







# *Escherichia coli* phylogeny drives co-amoxiclav resistance through variable expression of TEM-1 beta-lactamase

Received: 6 November 2024

Accepted: 27 August 2025

Published online: 30 September 2025



William Matlock<sup>1,2,3,7</sup> , Gillian Rodger<sup>2,3,7</sup>, Emma Pritchard<sup>2,3</sup>, Matthew Colpus<sup>3</sup> , Natalia Kapel<sup>3</sup> , Lucinda Barrett<sup>4</sup>, Marcus Morgan<sup>4</sup>, Sarah Oakley<sup>4</sup>, Katie L. Hopkins<sup>5</sup>, Aysha Roohi<sup>2,3</sup>, Drosos Karageorgopoulos<sup>4</sup> , Matthew B. Avison<sup>6</sup>, A. Sarah Walker<sup>2,3</sup>, Samuel Lipworth<sup>2,3,4,8</sup>  & Nicole Stoesser<sup>2,3,4,8</sup> 

Co-amoxiclav (amoxicillin and clavulanate) is a commonly used combination antibiotic, with resistance in *Escherichia coli* associated with increased mortality. The class A beta-lactamase *bla*<sub>TEM-1</sub> is often carried by resistant *E. coli* but exhibits high phenotypic heterogeneity, complicating genotype-phenotype predictions. We curated a dataset of  $n = 377$  diverse *E. coli* isolates where the only acquired beta-lactamase was *bla*<sub>TEM-1</sub>. We generated hybrid assemblies and co-amoxiclav minimum inhibitory concentrations (MICs), and *bla*<sub>TEM-1</sub> qPCR expression data for a subset ( $n = 67/377$ ). We first tested whether intrinsic expression of *bla*<sub>TEM-1</sub> varied between *E. coli* lineages, for example, from regulatory system differences, which are challenging to genomically quantify. Using genotypic features, we built a hierarchical Bayesian model for *bla*<sub>TEM-1</sub> expression, controlling for phylogeny. Expression varied across the phylogeny, with some lineages (phylogroups B1 and C, ST12) expressing *bla*<sub>TEM-1</sub> more than others (phylogroups E and F, ST372). Next, we built a second model to predict isolate MIC from genotypic features, again controlling for phylogeny. Phylogeny alone shifted MIC past the clinical breakpoint in 19% (55/292) of isolates with greater-than-chance probability, mostly representing ST12, ST69 and ST127. A third causal model confirmed that phylogenetic influence on *bla*<sub>TEM-1</sub> expression drove variation in MIC. We speculate that intergenic variation underlies this effect.

The class A beta-lactamase *bla*<sub>TEM-1</sub> was first identified in 1965 in a clinical *Escherichia coli* isolate<sup>1</sup>. Originally, it was mobilised by two of the earliest named transposons, Tn2 and Tn3, located on plasmids<sup>2</sup>. In the decades since, the genetic context of *bla*<sub>TEM-1</sub> has evolved<sup>3</sup>, and

other mobile genetic elements such as IS26<sup>4</sup> and a diverse array of plasmids<sup>5</sup> contribute to its dissemination. At the time of writing, NCBI contains over 220,000 unique isolates carrying *bla*<sub>TEM-1</sub> distributed across 26 genera, including the common clinical pathogens

<sup>1</sup>Department of Biology, University of Oxford, Oxford, UK. <sup>2</sup>Nuffield Department of Medicine, University of Oxford, Oxford, UK. <sup>3</sup>The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, Oxford, UK. <sup>4</sup>Oxford University Hospitals NHS Foundation Trust, Oxford, UK. <sup>5</sup>UK Health Security Agency, Colindale, UK. <sup>6</sup>School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK. <sup>7</sup>These authors contributed equally: William Matlock, Gillian Rodger. <sup>8</sup>These authors jointly supervised this work: Samuel Lipworth, Nicole Stoesser. ✉e-mail: [william.matlock@biology.ox.ac.uk](mailto:william.matlock@biology.ox.ac.uk); [nicole.stoesser@ndm.ox.ac.uk](mailto:nicole.stoesser@ndm.ox.ac.uk)

*Escherichia coli*, *Klebsiella pneumoniae* and *Acinetobacter baumannii*<sup>6</sup>. The emergence and dissemination of beta-lactam resistance has been a major healthcare challenge<sup>7</sup>, and *bla*<sub>TEM-1</sub> represents a key example.

