The distribution of antimicrobial resistance genes across phylogroup, host species and geography in 16,000 publicly-available *E. coli* genomes

Elizabeth Pursey[1,2*], Tatiana Dimitriu[1], William H. Gaze[3], Edze R. Westra[1] and Stineke van Houte[1]

[1]Environment and Sustainability Institute, Biosciences, University of Exeter, Penryn, Cornwall, United Kingdom

[2]Department of Experimental Medicine, University of Lund, Sweden

[3]European Centre for Environment and Human Health, University of Exeter Medical School, University of Exeter, Penryn, Cornwall, United Kingdom

*****Correspondence to: ellie.pursey@gmail.com

Key words: antimicrobial resistance, *E. coli*, phylogroups, hosts, regions

Abstract

E. coli is a highly diverse bacterial species that generates a huge global burden of antimicrobial-resistant infections. A wealth of whole genome sequence data is available on public databases for this species, presenting new opportunities to analyse the distribution of antimicrobial resistance (AMR) genes across its genetic and ecological diversity. We extracted and categorised metadata on host species and geographic location and combined this with *in silico* phylogrouping to describe the characteristics of ~16,000 assembled *E. coli* genomes from the NCBI RefSeq database. We estimated AMR carriage using various metrics: counts of overall genes, multidrug- and extensively drug-resistant categories, and selected β-lactamases of current global concern - $bla_{\text{CTX-M}}$ and carbapenemase genes. We present estimates of AMR carriage for these metrics by species type (human, agricultural/domestic animal, wild birds and other wild animals), geographic subregion, and across phylogroups. In addition, we describe the distribution of phylogroups within host types and geographic subregions. Our findings show high AMR carriage in commensal-associated phylogroups, agricultural and wild animal hosts and in many subregions. However, we also quantify large biases in sequencing data, the substantial gaps in our knowledge of AMR in many hosts, regions and environmental settings, and the need for systematic sampling to gain a more accurate picture.

Introduction

Escherichia coli is a remarkably diverse species both genetically and ecologically. It occupies various niches, ranging from a commensal of many warm-blooded organisms, to a globally devastating pathogen, to a free-living environmental bacterium [1]. It is a leading cause of mortality associated with drug-resistant infections and was the pathogen responsible for the most deaths attributed to antimicrobial resistance (AMR) in 2019 [2]. In addition, carbapenem-resistant and extended-spectrum β-lactamase (ESBL)-producing *E. coli* are World Health Organization priority pathogens, for which development of new antibiotics is urgently needed [3]. In light of the ongoing AMR crisis, it is necessary to understand how resistance genes are distributed across the ecological and genetic diversity of this species, highlighting the potential risks they pose across contexts.

AMR genes (ARGs) are part of the substantial accessory gene content found in *E. coli* [4], and therefore are not distributed uniformly across the species phylogeny. Acquired ARGs are particularly important due to their extensive ability to transmit horizontally in populations on mobile elements such as plasmids, leading to AMR spread across more distant locations and genetic lineages [5]. Some of these ARGs present significant clinical risks in humans, particularly those with current or potential spread across bacterial host genetic backgrounds, including to other species, and along the commensal-pathogen continuum [6].

Most research has focused on AMR in human-associated and pathogenic isolates [7]. A study focusing on ~1,000 isolates selected to represent the diversity of *E. coli*, mostly including isolates from non-clinical origins, found ARGs were associated with strains from humans and domesticated animals, independent of phylogroup [8]. Therapeutic and growth promoter usage of antibiotics in agriculture is well-documented, and has been linked to high rates of AMR in livestock-associated isolates [9]. Resistance may also spill over into wild animal populations, and there is increasing evidence of wildlife as potential reservoirs of AMR [10]. A One Health approach to AMR emphasises the interconnectedness of human, animal and environmental microbiomes [11], with the potential for bidirectional flow of ARGs between animal (wild or domesticated), crop, natural and built environment, and human microbiomes. There are also geographic and socioeconomic biases in studies of AMR in *E. coli*. Whilst antibiotics are used most in high-resource settings, the burden of AMR is estimated to

disproportionately affect lower and middle income countries, despite the scarcity of data in these regions [2].

A wealth of previous work on *E. coli* has investigated AMR trends in specific settings, such as individual hospitals or farms. As sequencing technologies become more accessible and public genomic databases grow, much larger genomics-based studies of AMR in highly-sampled species like *E. coli* are becoming increasingly feasible. For example, a recent large-scale study of over 70,000 *E. coli* genomes found that the number of distinct ARGs varied between phylogroups, and that specific resistance genes were found more frequently in some, though the pattern of resistance to antibiotic classes remained stable across groups [12]. Another study curated a collection of ~10,000 *E. coli* isolates sampled from human sources, and identified lineages where >50% of isolates were MDR [13].

In this work we expand on previous large-scale genomic studies of ARGs in *E. coli* by investigating diversity across different host categories and geographic subregions, as well as phylogroups. We characterise a final dataset of ~16,000 publicly-available assembled *E. coli* genomes, documenting the distribution of ARGs according to several metrics: counts of ARGs, multidrug- and extensively drug-resistant (MDR and XDR) classifications, and presence of clinically-important β-lactamases ($bla_{\text{CTX-M}}$ and carbapenemases). First, we used model comparison to assess the contributions of host, subregion and phylogroup towards explaining ARG variation in *E. coli*. Then, we used generalised linear models (GLMs) to describe the probabilities of MDR, $bla_{\text{CTX-M}}$ and carbapenemase presence according to the same predictors. To do so, we used subsampling and resampling approaches to correct for sample size disparities between groupings, and investigate the sensitivity of trends according to subsample sizes.

Methods

Genomes

The initial complete dataset of 26,881 *E. coli* genomes was retrieved from the National Center for Biotechnology Information (NCBI) RefSeq database in February 2022 in nucleotide fasta format using ncbi-genome-download v0.3.0 (https://github.com/kblin/ncbi-genome-download).

Detection of ARGs, phylogrouping and retrieval of metadata

NCBI AMRFinderPlus v3.10.21 was used to predict the identity of acquired ARGs [14]. Clermont phylogrouping [15] was performed *in silico* using the EzClermont command-line tool [16]. Snakemake v6.2.1 [17] was used to generate a pipeline for phylogrouping and detecting ARGs, which was run using the University of Exeter's Advanced Research Computing facilities. We removed efflux (*acrF*, *emrD* and *mdtM*) and *blaEC* β-lactamase genes from the dataset, as they were either very common or ubiquitous. As we could not use phenotypic resistance categories, we used the classes included in NCBI AMRFinderPlus output as a proxy for categorising genomes as multidrug-resistant (MDR; ARGs conferring resistance to ≥ 3/20 antibiotic classes) and extensively drug-resistant (XDR; ARGs conferring resistance to ≥ 10/20 antibiotic classes) [18].

