# Genomic analysis of extended-spectrum beta-lactamase producing *Escerichia coli* colonising adults in Blantrye, Malawi reveals recent arrival of global *bla*NDM-5-associated lineages but loss of carbapenemase-encoding genes in the absence of antimicrobial selection pressure

Joseph M. Lewis1,2,3,4, Madalitso Mphasa1, Rachel Banda1, Mathew Beale4, Eva Heinz2, Jane Mallewa5, Nicholas Thomson4, Nicholas A Feasey1,2

1 Malawi-Liverpool Wellcome Clinical Research Programme

2 Liverpool School of Tropcial Medicine

3 University of Liverpool

4 Wellcome Sanger Institute

5 College of Medicine, University of Malawi

6 University of Lancaster

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](mailto:jmlewis@liverpool.ac.uk)

+44 (0)151 795 9687

### Abstract

**Background**

*Escherichia coli* is an antimicrobial resistance (AMR) associated global pathogen but genomes from sub-Saharan Africa (sSA) are underrepresented in global sequencing efforts.

**Objectives**

We carried out a genomic investigation of extended-spectrum beta-lactamase (ESBL)-producing *E.coli* colonising adults in Blantyre, Malawi to describe diversity and AMR determinants and to place these isolates in a global genomic context.

**Methods**

We carried out short-read whole-genome sequencing of 473 ESBL *E. coli*  isolates. Presence of AMR genes, phylogroup, multilocus sequence type were determined and diversity was compared to a global collection of 800 genomes with a core-gene phylogeny. The genomic epidemiology of Malawian ST167 and ST410 - global carbapenemase-associated lineages - were explored by mapping to lineage-specific references along with genomes from Enterobase.

**Results**

Diversity of Malawian isolates reflects global *E. coli* diversity. The globally disseminating carbapenemase-associated ST167 and ST410 lineages have recently arrived in Malawi but, in Malawi, have either lost their carbapenemase-associated plasmid or arise from non-carbapenemase associated ancestors. However, ST167 and ST410 associated *bla*NDM-5 carrying plasmids are circulating in Blantyre.

**Conclusions**

Blantyre samples global *E. coli* diversity. The carbapenemase phenotype is not necessary for the success of the global carbapenemase-associated ST167 and ST410 lineages, and maintaining the carbapemase containing plasmid likely confers a fitness disadvantage. All elements necessary for rapid dissemination of blaNDM-5 are present in Blantyre, highlighting the critical need for carbapenem stewardship in this setting,

### Introduction

*Escherichia coli*, the ubiquitous human gut commensal and common human pathogen, is demonstrating increasing resistance to antimicrobials worldwide1,2. *E. coli* producing extended-spectrum beta lacatamase (ESBL) and carbapememase enzymes present particular therapeutic challenges and have been identified as priority pathogens by the World Health Organisation3. 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 dissemination4. This phenomenon is well described in *E. coli* sequence type (ST) 1315, associated with the ESBL-encoding gene *bla*CTX-M-15, but has also been recently described in other *E. coli* lineages such as ST1676 and ST4107 which are associated with the carbapenemase-encoding genes *bla*NDM-5 and *bla*OXA-181.

However, global *E. coli* sequencing efforts are biased towards isolates from high-income settings8 and 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 may differ - are comparatively poorly described. In Blantyre, Malawi, for example, carbapenem use is unusual but the third-generation cephalosporin antimicrobial ceftriaxone is very widely used in the hospital setting since its introduction in the Malawian national formulary in 20059. Since that time, ESBL- producing (and hence ceftriaxone resistant) *E. coli* are an increasing problem and now make up 31% of invasive *E. coli* in Blantyre10; carbapenem resistance is unusual, though is now sporadically described11. There is a significant unmet need for access to carbapemem antimicrobials to treat resistant infections, but the example of ceftriaxone shows that carbapenem resistance may be likely to rapidly disseminate. In this context, both robust stewardship protocols and ongoing genomic AMR surveillance are critical.

To that end, we present a genomic investigation of ESBL-producing *E. coli* from a study of colonisation with ESBL Enterobacterales in Blantyre, Malawi, aiming to describe the diversity and AMR determinants of ESBL *E. coli* in Blantyre and to place these isolates in a global genomic context.