In the UK, beta-lactam and beta-lactamase inhibitor combinations such as co-amoxiclav (amoxicillin and clavulanate), are commonly used as a first-line treatment for severe infections<sup>8</sup>. For Enterobacteriales, the current EUCAST co-amoxiclav minimum inhibitory concentration (MIC) clinical breakpoint for resistance is 8/2 µg/mL<sup>9</sup> across all indications, with a recent study concluding that empiric co-amoxiclav treatment of *E. coli* bacteraemia with MICs >32/2 µg/mL was associated with significantly higher mortality<sup>10</sup>. However, the carriage of *bla*<sub>TEM-1</sub> is associated with high phenotypic heterogeneity, making genotype-phenotype predictions challenging.

Small-scale, experimental *bla*<sub>TEM-1</sub> systems have demonstrated that the interplay of location (plasmid or chromosome) and copies in the genome, through varying dosage, contributes to variable resistance<sup>11,12</sup>. In addition, other determinants such as mutations in the promoter of *bla*<sub>TEM-1</sub><sup>13</sup> and the chromosomally intrinsic *ampC* gene<sup>14</sup>, and efflux pumps<sup>15</sup>, are associated with *E. coli* beta-lactam resistance. Moreover, different regulatory systems<sup>16,17</sup>, epistasis (interaction between genes)<sup>18,19</sup> and epigenetics (heritable phenotypic changes without alterations to the underlying DNA sequence)<sup>20</sup>, might also influence co-amoxiclav resistance. For example, five different *E. coli* strains carrying the same pLL35 plasmid (which carries *bla*<sub>CTX-M-15</sub> and *bla*<sub>TEM-112</sub>) varied in cefotaxime resistance<sup>21</sup>. Likewise, the introduction of a pOXA-48, a common conjugative plasmid in carbapenem-resistant clinical Enterobacteriales, to six different *E. coli* strains resulted in variable co-amoxiclav resistance<sup>22</sup>. This indicates that strain background plays a role in resistance.

Successful genotype-to-phenotype prediction requires a comprehensive understanding of not only individual resistant determinants but also their combined effects. Moreover, this understanding must be translated to clinically relevant pathogens. Yet, to accurately model resistance in these systems, a large sample of linked genomic and phenotypic data is required, which until recently has been limited by sequencing technology and costs.

In this study, we curated and completely reconstructed the genomes of 377 clinical *E. coli* bacteraemia isolates to reflect a real-world but relatively simple genetic scenario where the only acquired beta-lactamase gene identified was *bla*<sub>TEM-1</sub>, all identical at the amino acid level. We quantified the co-amoxiclav MICs for these isolates and generated *bla*<sub>TEM-1</sub> qPCR expression data for a subset. We then modelled *bla*<sub>TEM-1</sub> expression and co-amoxiclav MIC whilst controlling for confounding genetic mechanisms and chromosomal phylogeny, and characterised differences in intergenic content between lineages that may be contributing to differential phenotypic effects.

## Results

### A curated dataset of *E. coli* isolates with hybrid assemblies and co-amoxiclav MICs

We began with *n* = 548 candidate *E. coli* isolates, which following hybrid assembly, were curated into a final dataset of *n* = 377/548 (see “Methods” and Supplementary Information). In total, 77% (291/377) of assemblies were complete (all contigs were circularised), with the remaining 27% (86/377) having at least a circularised chromosome to confidently distinguish between chromosomal and plasmid-associated *bla*<sub>TEM-1</sub>. Assemblies contained median = 3 (IQR = 2–5) plasmid contigs.