Host and location metadata were retrieved and categorised using the Bio.Entrez utilities from Biopython v1.77. We filtered genomes for those with complete metadata for host species and geographic location, as well as those not typed as *E. coli* by phylogrouping, which excluded 10,609 genomes, resulting in a final dataset of 16,272 genomes. Geographic locations were split into 20 subregions (Fig. S1) according to Natural Earth data (https://www.naturalearthdata.com/).

All genomes were sorted into the following host species categories: ‘Human’, ‘Agricultural/Domestic animals’, ‘Wild birds’ and ‘Wild animals’. This was achieved using regular expressions constructed by manually reviewing text in the ‘host’ field of the biosample data for each accession number. Wild birds were separated from other wild animals due to their comparatively large numbers of genomes, high mobility and potential for ARG spread [19]. Any animal that is most likely to be associated with humans, such as ‘mouse’ or ‘canine’, was assigned to the ‘Agricultural/Domestic animals’ category, which also included farm animals such as cows and chickens. Broader responses that could not be reliably classified into any specific group, such as ‘Animal’, and ‘Avian’, were not included in further analyses, as well as animals likely to be human food sources such as mussels. Though a conservative approach was taken when assigning genomes to wild animal or bird host groups, these classifications determine the likely category of the host based on species, and cannot rule out cases such as zoo animals.

Statistical analysis

Data tidying and statistical analysis was done using R v4.0.4, with dplyr v1.0.1, ggplot2 v3.3.6 [20] and MetBrewer [21] for colour palettes. Maps additionally used the R packages sf 1.0-7 [22] and rnaturalearth v0.1.0.

In all cases, binomial GLMs were fitted using glmmTMB v1.1.5 [23]. To assess model fit, simulated residuals were plotted and their dispersion tested using DHARMa v0.4.5 [24]. Prior to all statistical modelling, subgroupings with sample sizes ≤ 100 (e.g. Central Asia) were removed from the dataset. Sample sizes for each predictor in the full dataset are shown in Figure 1A-C.

Initially, maximal models were fitted using the entire dataset with MDR, XDR or $bla_{\text{CTX-M}}$ genes as a binomial response variable (presence/absence) and all explanatory variables included (host, phylogroup and subregion). For carbapenemase genes, the same response variables were used, but only host and phylogroup were included as predictors due to lack of data coverage for many subregions for these comparatively rare genes. Sample sizes for all combinations of response variable and predictor are shown in Tables S1-3. Next, Akaike information criterion (AIC) scores were used to assess the fit of all possible models using the *dredge* function from MuMIn v1.46.0. In all cases, large ΔAIC values (>50) were seen between the maximal model and the next best model (Table S4), emphasising the contribution of all variables to the distribution of AMR gene measures. Model coefficients for all maximal models are shown in Table S5.

Next, subsampling approaches were used to circumvent disparities in sampling sizes between groups. Separate models were made for each predictor due to the difficulty in 1) taking single representative subsamples spanning all hosts, phylogroups and subregions and 2) estimating means from models with multiple categorical explanatory variables. A subsample was taken for each predictor, giving datasets with equal numbers of genomes from each phylogroup, host or subregion. Several subsample sizes were used to determine the sensitivity of trends to the sample size. Firstly, the dataset was subsampled without replacement down to the lowest group size for each model (e.g. all host categories were sampled to n=234, the sample size for wild birds). In addition, samples with replacement were taken with increasing sizes (n=500,

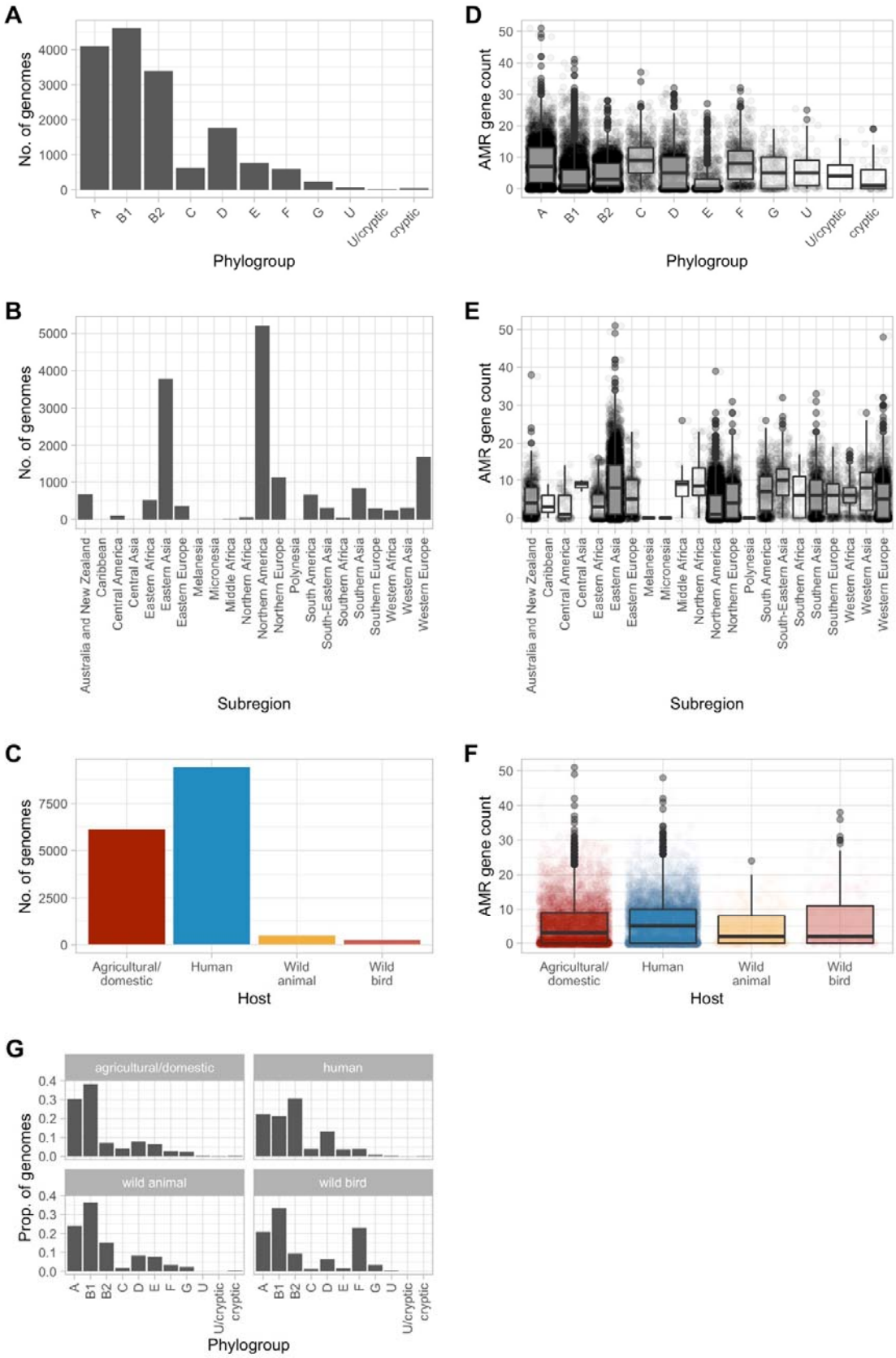
1000, 2000 and 4000) for hosts, phylogroups and subregions, to investigate the robustness of trends to changing sample sizes. All single predictor models were also run with the full, non-subsampled dataset.