### 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[REF carriage paper]. Briefly, 225 adults (≥ 16 years) with sepsis were recruited in the emergency department of Queen Elizabeth Central Hospital (QECH), Blantyre, Malawi, along with 100 adults who were planned to be admitted to QECH but with no plans for antimicrobial administration and 100 community dwelling adults with no antimicrobial exposure (except for longterm 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 shipped at ambient to the Wellcome Sanger Institute for paired-end short-read 150bp sequencing on the Illumina HiSeq X10 instrument. Species was confirmed with Kraken v0.10.6 and Braken v1.012. We first reconstructed a core gene phylogeny for the study isolates: *de novo* assembly was undertaken with SPAdes v3.14.013 and the quality of the assemblies assessed with QUAST v5.0.214. Assembly failures with a total assembled length of < 4Mb or assemblies with a QUAST-defined contamination of ≥ 10% were excluded from further analysis. Assemblies were annotated with Prokka v1.5 with a genus-specific database from RefSeq15 and the Roary v1.007 pangenome pipeline16 used to identify core genes. Genes present in ≥ 99% samples considered to be core, and were concatenated to a core gene pseudosequence. Variable sites were identified with snp-sites v2.4.117 and used to infer a maximum-likelihood phylogenetic tree in IQ-TREE v1.6.318 with ascertainment bias correction, the ModelFinder module used to select the best fitting nucleotide substitution model, and 1000 bootstrap replicates generated.

ARIBA v.2.12.119 was used to identify AMR-associated genes using the SRST2 curated version of the ARG-ANNOT database20 on the reads, and was used to call SNPs in the quinolone-resistance determining regions (QRDR) *gyrA, gyrB, parC* and *parE*, after downloading wild-type genes from the *Escherichia coli* K-12 substr. MG1655 (NC\_000913.3) reference genome. For the purposes of describing aggregate quinolone resistance, any mutation in any QRDR was considered to confer quinolone resistance. Beta lactamases were considered to be extended spectrum or carbapenemases based on the phenotypic classifications at <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 scheme21 hosted at pubMLST (<https://pubmlst.org/>). E. coli phylogroups were determined using the Clermont scheme and primers22 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 a local and global context, we reconstructed a global *E. coli* core-gene phylogeny using publicly available using genomes, selected for global diversity and enriched for ESBL production. We included 362 enterotoxogenic *E. coli* (ETEC) from a genomic study of global ETEC isolates from 1980-201123; 185 atypical enteropathogenic *E. coli* (aEPEC) sequenced for a study of aEPEC from the Global Enteric Multicentre Study (GEMS) in seven centres in Africa and Asia between 2007-201124; 149 ESBL-producing *E. coli* from a single centre study in Chachoengsao province, eastern Thailand25 in which the study isolates were found to represent global diversity; and 94 E. coli from our centre in Blantyre where a combination of invasive and carriage isolates sequenced, selected for diversity in AMR phenotype between 1996-201426. Accession numbers for all included isolates are provided in the supplementary appendix. De novo assembly, construction of a core gene pseudosequence was performed as above and a phylogeny inferred using the same nucleotide substitution model as was identified to be the best fitting model for the Malawian isolates, with 1000 bootstrap replicates. AMR gene presence and ST were also determined as above.

To better describe the genomic epidemiology of two commonly identified STs not seen in the previous Malawian collection – ST410 and ST167 – we downloaded all ST140 and all ST167 genomes from Enterobase27 on 1st March 2021 with 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/>), multiqc v1.828 and by trimmed raw reads with Trimmomatic v0.3929, removing adapter sequences (using TriSeq3-PE.fa and NexteeraPE.fa as references provided with Trimmomatic), leading or trailing bases with phred score below 4, bases with a mean score below 20 with 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 database30 (GenBank accession CP023870.1 for ST167 and CP023870.1 for St410) using the snippy v4.6.031 pipeline with default settings; mapped assemblies with mean mapped depth below 20x were excluded, areas of putative recombination were screened with Gubbins v 3.0.032, then SNP sites extracted and a phylogeny reconstructed with IQ-TREE as above. Finally, to explore the presence of the pNDM-MGR194 *bla*NDM-5 associated plasmid previously identified in Blantyre11, we mapped all ST167 and ST410 reads to the plasmid reference (GenBank accession KF220657.1) using snippy, calculated sequence coverage using the number of mapped bases excluding low quality bases as per snippy defaults, and sequence identity to the reference using the snippy consensus substitution SNP-calls.