We identified *n* = 451 *bla*<sub>TEM-1</sub> genes on 431 contigs (13% [58/431] chromosomal versus 87% [373/431] plasmid). Isolates carried a median = 1 copy of *bla*<sub>TEM-1</sub> (range = 1–6). Carrying more than one copy of *bla*<sub>TEM-1</sub> on a single contig was rare: of all *bla*<sub>TEM-1</sub>-positive contigs, 97% (400/412) versus 3% (12/412) had no duplications versus at least one. The *bla*<sub>TEM-1</sub> genes had synonymous single-nucleotide polymorphisms (SNPs) in positions 18, 138, 228, 396, 474, 705 and 717, totalling *n* = 7 single-nucleotide variant (SNV) profiles across the replicons, yet

diversity was dominated by *bla*<sub>TEM-1b</sub> at 73% (329/451; SNV profile TATTTCC; see “Methods”) <sup>23</sup>. Where *bla*<sub>TEM-1</sub>-positive contigs had at least two copies of *bla*<sub>TEM-1</sub>, they were almost always the same SNV duplicated (11/12).

By examining the genomic arrangement of *bla*<sub>TEM-1</sub> (namely the replicons it was found on as well as any copies), we found most isolates carried a single non-chromosomal copy (73% [277/377]; Fig. 1a, b). More generally, whilst the plasmid contigs totalled only 3.6% of the total sequence length (bp) across the assemblies (71,126,646 bp/1,969,804,202 bp), they carried 85.8% of the *bla*<sub>TEM-1</sub> genes (387/451). Such *bla*<sub>TEM-1</sub>-carrying plasmids were represented across the *E. coli* phylogeny (Fig. 1c). Overall, the dataset comprised 5.6% (21/377) phylogroup A, 8.0% (30/377) B1, 48.5% (183/377) B2, 4.8% (18/377) C, 27.3% (103/377) D, 0.5% (2/377) E, 4.2% (16/377) F and 1.1% (4/377) G. In total, we manually corrected *n* = 5 EzClermont phylogroup classifications using a chromosomal core gene phylogeny (see “Methods”): OXEC-108 (G to D), OXEC-317 (B2 to D), OXEC-333 (U to B1), OXEC-344 (U to B1) and OXEC-406 (U to B1). The EzClermont publication presented a 98.4% (123/125) true-positive rate on their validation set, which is in line with our 98.7% (372/377) <sup>24</sup>.

Five known upstream promoters modulate the expression of *bla*<sub>TEM-1</sub>: *P*<sub>3</sub>, *P*<sub>a/Pb</sub>, *P*<sub>4</sub>, *P*<sub>5</sub> and *P*<sub>c/Pd</sub> <sup>13,25</sup>. We linked 91% (409/451) *bla*<sub>TEM-1</sub> genes to a promoter immediately upstream, of which a majority, 64% (262/409), were identical to the *P*<sub>3</sub> reference. More generally, excluding *n* = 2 different *P*<sub>c/Pd</sub>-like promoters which have large deletions, we identified SNPs in positions 32, 43, 65, 141, 162 and 175, totalling *n* = 8 SNV profiles (by Sutcliffe numbering<sup>26</sup>). Notably, 15% (39/262) of promoters had the *P*<sub>a/Pb</sub>-associated C32T mutation, which produces two overlapping promoter sequences. Figure S1 visualises the joint distribution of isolate phylogroup and linked *bla*<sub>TEM-1</sub> promoter SNV.