Model estimated means and 95% confidence intervals (CIs) were calculated using *ggpredict* from the ggeffects package [25]. Model coefficients are presented in Table S6 for host, Table S7 for subregion and Table S8 for phylogroup. We did not perform post-hoc comparisons or report *P*-values in the manuscript due to the large number of pairwise comparisons that would be necessary (e.g. 91 comparisons for each model of subregion data). In addition, we did not have *a priori* hypotheses about group differences to test, but rather aimed to describe the patterns in the data with our modelling approaches. Finally, many groups had large sample sizes, which can produce a significant result even for a biologically insignificant effect size [26, 27].

Results

Quantification of sample sizes across hosts, phylogroups and subregions in the RefSeq dataset

Initially, we characterised the RefSeq dataset by looking at sample sizes across phylogroups, host categories, and geographic subregions (Fig. 1A-C, Tables S1-3). Phylogroups B1, A and B2 collectively made up 74% (12,109/16,272) of genomes, with smaller numbers found across C, E, F, G and U (Fig. 1A). Sixty-five genomes were typed as U/cryptic or cryptic. The majority of genomes had a sampling location in North America (n=5,202) or Eastern Asia (n=3,790), whilst many regions were represented poorly such as Micronesia, Melanesia and Polynesia (n=1 for all), Central Asia and Caribbean (n=3 for both) and Middle Africa (n=10) (Fig. 1B). Finally, 96% of the *E. coli* genomes were isolated from humans and agricultural or domestic animals (15,549/16,272) with only 234 from wild birds and 489 from other wild animals (Fig. 1C).


A
No. of genomes
Phylogroup
A
B1
B2
C
D
E
F
G
U
U/cryptic
cryptic
D
AMR gene count
Phylogroup
B
No. of genomes
Subregion
Australia and New Zealand
Caribbean
Central America
Central Asia
Eastern Africa
Eastern Asia
Eastern Europe
Melanesia
Micronesia
Middle Africa
Northern Africa
Northern America
Northern Europe
Polynesia
South America
South-Eastern Asia
Southern Africa
Southern Asia
Southern Europe
Western Africa
Western Asia
Western Europe
E
AMR gene count
Subregion
C
No. of genomes
Agricultural/ domestic
Human
Wild animal
Wild bird
Host
F
AMR gene count
Host
G
agricultural/domestic
human
wild animal
wild bird
Prop. of genomes
Phylogroup

200

Figure 1 – Total number of genomes per A) phylogroup, B) geographic subregion and C) host species category. Counts of ARGs for D) phylogroup, E) geographic subregion and F) host species category, with raw data plotted below boxplots. G) Proportion of genomes in each phylogroup per host category.

ARG counts across the diversity of E. coli *sequences*

We measured counts of ARGs detected across the diversity of genomes to get an overall measure of how AMR carriage was distributed across the dataset (Fig. 1D-F). Overall, the mean number of ARGs per genome for the entire dataset was 5.75. Median ARG counts were highest in phylogroups C, F and A (9, 8 and 7 respectively; Fig. 1D). Meanwhile the lowest median counts were in groups E (0) and B1 (1). A wide range of ARG counts was also seen between and within subregions (Fig. 1E). Among the most well-sampled regions, genomes from Eastern Asia had the highest median ARG count (8), though relatively high numbers were also seen in Western Europe (median ARG count = 5). Additionally, higher median ARG numbers were seen in less represented regions such as Northern (8.5), Middle (9) and Southern Africa (6) and Central Asia (9). High within-group diversity in ARG counts was again seen for host categories, with highest median counts in genomes from humans (5) and agricultural animals (3).

Hosts and subregions show different phylogroup patterns

We investigated the distribution of *E. coli* phylogroups within hosts (Fig. 1G) and geographic subregions (Fig. S2). *E. coli* isolates from agricultural and domestic animals mostly belonged to phylogroups A and B1 (n=1,863 and n=2,354, respectively). Those from humans were most likely to belong to phylogroup B2 (n=2,870), as well as A and B1 (n=2,071 and n=2,000 respectively). There were also comparatively high numbers of phylogroup D isolates in genomes from human hosts (n=1,237). A higher proportion of phylogroup F isolates were seen in wild birds (n=54), whilst wild animals had slightly higher proportions of B2 isolates (n=74) than agricultural or domesticated animals.

The frequencies of different phylogroups varied between geographic subregions (Fig. S2). While A, B1 and B2 were typically the largest groups, their relative proportions varied. For example, B1 was more common in Eastern (n=195) and Western (n=94) Africa, B2 was

highest in Australia and New Zealand (n=216) and Eastern Europe (n=112), while A was most common in South-Eastern (n=131) and Western (n=117) Asia. Isolates belonging to phylogroup D were also relatively common, with notably higher proportions in regions including Northern Africa (n=17) and Eastern Europe (n=86).

Specific measures of AMR carriage

Following characterisation of the dataset, we began analysing more specific measures of AMR carriage. First, we categorised genomes as MDR (≥ 1 ARGs conferring resistance to ≥ 3 classes) and XDR (≥ 1 ARGs conferring resistance to ≥ 10 classes) following detection of 402 unique resistance genes conferring predicted resistance to 19 classes. We also looked at the presence of selected β-lactamase genes of current global health concern. We detected 40 unique $bla_{\text{CTX-M}}$ genes across the final dataset, most commonly $bla_{\text{CTX-M-15}}$ (n=2,536), $bla_{\text{CTX-M-14}}$ (n=1,063) and $bla_{\text{CTX-M-55}}$ (n=903). Meanwhile, a total of 46 different carbapenemases from the NDM, KPC, OXA, IMP, VIM, IMI and GES classes were identified, the most frequent being NDM-5 (n=893), KPC-2 (n=264), NDM-1 (n=219) and OXA-48 (n=189).

We generated model estimated mean proportions with 95% CIs for models of MDR (Fig. 2), XDR (Fig. S3) and $bla_{\text{CTX-M}}$ (Fig. 3) presence, according to host, phylogroup and subregion for both the full dataset, and a dataset subsampled down to the smallest group size without replacement to adjust for large sample size differences between groups. The same process was repeated for carbapenemase gene presence (Fig. 4), but subregion was not used as a predictor due the rarity of carbapenemase-positive genomes across many regions. Trends were broadly consistent across the two datasets and estimates for the full dataset are quoted in the subsequent text.