All statistical analyses were carried out in R v4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) and trees were visualized in the *ggtree* v2.2.433package. 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 (16-062) and Malawi College of Medicine (16-062). Reads from all isolates sequenced in this study have been deposited in the European Nucleotide Archive: accession numbers and associated metadata are provided in the supplementary appendix.

### Results

Following quality control, 473 E. coli genomes were included in the analysis, 440 from participants who were enrolled in the hospital, and 33 from community members, with a median 2 (IQR 1-5) samples per participant. The commonest 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 1B). 56 recognised STs were identified, with a median 20 (IQR 13-38) samples per ST (Figure 1A). 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%]). De-novo assembly resulted in assemblies with a median 92 (IQR 68-122) contigs and N50 of 180kbp (IQR 123-234kbp, see Supplementary Appendix for assembly statistics), and annotation with Prokka followed by the Roary pan-genome pipeline identified a pan-genome of 26,840 genes of which 2966 were core. A maximum-likelihood phylogeny inferred from the core gene psedosequence (fit with the general time reversible nucleotide substitution model with FreeRate site heterogeneity with 5 parameters, as selected by the IQ-TREE ModelFinder module) showed that the STs, as expected, were largely monophyletic and mapped well to the tree topography (Supplementary figure 1).

The identified AMR determinants are shown in Figure 2. Only one isolate contained a carbapenemase encoding gene: *bla*NDM-5  carried on a 46.2kbp Inc-X3 plasmid (closely related – 11 SNPs - to the blaNDM-5 associated plasmid, pNDM-MGR194), which we have described previously11. The remainder (n = 472) carried at least one ESBL-encoding gene, most commonly *bla*CTX-M-15 which was present in 319/473 (67%) of isolates. All other identified ESBL-encoding genes were members of the *bla*CTX-M family with the exception of *bla*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*SHV-12. Determinants of resistance to aminoglycosides (99% [469/473] of isolates), trimethoprim (97% [459/473]), sulphonamides (99% [468/473]) and quinolones were very common (88% [418/473] ); genes conferring resistance to chloramphenicol less so (52% [248/473]).

We next placed the Malawian carriage isolates in the context of the global collection with a core gene phylogeny. In the 1273 included isolates, the Roary pan-genome pipeline identified 2872 core genes in a pangenome of 44,840 genes, which were concatenated to a pseudosequence of 604,817 bases with 77,194 variable sites, which were used to infer a phylogeny (Figure 3). The majority of Malawian isolates clustered closely with global isolates; Malawian isolates of the established ESBL-associated global high risk clone ST131, for example, were closely related to global ST131 strains, suggesting multiple importations to Malawi. However, there were two areas which, in this collection, seemed to be Malawi-restricted: ST410, and a group of phylogroup A isolates, especially ST167. Both of these STs (the third- and second- most frequently identified STs in the current study) were not seen at all in the previous study of Malawian *E. coli* isolates from 1996-2014, suggesting a possible recent introduction.

To further explore the genomic epidemiology of ST410 and ST167 we mapped reads from all ST410 and ST167 isolates from this study to ST-specific references along with all ST410 and ST167 isolates from Enterobase with available Illumina short reads and metadata. After quality control, 281 ST167 and 511 ST410 Enterobase isolates were included; median mapped depths were 59x (IQR 43-79) and 65x (IQR 50-83) respectively. Following screening of recombinations with Gubbins,, 42,526 variable sites for ST167 and 39,802 sites for ST410 were used to infer phylogenies (Figures 4-6). The phylogenies demonstrate that ST167 and ST410 are globally distributed carbaopenemase-associated lineages; ST167 is associated with *bla*NDM-5 and ST410 with both *bla*NDM-5 and *bla*OXA-181, as previously described6,7. The B4/H24RxC lineage of ST410 has been described as a carbapenemase-associated high-risk global clone7, as is evident from the phylogeny, but in fact acquisition of carbapenemase-encoding genes has occurred multiple times independently in both ST410 and ST167. All Malawian ST410 (with the exception of a single isolate) were monophyletic, not clustering with global isolates, and descended from a single ancestor in the B4/H24RxC lineage with good (>95%) bootstrap support. These Malawian ST410 have lost their carbapenemase resistance determinants whilst maintaining the *bla*CTX-M-15 and *bla*CMY-94 genes characteristic of the lineage. In contrast, Malawian ST167 isolates clustered in a clade with global ST167 isolates, consistent with repeated introductions to Malawi. This clade was a sister to the globally successful carbapenemase-associated ST167 isolates, with only sporadic presence of carbapenemase genes.

