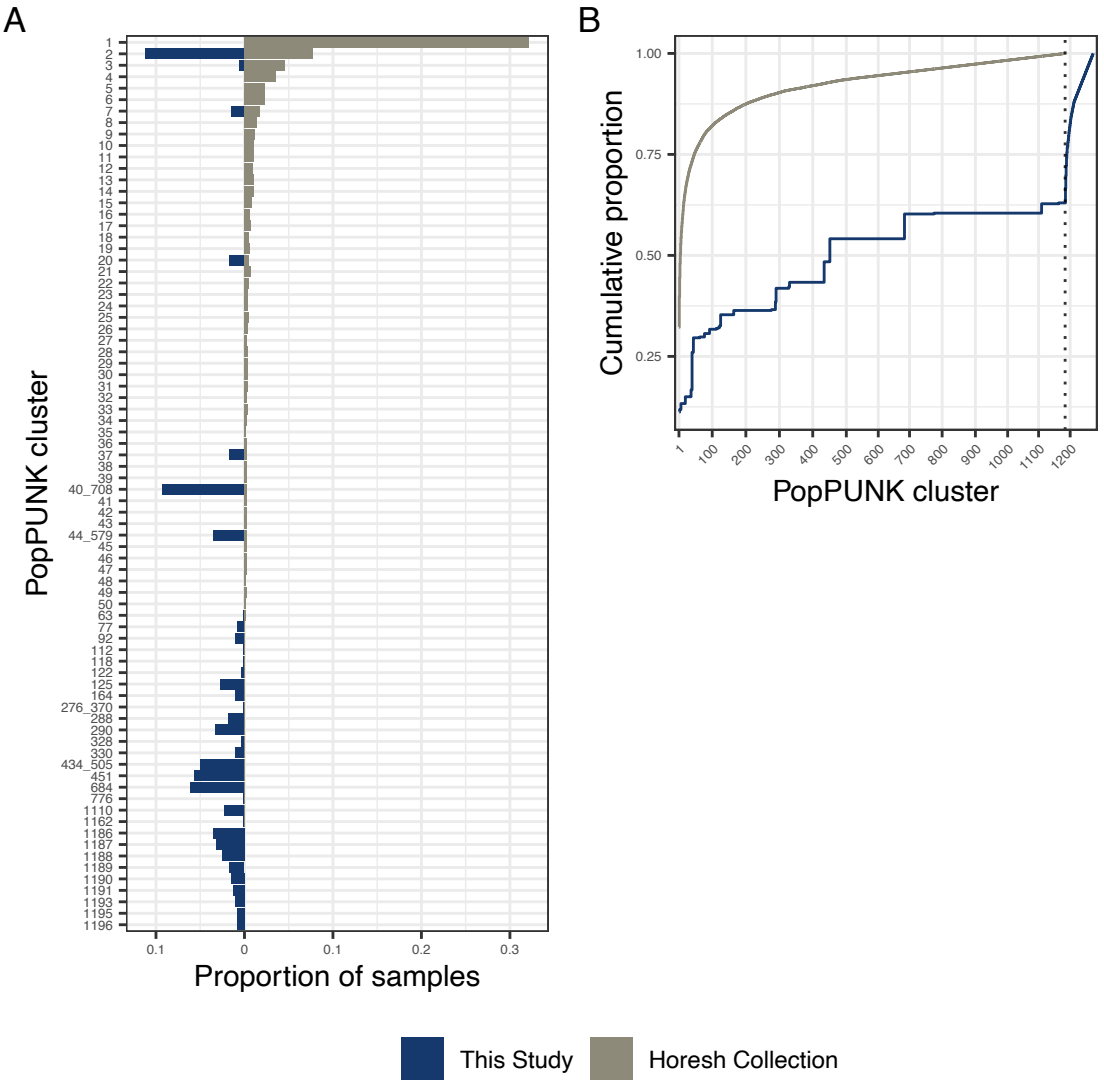


Figure 2: Comparing distribution of popPUNK clusters in Malawian and global collections. A: proportion of samples assigned to a given popPUNK cluster in Malawian (left) and global (right) isolates. Clusters are arranged in numeric order which, by definition, is size order from the original publication from largest to smallest. Clusters 1-50 (accounting for 76% of global isolates) are shown along with any cluster containing at least Malawian isolates. B: Cumulative proportion of samples with given cluster membership, stratified by study; clusters are again numerically ordered on x-axis. Dotted line shows the maximum cluster identifier that was defined in the global collection (n= 1184); clusters with an identifier greater than 1184 (to right of dotted line) were not present in global collection. Clusters with an identifier made up of two numbers separated by an underscore are clusters that were two separate clusters in the original global collection but have been merged after Malawian genomes were added (e.g. 40_708).