1. *MDR and XDR*

MDR genomes were common, with a given genome from the majority of phylogroups, subregions and hosts having > 0.4 probability of being MDR (Fig. 2). Within phylogroups, multidrug resistance was most common in groups C (0.83, 95% CI=0.80-0.86), F (0.75, 95% CI=0.71-0.78) and A (0.72, 95% CI=0.70-0.73), whilst being comparatively uncommon in group E isolates (0.25, 95% CI=0.22-0.28). Isolates from humans (0.60, 95% CI=0.58-0.60) and agricultural/domestic hosts (0.51, 95% CI=0.50-0.52) had the highest probabilities of

MDR, though estimates from wild animals (0.44, 95% CI=0.39-0.48) and birds (0.43, 95% CI= 0.37-0.50) were reasonably high. Finally, South-Eastern Asia (0.86, 95% CI = 0.83-0.90) and Western Africa (0.78, 95% CI=0.15-0.72) had the highest estimated proportions of MDR genomes, while the lowest were in Northern America (0.40, 95% CI=0.38-0.41) and Central America (0.42, 95% CI=0.33-0.52).

XDR genomes were less common, and not detected at high enough frequencies to estimate their probabilities in the minimum subsampled dataset for phylogroup B2 and wild animal hosts (Fig. S3). The phylogroups with the highest probabilities of XDR were A (0.092, 95% CI=0.083-0.10) and F (0.07, 95% CI=0.052-0.093), whilst remaining very low for B2 (0.003, 95% CI=0.0018-0.0058) and E (0.018, 95% CI=0.012-0.03). The highest XDR probabilities were estimated in wild birds (0.13, 95% CI=0.091-0.18), with probabilities below 0.06 for other host categories. Finally, XDR genomes were by far the most likely to occur in Eastern Asia (0.16, 95% CI=0.14-0.17), and below 0.08 for all other regions.

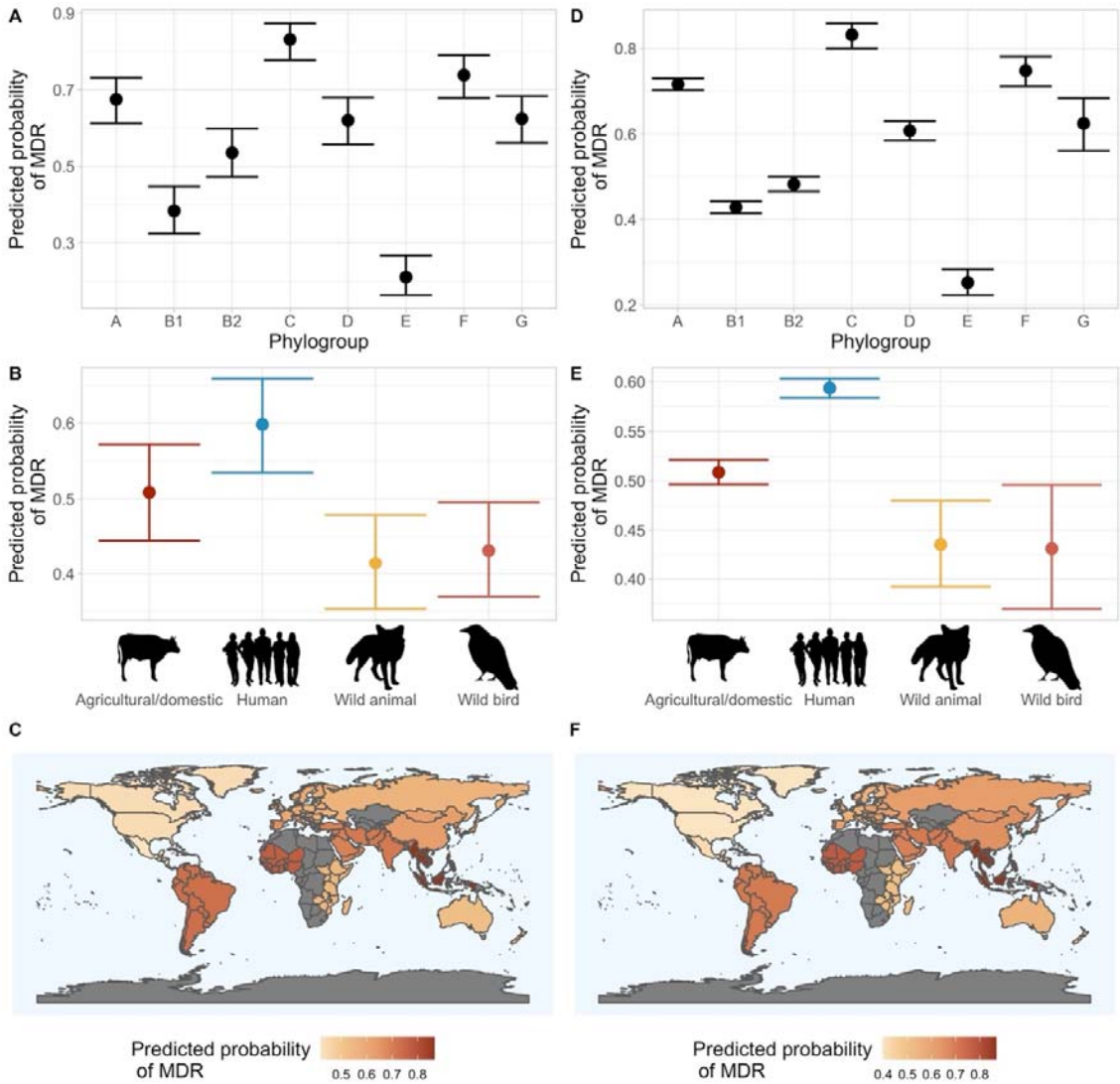


Figure 2 –Estimated mean probabilities and 95% CIs of MDR from binomial GLMs for datasets subsampled without replacement to the minimum group size for A) phylogroup, B) host category and C) geographic subregion. The same model types for the full non-subsampled dataset for D) phylogroup, E) host category and F) geographic subregion. CIs for subregion models can be found in Figure S6.

2. bla$_{CTX\text{-}M}$ *and carbapenemase genes*

The presence of the *bla*$_{\text{CTX-M}}$ class of β-lactamases was highly variable between groups (Fig. 3). Estimated probabilities ranged from 0.08 (95% CI = 0.066-0.11) in phylogroup E to 0.51 (95% CI=0.47-0.55) in phylogroup F, whilst *bla*$_{\text{CTX-M}}$ genes were more associated with human hosts (0.33, 95% CI=0.32-0.34) than other host categories. The probabilities of

bla$_{\text{CTX-M}}$ genes being present also varied widely between subregions, with the highest in South-Eastern Asia (0.59, 95% CI=0.53-0.64) and Eastern Europe (0.52, 95% CI=0.47-0.57), and all other subregions below 0.45.