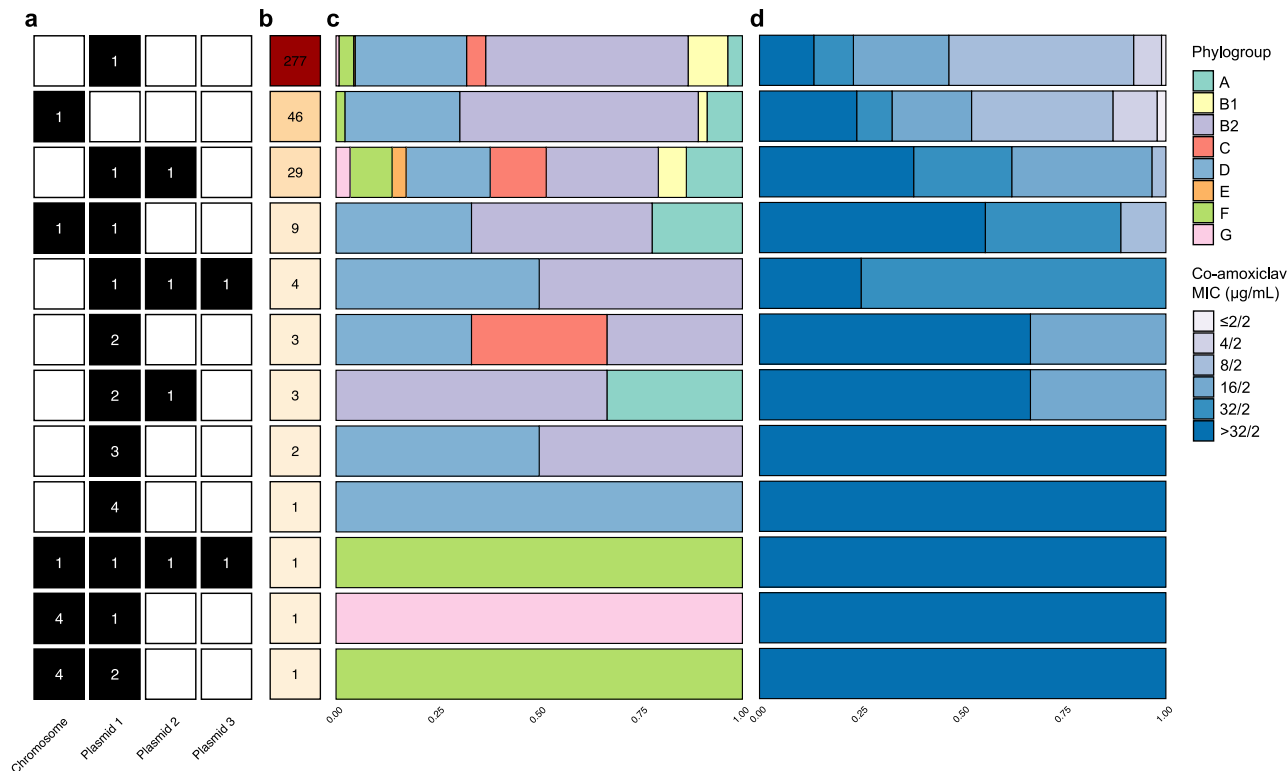
Isolates were associated with a diverse range of co-amoxiclav MICs (µg/mL; ≤2/2 [4 (1.1%)], 4/2 [24 (6.4%)], 8/2 [144 (38.2%)], 16/2 [86 (22.8%)], 32/2 [44 (11.7%)] and >32/2 [75 (19.9%)] (Fig. 1d; see “Methods”). Figure S2 visualises the joint distribution of isolate phylogroup and MIC. Figure S3 visualises the distribution of *bla*<sub>TEM-1</sub> genome and cell copy number, *bla*<sub>TEM-1</sub> expression, and co-amoxiclav MIC against the chromosomal core gene phylogeny.

### *bla*<sub>TEM-1</sub> associated with conjugative plasmids

Whilst chromosomal copies of *bla*<sub>TEM-1</sub> can remain with a lineage over time, plasmidic copies might come and go. This could give the host cell access to a transient boost in resistance without impeding long-term fitness.