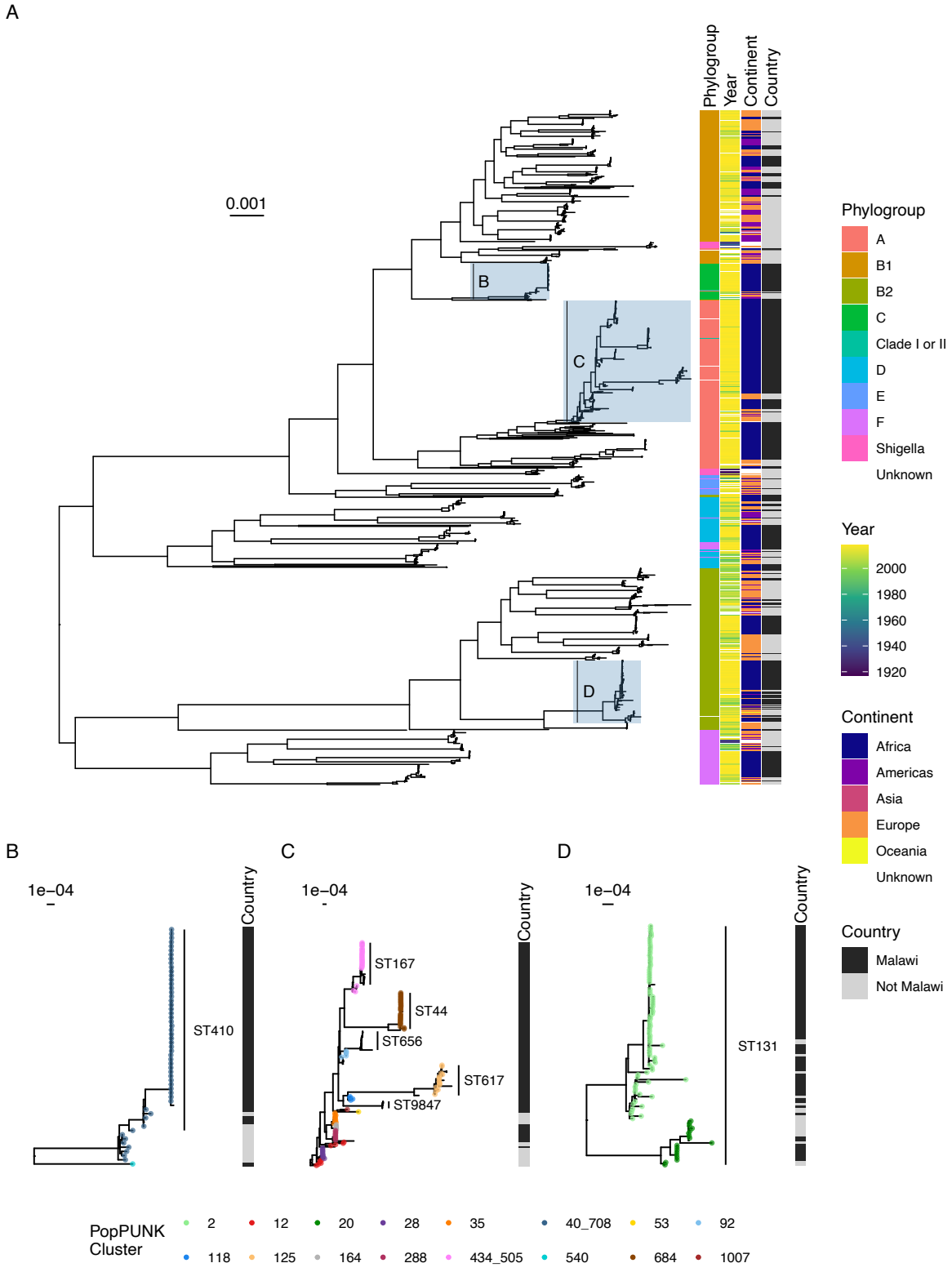


Figure 3: Midpoint rooted core-gene maximum likelihood phylogenetic tree of Malawian isolates with global context isolates (10 isolates each from the top 50 popPUNK clusters in the global collection, along with a further 92 Malawian isolates

form a previous study), showing phylogroup, year of collection, continent and country (Malawi vs not Malawi). Generally, Malawian isolates are distributed throughout the tree. B-D show magnified subtrees of the three most frequently identified STs: ST 410 (B), ST167 with surrounding phylogroup A isolates (C) and ST131 (D) with tip points coloured based on popPUNK cluster allocation. Lack of coloured point indicates that the isolate was assigned a novel cluster not present in the original global collection.