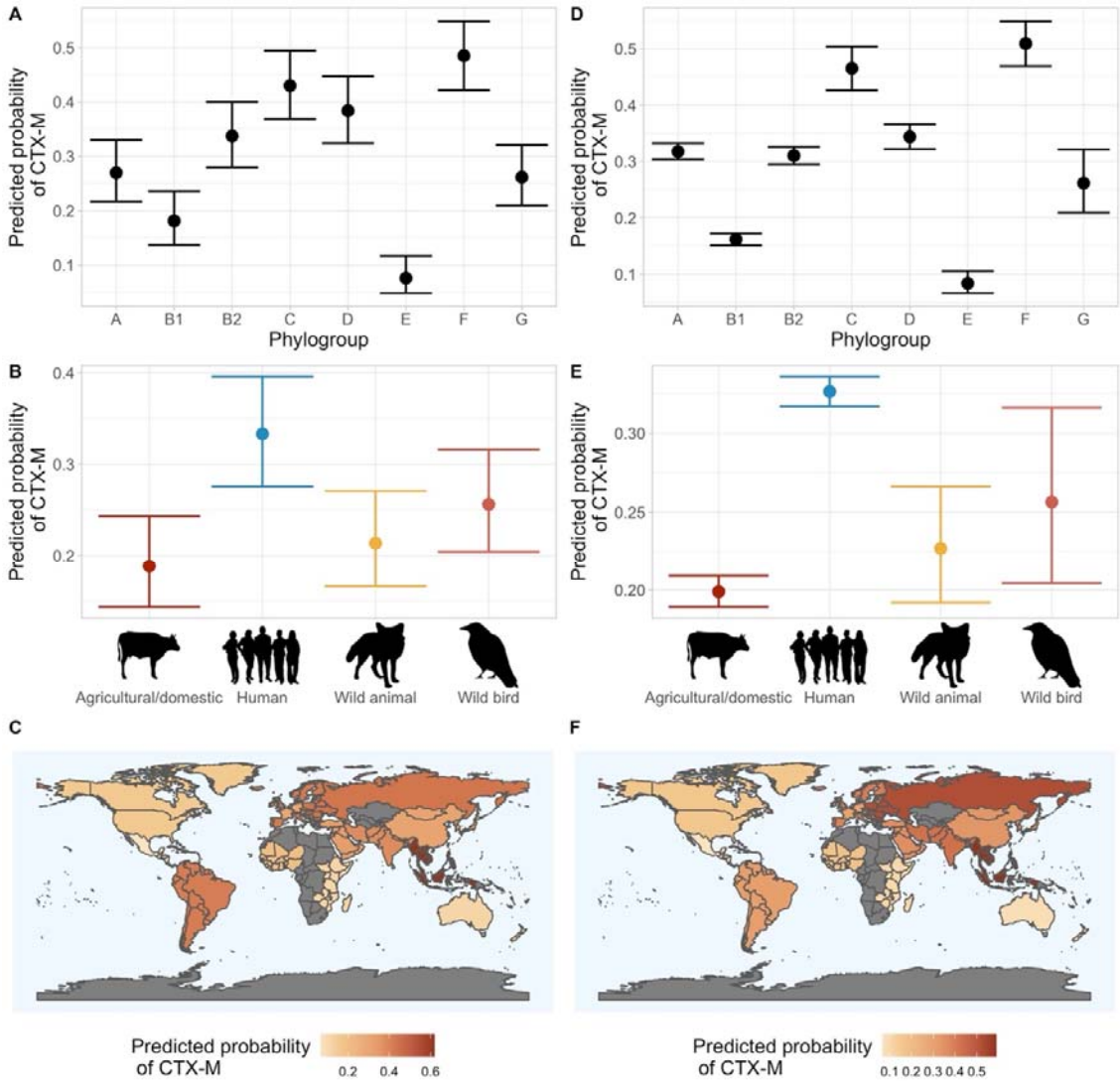


Figure 3 – Estimated mean probabilities and 95% CIs of *bla*$_{\text{CTX-M}}$ presence from binomial GLMs for datasets subsampled without replacement to the minimum group size for A) phylogroup, B) host category and C) geographic subregion. The same model types for the full non-subsampled dataset for D) phylogroup, E) host category and F) geographic subregion. CIs for subregion models can be found in Figure S12.

Carbapenamase genes were far less common than *bla*$_{\text{CTX-M}}$ genes (Fig. 4). There was a clear split between phylogroups C, F, A and D, which were more likely to possess carbapenemases (probabilities >0.15), and the remaining groups which had probabilities of 0.05 or below. The

highest proportions of carbapenemases were estimated in wild bird (0.25, 95% CI=0.20-0.31) and human (0.14, 95% CI= 0.13-0.15) hosts.

We further investigated sampling bias for carbapenemase-positive isolates in wild birds due to the high proportion of genomes possessing these genes. Isolates from wild birds possessing carbapenemases were only represented in 3 subregions: Australia and New Zealand (n=25), Eastern Asia (n=32) and Western Asia (n=1). Sequencing bias was evident; for example, 68 out of 234 wild bird isolates were from one study of silver gulls in Australia (which made up 68 out of the 93 isolates from this region). This study (PRJNA630096) specifically sought and sequenced isolates conferring resistance to critically important antimicrobials. Further carbapenemase-producing isolates came from studies in China (PRJNA669620 and PRJNA349231) that specifically studied carbapenem-resistant *E. coli*.