Finally, given our previous identification of a globally disseminated pNDM-MGR194 *bla*NDM-5 associated plasmid in Blantyre, we mapped all ST167 and ST410 reads to that plasmid reference. 25/327 (7%) of ST167 and 32/557 (6%) of ST410 samples mapped to the reference at ≥ 95% coverage and sequence identity and the non-Malawian ST410 B4/H24RxC lineage very frequently mapped 40-42kb at 100% identity, consistent with a closely related plasmid (Supplementary Figures 2 and 3). This was not the case in any of the Malawian ST410 isolates, consistent with the loss of the carbapenem-containing B4/H24RxC-associated plasmid.

### Discussion

We present an analysis of 473 genomes of ESBL-producing *E. coli* colonising adults from Blantrye, Malawi. sSA is underrepresented in global *E. coli* collections: of 10,146 genomes in a recently published curated *E. coli* database8, only 246 were from the African continent, so this dataset and associated metadata contributes significantly to AMR genomic surveillance in Malawi and sSA. From these data we can draw several conclusions both with regard to the epidemiology of AMR in Malawi and the biology of carbapenem resistance in *E. coli*.

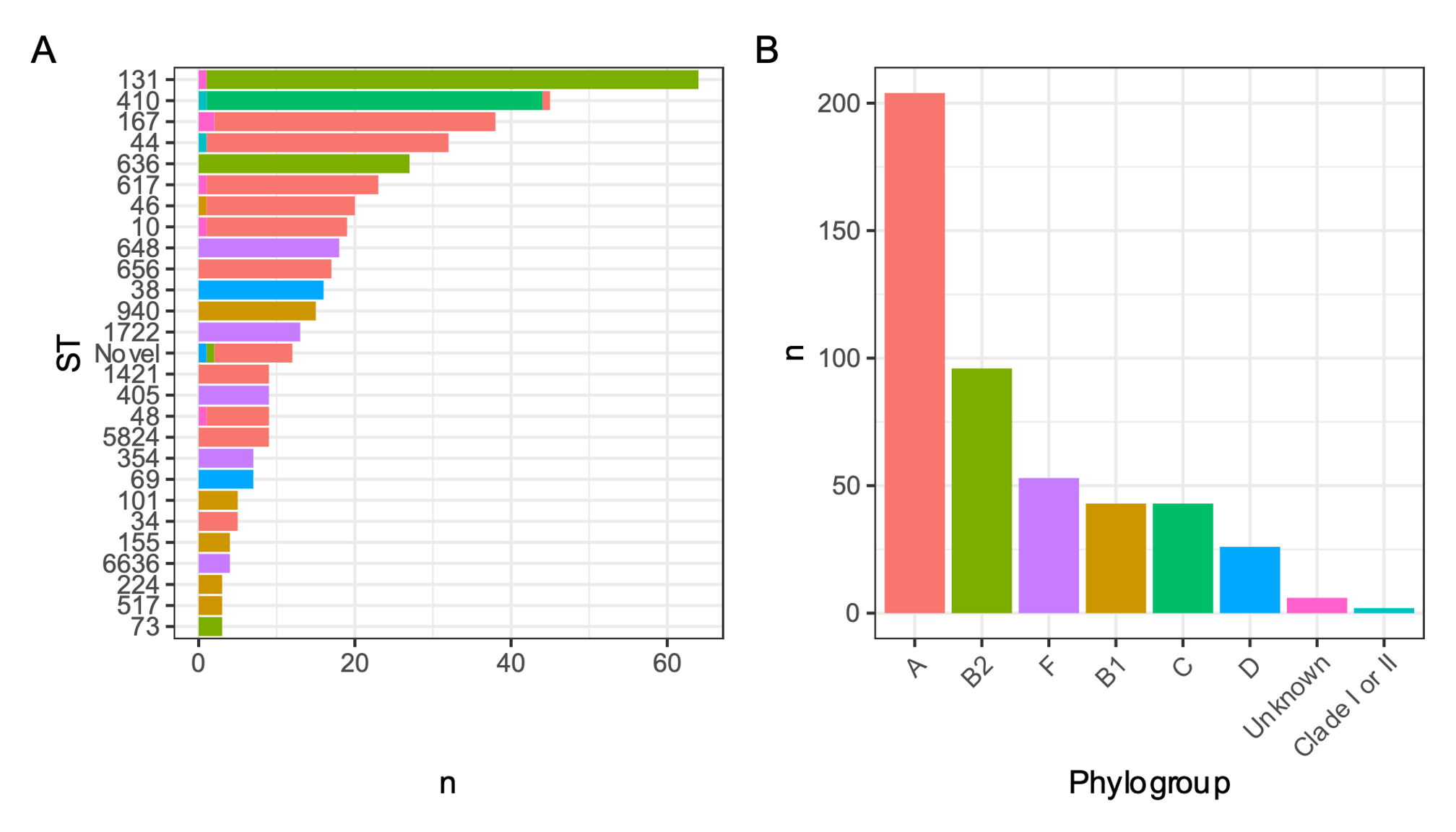
First, Malawi samples global *E. coli* diversity: the isolates in this study encompass the full diversity of the species with all major phylogroups and 56 STs represented. Malawian ESBL *E. coli* largely cluster with global isolates, suggesting unrestricted mixing of strains at a global scale. Global trends in