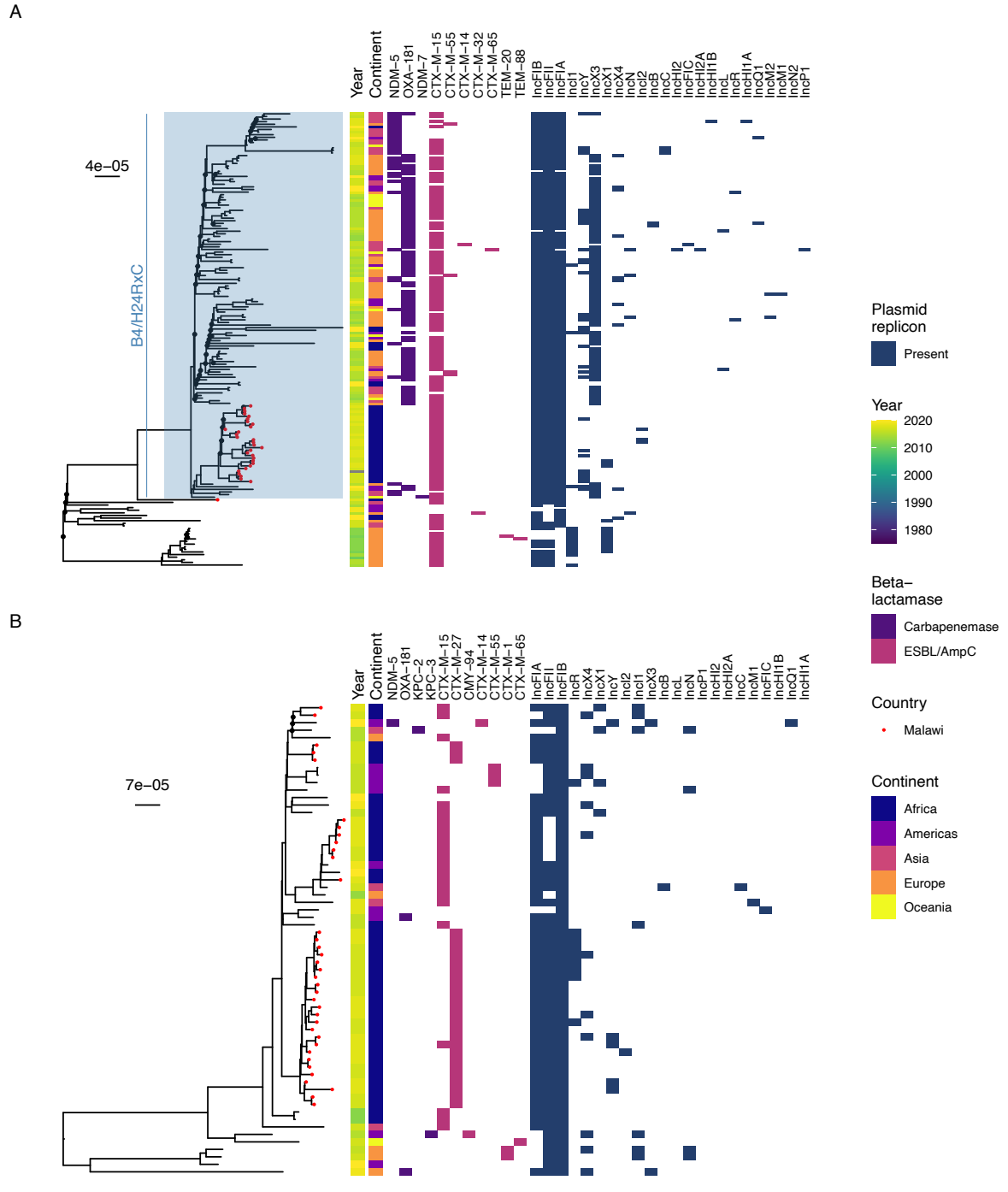


Figure 4: Subtrees of midpoint-rooted, maximum likelihood phylogenies of E. coli ST410 (A) and ST167 (B), showing the Malawian isolates (red tip points) in the context of all ST410 or ST167 isolates available in Enterobase. Assemblies were constructed by mapping to ST-specific reference genomes. ESBL/CPE genes and plasmid replicons are shown. Bootstrap support of less than 95% is shown by a black point at tree node