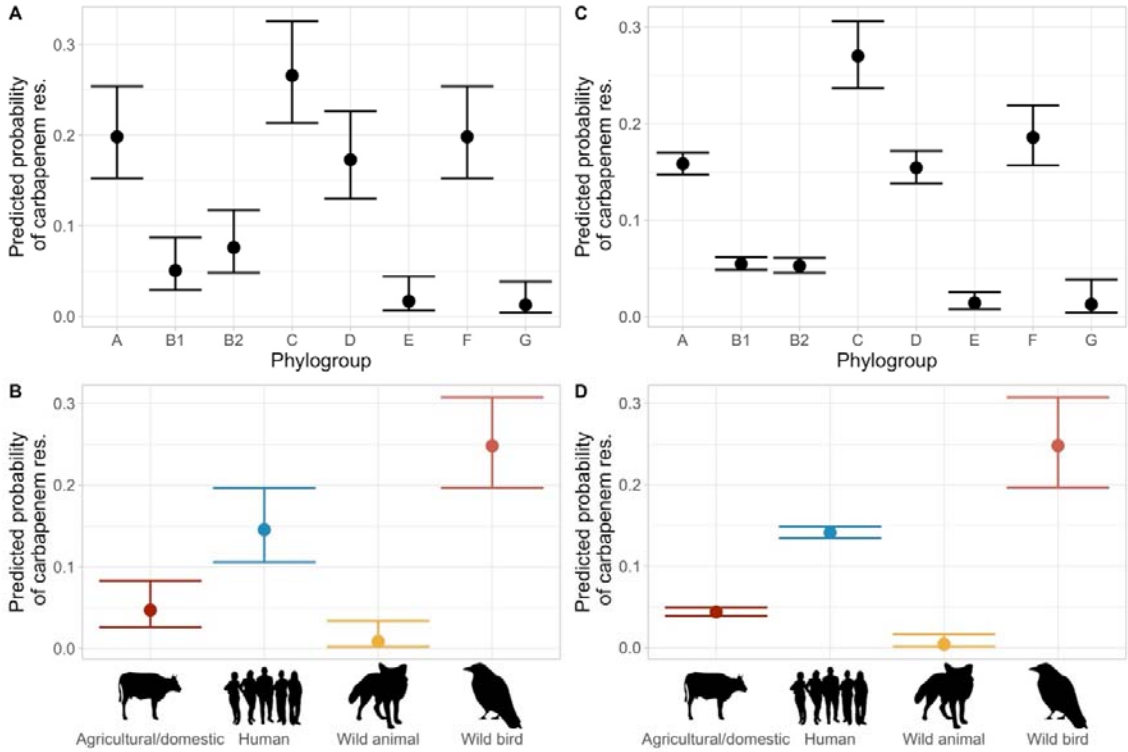


Figure 4 – Estimated mean probabilities and 95% CIs of carbapenemase presence from binomial GLMs for datasets subsampled without replacement to the minimum group size for A) phylogroup and B) host category. The same model types for the full non-subsampled dataset for C) phylogroup and D) host category.

Subsampling and resampling approaches do not alter major trends

Due to the large discrepancies in sample size between hosts, regions and phylogroups in this dataset, we used additional sampling with replacement to examine the sensitivity of the observed trends to changing sample sizes of 500, 1000, 2000 and 4000 (Fig. S4-S12). This showed that, although fluctuations in estimated proportions occurred (as well as expected reductions in CIs with increased sample size), broader trends were reproducible across wide variations in subsample size. In some smaller subsamples, insufficient genomes possessed the metric of AMR carriage (e.g. a carbapenemase gene) to make estimates in some subgroupings.

Discussion

Though we are in the era of large-scale genomics studies, with thousands of genomes available for species such as *E. coli*, it remains challenging to collate this information to answer questions about AMR prevalence. In this study, we characterised the RefSeq *E. coli* dataset for which metadata was available, investigating sampling bias and AMR carriage according to various metrics for different hosts, phylogroups and geographic subregions.

How evenly sampled are publicly-available E. coli *genomes?*

In first characterising this dataset, large sample size discrepancies in the locations and hosts *E. coli* genomes were sourced from were apparent. This was not unexpected. Previous large-scale work on *E. coli* genomes isolated from human hosts has emphasised the sampling biases towards clinically-important isolates and lineages, as well as high-income countries, even in the most well-sampled host species [13]. Other work has used literature searches to quantify how poorly sampled wild animals are for *E. coli*, with small within-group sample sizes in the few studies available [28]. Our data adds to this body of knowledge emphasising how little is known about ARG burden across the diversity of this species, despite the vast amount of sequencing data available.

We investigated the effects of subsampling all groups to the minimum sample size, as well as resampling up to larger sample sizes. This approach cannot replace unbiased sampling, particularly for those groups that are less well represented. However, it is interesting to note that estimates of AMR carriage for larger groups, such as phylogroup A or human hosts, did

not alter substantially regardless of the subsampling regime. Additionally, the spread of counts of ARGs was wide in these groups, indicating that they were not necessarily dominated by one or few lineages.

Previous work has also investigated the potential effect of sampling bias on AMR estimates using *E. coli* genomes from GenBank [29]. These authors found that 1) the average distance of a given newly sequenced *E. coli* genome is equivalent to that between known close relatives, such as O157:H7 genomes, and 2) the results of a population genetics analysis of AMR did not change substantially when all human or unknown source genomes were removed, or all genomes for which the GenBank record referenced antibiotic resistance. Therefore, it is possible that, for more well-sampled subgroupings of *E. coli*, our current sequence collection already spans a high diversity of this species.

Phylogroup distributions within hosts and subregions

Though the most common phylogroups were A, B1 and B2, their proportions varied substantially between hosts. Phylogroup B1 was the most frequent in agricultural and wild animals. This is broadly consistent with many smaller-scale studies for which B1 was the most frequent in agricultural settings. For example, B1 made up 71% of agriculture-associated isolates in a study in the Philippines [30], 50% of those in a study of sheep farming in China [31], and 35% of isolates from cattle and their attendants in Tanzania [32]. The wide variation in actual proportions observed in these different studies may be due to smaller sample sizes (n=17-100 in those referenced), as well as location and species effects. Human isolates were proportionally more likely to be B2 than those from other hosts. *E. coli* isolates from group B2 tend to possess more virulence traits than those from A and B1, though this virulence is proposed to be a by-product of commensalism [33]. Geographic differences in predominant commensal *E. coli* phylogroups have broadly represented a shift from A to B2 in industrialised nations over the past decades, with these being the most common phylogroups in humans [34]. We saw this trend to some extent – European regions and Australia and New Zealand had high proportions of B2. Contrastingly, we found that other regions such as North America and Eastern Asia had B1 as the predominant phylogroup.

Estimating AMR carriage across phylogroups, hosts and subregions

We used various metrics of AMR carriage to get an overview of ARG distribution in publicly-available *E. coli* genomes. When comparing phylogroups, A, C and F were repeatedly associated with more ARGs as well as clinically-important β-lactamases. Phylogroup A is a major group that appears to be more generalist. It is spread across vertebrate hosts [8], and contains mostly human commensal strains [35] as well as laboratory strains [36]. Although the presence of engineered strains may partly explain the high AMR metrics in this group, our results also imply that ARGs may be frequent in commensal lineages. Phylogroups commonly present in the human gut could be a reservoir for resistance genes and plasmids [37], potentially allowing their horizontal spread into other, more problematic lineages or species.

As they are minor phylogroups, comparatively less is known about groups C and F. In terms of pathogenic potential, the sequence type complex 88 (STc88) lineage of group C is one of the main avian pathogenic *E. coli* (APEC) strains [38]. This phylogroup also includes ST410, a recently-emerged MDR lineage with fluoroquinolone resistance, ESBL ($bla_{\text{CTX-M-15}}$) and carbapenemase ($bla_{\text{OXA-181}}$ and $bla_{\text{NDM-5}}$) genes [38]. Meanwhile, phylogroup F has been associated with fewer virulence traits than closely related groups [39], though it has also been linked to extraintestinal pathogenic (ExPEC) infections [38] and contains the MDR STc648 lineage [40]. Therefore, these phylogroups could represent highly virulent and/or drug-resistant lineages and may warrant further investigation.