ST distribution are reflected in Blantyre; ST131, the most frequently isolated pathogenic ST in many settings worldwide34, was the most commonly isolated ST. The recent globally emergent carbapenemase-associated STs ST167 and ST410 seem to have recently arrived in Blantyre, mirroring global trends6,7: these STs were not seen in the previous Blantyre *E. coli* collection26, suggesting recent arrival.

Second, though these globally carbapemase-associated STs have recently arrived in Blantyre, in the Malawian setting they are not carbapenemase-associated. This suggests that the global dissemination of these clones may be driven by virulence or other determinants of fitness and is not dependent on the AMR-phenotype; and that there is a fitness disadvantage in maintaining carbapenemase-associated mobile genetic elements in the absence of antimicrobial pressure. The evolution of carbapenemase absence differs between ST410 and ST167 lineages: the ancestral Malawian ST410 strain lost the B4/H24RxC-associated carbapenemase-carrying plasmid, whereas global non-carbapenemase carrying ST167 (arising from sister clades to the global carbapemase-associated lineages) have been introduced multiple times to Malawi. The relative contributions of virulence determinants (or other fitness determinants) versus antimicrobial resistance determinants in driving global spread of high risk clones is not clear4, but our data suggest that, in the case of carbapenemase-associated ST167 and ST410, the carbapenemase is not necessary for success.