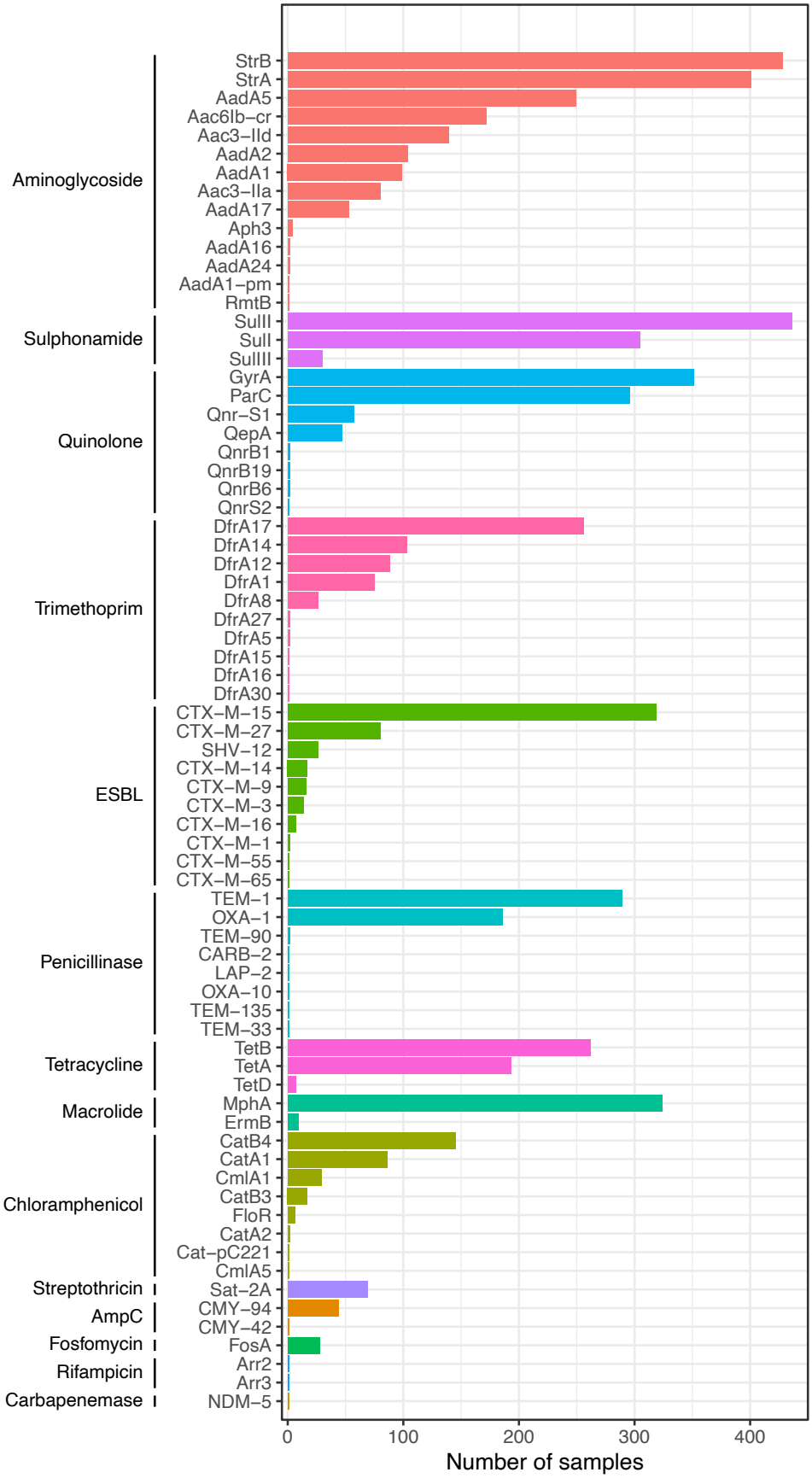


Figure 5: Distribution of identified antimicrobial resistance determinants.