Unsurprisingly, human hosts were associated with more ARGs than other host categories. However, high levels were also seen in agricultural and domestic animal species for some measures, such as MDR. In addition, $bla_{\text{CTX-M}}$ and carbapenemase genes were detected in all the animal host categories to some degree. Antibiotics are still extensively used in animal agriculture, selecting for ARGs both directly in the gut and through excretion of antibiotics into soil and the local environment [41]. Despite this there is no standardised global surveillance system used in animal agriculture that is equivalent to those used in humans [41]. There is evidence for transmission of $bla_{\text{CTX-M}}$ variants in *E. coli* from humans to farmed animals and the environment, which also occurs in the reverse direction via food products [42]. Only limited studies document the occurrence of carbapenemases in *E. coli* strains from wildlife, agricultural animals and soils, though they have been reported in these contexts [43, 44]. This highlights the need for further prevalence studies to determine the extent to which these genes are spreading in different environments.

The proportionally high numbers of carbapenemase-possessing isolates (and potentially XDR isolates) from wild birds in this work are likely due to studies that enriched for ARGs. However, this emphasizes our lack of knowledge on the true extent of AMR spread in the natural world. High levels of problematic ARGs in wild birds, particularly generalist species, could occur due to their relatively high mobility compared to other taxa, and their utilisation of diverse foraging habitats, both of which increase potential exposure to anthropogenic sources of resistant bacteria [45, 46].

Finally, estimates of ARG prevalence varied geographically depending on the metric used. The highest AMR carriage was generally estimated for Eastern and South-Eastern Asia across measures. However, other regions were particularly high for individual measures. For example, Eastern Europe had high levels of $bla_{\text{CTX-M}}$ genes and MDR genomes were more common in Western Africa. These trends could either reflect genuine differences in AMR levels and individual gene prevalence, or regional differences in sampling strategies and genomic surveillance. There are large gaps in AMR surveillance for low and middle income (LMIC) countries. A recent study utilised the relationship between socioeconomic characteristics and AMR prevalence to model AMR levels for underrepresented countries, estimating high levels of third-generation cephalosporin resistance in *E. coli* from Western Asia [47]. Genomic surveillance has been leveraged during the COVID-19 pandemic and for other infectious diseases, but next-generation sequencing capacity is low in many regions [48]. However, the development of more many sequence analysis tools to detect ARGs [49], as well as developing capacity to incorporate whole genome sequencing into AMR surveillance in LMICs [50], may lead to this becoming a powerful tool in the future.

Public genomic databases – limitations, benefits and future work

We show the void in sequencing data from *E. coli* isolates outside of human and agricultural settings in a very large dataset, as well as stark regional disparities, with scarce publicly-available data representing the majority of global regions. Whilst this study characterises this currently available dataset, representative estimates of ARG burden in *E. coli* sequences cannot be made without systematic, representative sampling. Notably, it is concerning that ARGs that inactivate our last-resort antibiotics are being detected and potentially spread in any environmental context, and that we do not have sufficient genomic data to investigate

this. Despite their limitations, publicly-available datasets are important, allowing access to sequencing data for scientists and the public across the globe, as well as improving reproducibility. Future systematic sampling and genomic surveillance in previously under-investigated settings will give a more accurate picture of the scale of the AMR problem.

Data Summary

All code used to generate and analyse data are available at https://github.com/elliekpursey/AMR-Ecoli.

Author contributions

EP, SVH and ERW conceptualised the study. WHG refined the directions of the analysis and methodology. EP performed all data and statistical analysis and wrote the first manuscript draft. All authors contributed to interpretation of the results and editing subsequent versions of the manuscript.

Conflicts of interest

We declare no conflicts of interest.

Acknowledgements and funding

EP was supported by a PhD studentship equally funded by a grant from the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC-STG-2016-714478 to ERW) and the College of Life and Environmental Sciences, University of Exeter, UK. EP thanks Liam Langley for statistical modelling advice, and Elze Hesse and Kate Baker for feedback on the analyses. TD is supported by European Research Council (ERC-2017-ADG-788405 to ERW). SVH acknowledges BBSRC for funding (BB/R010781/1 and BB/S017674/1).

References

1. **Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, *et al.*** Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. *Proc Natl Acad Sci* 2011;108:7200–7205.

2. **Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, *et al.*** Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 2022;399:629–655.

3. **World Health Organization**. *Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis*. 9789240026438 (electronic version); Geneva: World Health Organization; 2017.

4. **Hall RJ, Whelan FJ, Cummins EA, Connor C, McNally A, *et al.*** Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility. *Microb Genomics* 2021;7:000650.

5. **Von Wintersdorff CJH, Penders J, Van Niekerk JM, Mills ND, Majumder S, *et al.*** Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front Microbiol* 2016;7:1–10.

6. **Zhang A-N, Gaston JM, Dai CL, Zhao S, Poyet M, *et al.*** An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat Commun* 2021;12:4765.

7. **Woolhouse M, Ward M, van Bunnik B, Farrar J**. Antimicrobial resistance in humans, livestock and the wider environment. *Philos Trans R Soc B Biol Sci* 2015;370:20140083–20140083.

8. **Touchon M, Perrin A, Sousa JAM de, Vangchhia B, Burn S, *et al.*** Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLOS Genet* 2020;16:e1008866.

9. **Economou V, Gousia P**. Agriculture and food animals as a source of antimicrobial-resistant bacteria. *Infect Drug Resist* 2015;8:49–61.

10. **Arnold KE, Williams NJ, Bennett M**. 'Disperse abroad in the land': the role of wildlife in the dissemination of antimicrobial resistance. *Biol Lett* 2016;12:20160137.

11. **Robinson TP, Bu DP, Carrique-Mas J, Fèvre EM, Gilbert M, *et al.*** Antibiotic resistance is the quintessential One Health issue. *Trans R Soc Trop Med Hyg* 2016;110:377–380.

12. **Petitjean M, Condamine B, Burdet C, Denamur E, Ruppé E**. Phylum barrier and Escherichia coli intra-species phylogeny drive the acquisition of antibiotic-resistance genes. *Microb Genomics* 2021;7:000489.

13. **Horesh G, Blackwell GA, Tonkin-Hill G, Corander J, Heinz E, *et al.*** A comprehensive and high-quality collection of Escherichia coli genomes and their genes. *Microb Genomics* 2021;7:000499.

14. **Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, *et al.*** AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* 2021;11:12728.

15. **Clermont O, Christenson JK, Denamur E, Gordon DM**. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 2013;5:58–65.

16. **Waters NR, Abram F, Brennan F, Holmes A, Pritchard L**. Easy phylotyping of Escherichia coli via the EzClermont web app and command-line tool. *Access Microbiol*.