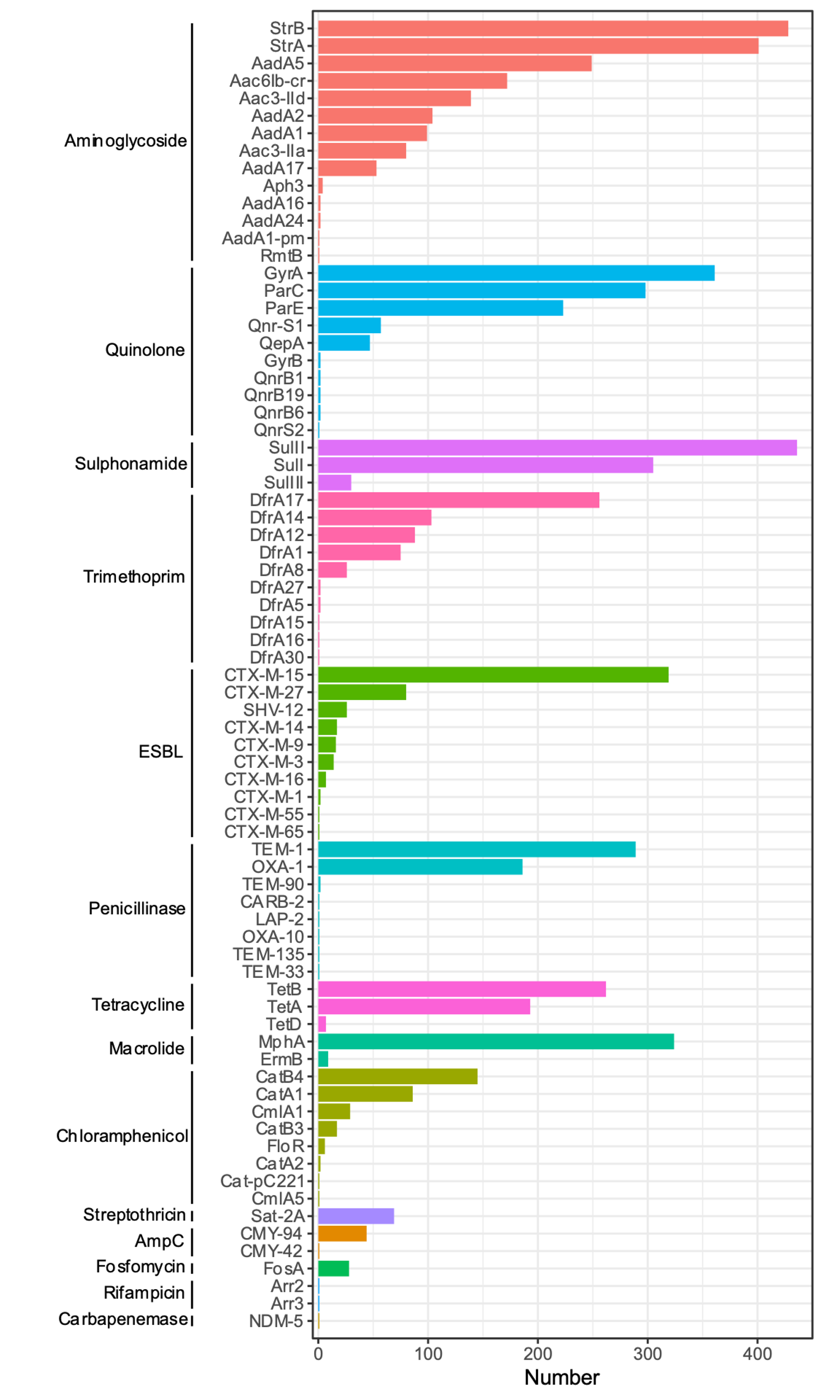
Third, all elements necessary for rapid emergence of *bla*NDM-5 are present in Blantyre, highlighting the urgent need for carbapenem stewardship. We previously described the identification of a globally-distributed *bla*NDM-5-associated plasmid in Blantyre, but we here demonstrate that very similar plasmids are strongly associated with the ST167 and (particularly) ST410 lineages. It seems likely that rapid dissemination of these plasmids across these lineages in Blantyre will occur in the presence of antimicrobial pressure. The high prevalence of ESBL-producers amongst invasive Enterobacterales in Blantyre10 means that there is already a significant unmet need for access to carbapenems. Balancing this need against the necessity for stewardship of these last line antimicrobials in a setting where much treatment of febrile illness is empiric is challenging, and needs evidence-based stewardship protocols that can be deployed in LMIC.

There are limitations to our study. Some participants provided multiple samples so it is possible that this introduced bias into the collection. Global collections of genomes are by their nature biased as they may be outbreak investigations (and hence include very clonal isolates, as can be seen in the ST167 and ST410 tree topography), may focus on a particular kind of isolate (i.e EPEC or ETEC), and tertiary centres or some centres or countries may be over-represented. Our study was based at a single urban centre, so may not be generalisable.