Table 1: Phylogroup, sequence type (ST), continent of collection and pathotype of popPUNK-defined clusters

Cluster	Phylogroup	This study		Horesh Collection			
		n	STs	n	STs	Location	Pathotype
2	B2	53	ST131 (1.00)	781	ST131 (0.99)	Europe (0.73);Unknown (0.12);Americas (0.11);Oceania (0.03);Asia (0.00)	ExPEC (0.54);Not determined (0.46)
40_708	C	44	ST410 (1.00)	29	ST23 (0.41);ST410 (0.34);ST2230 (0.07);ST369 (0.07);ST5491 (0.07);ST5286 (0.03)	Europe (0.62);Americas (0.31);Asia (0.07)	ExPEC (0.55);STEC (0.21);ETEC (0.17);Not determined (0.07)
684	A	29	ST44 (1.00)	1	ST44 (1.00)	Europe (1.00)	ExPEC (1.00)
451	B2	27	ST636 (1.00)	1	ST636 (1.00)	Americas (1.00)	ExPEC (1.00)
434_505	A	24	ST167 (0.92);ST10 (0.08)	3	ST10 (1.00)	Americas (0.67);Europe (0.33)	ExPEC (0.67);Not determined (0.33)
44_579	F	17	ST648 (1.00)	26	ST648 (1.00)	Europe (0.42);Unknown (0.38);Americas (0.19)	ExPEC (0.73);Not determined (0.27)
1186	A	17	ST656 (1.00)	0	-	-	-
290	A	16	ST46 (1.00)	2	ST46 (1.00)	Americas (1.00)	ExPEC (0.50);Not determined (0.50)
1187	B1	15	ST940 (1.00)	0	-	-	-
125	A	13	ST617 (0.92);ST4981 (0.08)	7	ST617 (1.00)	Americas (0.71);Europe (0.29)	ExPEC (0.57);Not determined (0.43)

1188	A	12	ST167 (1.00)	0	-	-	-
1110	F	11	ST1722 (1.00)	1	ST1722 (1.00)	Europe (1.00)	ExPEC (1.00)
288	A	9	ST5824 (1.00)	3	ST227 (1.00)	Americas (1.00)	Not determined (1.00)
20	B2	8	ST131 (1.00)	48	ST131 (0.96);ST5432 (0.02);ST5494 (0.02)	Europe (0.75);Americas (0.12);Unknown (0.10);Oceania (0.02)	ExPEC (0.71);Not determined (0.29)
37	D	8	ST405 (1.00)	27	ST405 (0.96);ST964 (0.04)	Europe (0.63);Americas (0.33);Asia (0.04)	ExPEC (0.81);Not determined (0.11);ETEC (0.04);STEC (0.04)
1189	A	8	ST1421 (1.00)	0	-	-	-
7	D	7	ST69 (1.00)	174	ST69 (0.94);ST106 (0.03)	Europe (0.83);Americas (0.12);Unknown (0.05)	ExPEC (0.79);Not determined (0.19);EAEC (0.02)
1190	A	7	ST9847 (1.00)	0	-	-	-
1191	D	6	ST38 (0.83);Novel (0.17)	0	-	-	-
92	A	5	ST10 (1.00)	10	ST10 (1.00)	Europe (0.60);Americas (0.40)	ExPEC (0.60);Not determined (0.40)
164	A	5	ST34 (1.00)	6	ST34 (1.00)	Africa (0.67);Europe (0.33)	EPEC (0.67);Not determined (0.33)
330	B1	5	ST101 (1.00)	2	ST101 (1.00)	Europe (0.50);Unknown (0.50)	Not determined (1.00)
1193	A	5	ST10 (1.00)	0	-	-	-

References

1. Temkin, E. *et al.* Estimating the number of infections caused by antibiotic-resistant Escherichia coli and Klebsiella pneumoniae in 2014: a modelling study. *The Lancet Global Health* **6**, e969–e979 (2018).
2. Alvarez-Uria, G., Gandra, S., Mandal, S. & Laxminarayan, R. Global forecast of antimicrobial resistance in invasive isolates of Escherichia coli and Klebsiella pneumoniae. *Int J Infect Dis* **68**, 50–53 (2018).
3. World Health Organisation. *Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis.* (2017).
4. Denamur, E., Clermont, O., Bonacorsi, S. & Gordon, D. The population genetics of pathogenic Escherichia coli. *Nature Reviews Microbiology* **19**, 37–54 (2021).
5. Stoesser, N. *et al.* Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. *mBio* **7**, e02162 (2016).
6. Zong, Z., Fenn, S., Connor, C., Feng, Y. & McNally, A. Complete genomic characterization of two Escherichia coli lineages responsible for a cluster of carbapenem-resistant infections in a Chinese hospital. *Journal of Antimicrobial Chemotherapy* **73**, 2340–2346 (2018).
7. Feng, Y. *et al.* Key evolutionary events in the emergence of a globally disseminated, carbapenem resistant clone in the Escherichia coli ST410 lineage. *Commun Biol* **2**, 322 (2019).
8. Horesh, G. *et al.* A comprehensive and high-quality collection of Escherichia coli genomes and their genes. *Microb Genom* **7**, (2021).
9. Lester, R. *et al.* Sustained Reduction in Third-generation Cephalosporin Usage in Adult Inpatients Following Introduction of an Antimicrobial Stewardship Program