17. **Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, *et al.*** *Sustainable data analysis with Snakemake*. 10:33; F1000Research; 18 January 2021.

18. **Magiorakos A-P, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, *et al.*** Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect* 2012;18:268–281.

19. **Skarżyńska M, Zaja□c M, Bomba A, Bocian Ł, Kozdruń W, *et al.*** Antimicrobial Resistance Glides in the Sky—Free-Living Birds as a Reservoir of Resistant Escherichia coli With Zoonotic Potential. *Front Microbiol*;12. https://www.frontiersin.org/articles/10.3389/fmicb.2021.656223 (2021, accessed 4 July 2022).

20. **Wickham H, Averick M, Bryan J, Chang W, McGowan L, *et al.*** Welcome to the Tidyverse. *J Open Source Softw* 2019;4:1686.

21. **Mills BR**. MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art.

22. **Pebesma E**. Simple Features for R: Standardized Support for Spatial Vector Data. *R J* 2018;10:439.

23. **Brooks M E, Kristensen K, Benthem K J ,van, Magnusson A, Berg C W, *et al.*** glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *R J* 2017;9:378.

24. **Hartig F**. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. http://florianhartig.github.io/DHARMa/ (2020).

25. **Lüdecke D**. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *J Open Source Softw* 2018;3:772.

26. **Sullivan GM, Feinn R**. Using Effect Size—or Why the P Value Is Not Enough. *J Grad Med Educ* 2012;4:279–282.

27. **Berner D, Amrhein V**. Why and how we should join the shift from significance testing to estimation. *J Evol Biol*;n/a.

28. **Lagerstrom KM, Hadly EA**. The under-investigated wild side of Escherichia coli: genetic diversity, pathogenicity and antimicrobial resistance in wild animals. *Proc R Soc B Biol Sci* 2021;288:20210399.

29. **Goldstone RJ, Smith DGE**. A population genomics approach to exploiting the accessory 'resistome' of Escherichia coli. *Microb Genomics* 2017;3:e000108.

30. **Zara ES, Vital PG**. Phylogroup typing and carbapenem resistance of Escherichia coli from agricultural samples in Metro Manila, Philippines. *J Environ Sci Health Part B* 2022;57:644–656.

31. **Zhao X, Lv Y, Adam FEA, Xie Q, Wang B, *et al.*** Comparison of Antimicrobial Resistance, Virulence Genes, Phylogroups, and Biofilm Formation of Escherichia coli Isolated From Intensive Farming and Free-Range Sheep. *Front Microbiol*;12. https://www.frontiersin.org/articles/10.3389/fmicb.2021.699927 (2021, accessed 19 July 2022).

32. **Madoshi BP, Kudirkiene E, Mtambo MMA, Muhairwa AP, Lupindu AM, *et al.*** Characterisation of Commensal Escherichia coli Isolated from Apparently Healthy Cattle and Their Attendants in Tanzania. *PLOS ONE* 2016;11:e0168160.

33. **Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, *et al.*** Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group Escherichia coli Strains. *Mol Biol Evol* 2007;24:2373–2384.

34. **Tenaillon O, Skurnik D, Picard B, Denamur E**. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* 2010;8:207–217.

35. **Smati M, Clermont O, Le Gal F, Schichmanoff O, Jauréguy F, *et al.*** Real-Time PCR for Quantitative Analysis of Human Commensal Escherichia coli Populations Reveals a High Frequency of Subdominant Phylogroups. *Appl Environ Microbiol* 2013;79:5005–5012.

36. **Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, *et al.*** Mash-based analyses of Escherichia coli genomes reveal 14 distinct phylogroups. *Commun Biol* 2021;4:1–12.

37. **Hu Y, Yang X, Qin J, Lu N, Cheng G, *et al.*** Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun* 2013;4:2151.

38. **Denamur E, Clermont O, Bonacorsi S, Gordon D**. The population genetics of pathogenic Escherichia coli. *Nat Rev Microbiol*. Epub ahead of print 21 August 2020. DOI: 10.1038/s41579-020-0416-x.

39. **Vangchhia B, Abraham S, Bell JM, Collignon P, Gibson JS, *et al.*** Phylogenetic diversity, antimicrobial susceptibility and virulence characteristics of phylogroup F Escherichia coli in Australia. *Microbiology* 2016;162:1904–1912.

40. **Johnson JR, Johnston BD, Gordon DM**. Rapid and Specific Detection of the Escherichia coli Sequence Type 648 Complex within Phylogroup F. *J Clin Microbiol* 2017;55:1116–1121.

41. **Jadeja NB, Worrich A**. From gut to mud: dissemination of antimicrobial resistance between animal and agricultural niches. *Environ Microbiol* 2022;24:3290–3306.

42. **Bevan ER, Jones AM, Hawkey PM**. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *J Antimicrob Chemother* 2017;72:2145–2155.

43. **Furlan JPR, Sellera FP, Stehling EG**. Strengthening genomic surveillance of carbapenemases in soils: a call for global attention. *Lancet Microbe*;0. Epub ahead of print 23 March 2023. DOI: 10.1016/S2666-5247(23)00093-9.

44. **Köck R, Daniels-Haardt I, Becker K, Mellmann A, Friedrich AW, *et al.*** Carbapenem-resistant Enterobacteriaceae in wildlife, food-producing, and companion animals: a systematic review. *Clin Microbiol Infect* 2018;24:1241–1250.

45. **Navarro J, Grémillet D, Afán I, Miranda F, Bouten W, *et al.*** Pathogen transmission risk by opportunistic gulls moving across human landscapes. *Sci Rep* 2019;9:10659.

46. **Migura-Garcia L, Ramos R, Cerdà-Cuéllar M**. Antimicrobial Resistance of *Salmonella* Serovars and *Campylobacter* spp. Isolated from an Opportunistic Gull Species, Yellow-legged Gull (*Larus michahellis*). *J Wildl Dis* 2017;53:148–152.

47. **Oldenkamp R, Schultsz C, Mancini E, Cappuccio A**. Filling the gaps in the global prevalence map of clinical antimicrobial resistance. *Proc Natl Acad Sci* 2021;118:e2013515118.

48. **Inzaule SC, Tessema SK, Kebede Y, Ogwell Ouma AE, Nkengasong JN**. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *Lancet Infect Dis* 2021;21:e281–e289.

49. **Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, *et al.*** Using Genomics to Track Global Antimicrobial Resistance. *Front Public Health*;7. https://www.frontiersin.org/articles/10.3389/fpubh.2019.00242 (2019, accessed 22 May 2023).

50. **NIHR Global Health Research Unit on Genomic Surveillance of AMR**. Whole-genome sequencing as part of national and international surveillance programmes for antimicrobial resistance: a roadmap. *BMJ Glob Health* 2020;5:e002244.