In conclusion, we present an analysis 473 genomes of ESBL-producing *E. coli* colonising adults from Blantrye, Malawi, significantly contributing to the genomic surveillance of AMR in this low-income setting. We find that the diversity of *E. coli* in Blantyre reflects global diversity, and that recently globally disseminating carbapenemase-associated lineages have also recently arrived in Malawi. In the Malawian setting these lineages have either lost their carbapenemase gene, or have arisen from non-carbapenem associated ancestral strains, suggesting a fitness disadvantage in maintaining carbapenemase-assaociated mobile genetic elements in a setting where carbapenem antimicrobials are rarely used. Finally, we demonstrate that all elements necessary for rapid dissemination of the *bla*NDM-5 carbapenemase encoding gene – *bla*NDM-5 associated high risk clones and associated plasmids – are present in Blantyre, highlighting a critical need for stewardship as carbapenems are introduced to this setting.



**Figure 1:** **ST (A) and phylogroup (B) distribution of included isolates.**



**Figure 2:** **Distribution of identified antimicrobial resistance determinants.**

**Figure 3: Midpoint-rooted maximum-likelihood global *E. coli* phylogeny.** Malawian *E. coli* isolates in the context of a global collection showing year, country and continent of isolation, and isolate source. Most Malawian isolates are distributed through the tree and the Malawian ST131 isolates, for example, are closely related to global isolates. However, there are exceptions: ST410 and a group of phylogroup A isolates including ST167 which, in this collection, seem to be Malawi-restricted and both are not represented in the previous Malawian collection of isolates form 1996-2014.



**Figure 4: Midpoint-rooted maximum-likelihood global *E. coli* ST167 phylogeny.** Malawian ST167 *E. coli* isolates in the context of all ST167 isolates downloaded from Enterobase with available metadata, showing year and continent of isolation along with ESBL and carbapenemase genes. Malawian isolates are shown by red points on tree tips. ST167 is a globally successful *bla*NDM-5 associated clone, with two lineages in which *bla*NDM-5 has arisen independently; but the Malawian ST167 isolates belong to a sister clade to the strongly *bla*NDM-5 associated lineages. Nevertheless, global isolates are closely related to Malawian ST167 isolates, including sporadic *bla*NDM-5 and *bla*OXA-181 -associated isolates, suggesting multiple arrivals of ST167 to Malawi.



**Figure 5: Midpoint-rooted maximum-likelihood global *E. coli* ST410 phylogeny.** Malawian ST167 *E. coli* isolates in the context of all ST167 isolates downloaded from Enterobase with available metadata, showing year and continent of isolation along with ESBL and carbapenemase genes. Malawian isolates are shown by red points on tree tips. The globally distributed *bla*NDM-5 and *bla*OXA-181 associated B4/H24RxC lineage is shown highlighted in blue. Malawian ST410 isolates (with the exception of one isolate) are descended from B4/H24RxC but have lost the carbapenemase gene whilst maintaining the ESBL resistance determinents *bla*CTX-M-15 and *bla*CMY-92.



**Figure 6: Expansion of global E coli ST410 B4/H24RxC phylogeny.** Year and continent of isolation along with ESBL and carbapenemase genes are shown, and Malawian isolates are shown by red points on tree tips. Bootstrap support of less than 95% is shown by a black point at tree node, showing good bootstrap support for the Malawian ST410 arising from B4/H24RxC.

**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. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**, R46 (2014).

13. 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).

14. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

15. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

16. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

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

18. 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).

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

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

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

22. 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).

23. von Mentzer, A. *et al.* Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution. *Nature Genetics* **46**, 1321–1326 (2014).

24. Ingle, D. J. *et al.* Evolution of atypical enteropathogenic E. coli by repeated acquisition of LEE pathogenicity island variants. *Nature Microbiology* **1**, 15010 (2016).

25. Runcharoen, C. *et al.* Whole genome sequencing of ESBL-producing Escherichia coli isolated from patients, farm waste and canals in Thailand. *Genome Medicine* **9**, 81 (2017).

26. 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).

27. 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).

28. 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).

29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

30. 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).

31. Torsten Seemann. snippy: fast bacterial variant calling from NGS reads. (2015).

32. 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).

33. 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).

34. 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).