in a Large, Urban Hospital in Malawi. *Clinical Infectious Diseases* **71**, e478–e486 (2020).

10. Musicha, P. *et al.* Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998–2016): a surveillance study. *The Lancet Infectious Diseases* **17**, 1042–1052 (2017).
11. Lewis, J. M. *et al.* Emergence of carbapenemase producing Enterobacteriaceae, Malawi. *J Glob Antimicrob Resist* **20**, 225–227 (2019).
12. Lewis, J. *et al.* Dynamics of gut mucosal colonisation with extended spectrum beta-lactamase producing Enterobacterales in Malawi [submitted]. *bioRxiv*.
13. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).
14. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455–77 (2012).
15. Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microbial Genomics* **2**, e000083.
16. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–55 (2015).
17. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
18. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
19. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

20. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* **2**, e000056 (2016).

21. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

22. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics* **3**, e000131 (2017).

23. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine* **6**, 90 (2014).

24. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517–D525 (2020).

25. Wirth, T. *et al.* Sex and virulence in Escherichia coli: an evolutionary perspective. *Mol Microbiol* **60**, 1136–1151 (2006).

26. Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial agents and chemotherapy* **58**, 3895–903 (2014).

27. Joensen, K. G. *et al.* Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *J Clin Microbiol* **52**, 1501–1510 (2014).

28. Robins-Browne, R. M. *et al.* Are Escherichia coli Pathotypes Still Relevant in the Era of Whole-Genome Sequencing? *Front Cell Infect Microbiol* **6**, 141 (2016).

29. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and

detection of new phylo-groups. *Environmental Microbiology Reports* **5**, 58–65 (2013).

30. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* (2019) doi:10.1101/gr.241455.118.
31. Musicha, P. *et al.* Genomic landscape of extended-spectrum β-lactamase resistance in Escherichia coli from an urban African setting. *Journal of Antimicrobial Chemotherapy* **72**, 1602–1609 (2017).
32. Zhou, Z. *et al.* The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
33. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
34. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
35. Sichtig, H. *et al.* FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nature Communications* **10**, 3313 (2019).
36. Torsten Seemann. snippy: fast bacterial variant calling from NGS reads. (2015).
37. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).
38. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).

39. Lewis, J. *joelewis101/blantyreESBL: v1.0.0*. (Zenodo, 2021). doi:10.5281/zenodo.5554082.

40. Bevan, E. R., Jones, A. M. & Hawkey, P. M. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *Journal of Antimicrobial Chemotherapy* **72**, 2145–2155 (2017).

41. Mathers, A. J., Peirano, G. & Pitout, J. D. D. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clin Microbiol Rev* **28**, 565–591 (2015).

42. Kallonen, T. *et al.* Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Research* **27**, 1437–1449 (2017).

43. Dunn, S. J., Connor, C. & McNally, A. The evolution and transmission of multi-drug resistant Escherichia coli and Klebsiella pneumoniae: the complexity of clones and plasmids. *Current Opinion in Microbiology* **51**, 51–56 (2019).

44. World Health Organisation. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach. Second edition*. (2016).

45. Kumwenda, G. P. *et al.* First Identification and genomic characterization of multidrug-resistant carbapenemase-producing Enterobacteriaceae clinical isolates in Malawi, Africa. *J Med Microbiol* **68**, 1707–1715 (2019).

46. Ministry of Health. Government of Malawi. Malawi standard treatmemt guidelines (MSTG) 5th edition 2015. (2015).

47. Feasey, N. A. *et al.* Three Epidemics of Invasive Multidrug-Resistant Salmonella Bloodstream Infection in Blantyre, Malawi, 1998-2014. *Clinical infectious*

diseases : an official publication of the Infectious Diseases Society of America **61 Suppl 4**, S363-71 (2015).