Confining the analysis to circularised plasmids (1036/1512), in silico replicon typing revealed the most common plasmid families to be ColRNAI-like, Col156-like, B/O/K/Z-like and Col(MG828)-like at 11% (117/1036), 7.4% (77/1036), 6.8% (70/1036) and 5.2% (54/1036), respectively. All other plasmid families had fewer than 50 representatives. Figure S4 visualises the relationship between strain background (sequence type and phylogroup) and replicon types (PlasmidFinder output; see “Methods”) for the circularised plasmids. Using a 2-sample test for equality of proportions with continuity correction, the *bla*<sub>TEM-1</sub>-positive plasmids (333/1036) were significantly more likely to be putatively conjugative (85.9% [286/333]) compared to the *bla*<sub>TEM-1</sub>-negative plasmids (28.4% [200/703]; prop = 200/703;  $\chi^2 = 297.02$ , df = 1, *p*-value < 2.2e-16). Moreover, for the most common genomic arrangement of *bla*<sub>TEM-1</sub> (i.e., a single plasmid [277/377]), amongst circularised plasmids (252/277), 90% were putatively conjugative (227/252).

For a genome carrying *bla*<sub>TEM-1</sub> on the chromosome, the gene's copy number and total number of genes in the genome are equivalent. For a genome carrying *bla*<sub>TEM-1</sub> on a plasmid, this might not be the case. This is because plasmids can exist as multiple copies. The calculated copy number of all plasmidic contigs (*n* = 1512) was median = 3.13



**Fig. 1 | A genotypically and phenotypically heterogeneous population of *bla*<sub>TEM-1</sub>-carrying *E. coli*.** **a** The genomic arrangement of *bla*<sub>TEM-1</sub> in the genomes ordered in descending **(b)** frequency. **c** Phylogroup and **d** co-amoxiclav MIC distribution for each genomic arrangement.

(range=0.04–57.00). Of these,  $n = 19/1512$  contigs (with 6/19 circularised) had calculated copy numbers less than one (see “Methods”). This was potentially due to uneven short-read coverage. However, none carried *bla*<sub>TEM-1</sub> so were not used in the later modelling. Taking the circularised plasmids with copy number at least one (1030/1512), longer plasmids (> 10kbp; 668/1030) were generally low copy number (median = 2.36), whilst shorter plasmids (≤ 10kbp; 362/1030) were generally high copy number (median = 11.01, see Fig. S5), consistent with previous studies<sup>27</sup>.

### *E. coli* phylogeny shapes *bla*<sub>TEM-1</sub> expression

Within different *E. coli* lineages, *bla*<sub>TEM-1</sub> and its promoter are potentially subject to different regulatory systems and epigenetic interactions, which may in turn affect *bla*<sub>TEM-1</sub> expression. To test this, we first selected a random subsample of  $n = 67/377$  isolates with a single copy of *bla*<sub>TEM-1</sub> in the genome, either on a chromosome (15/67) or a plasmid (52/67). Moreover, we only selected isolates with zero, one, or two mutations in the *bla*<sub>TEM-1</sub> promoter sequence: C32T, a well-studied but rare mutation which produces two overlapping promoters and is known to increase expression<sup>13</sup>, and G175A, a less studied but common mutation in our dataset (according to Sutcliffe numbering based on the PBR322 plasmid<sup>26</sup>; see Table 1). Focussing on only two mutations enabled us to statistically explore the effect of their interaction. Isolates were distributed across the entire *E. coli* phylogeny (phylogroup A [6/67], B1 [4/67], B2 [40/67], C [4/67], D [8/67], E [1/67] and F ([4/67])). We then performed qPCR to evaluate for *bla*<sub>TEM-1</sub> expression (see “Methods”). Every isolate had at least two replicates (2 [48/67], 4 [1/67], or 9 [18/67]), giving a total of  $n = 262$  *bla*<sub>TEM-1</sub> ΔCt observations (TEM-1 Ct – 16S Ct; see “Methods”) for modelling.

To test for the effects of *E. coli* lineage, we built a maximum likelihood core gene phylogeny for all  $n = 377$  chromosomes (see “Methods”). In total, we identified 17,836 gene clusters, of which 18.7% (3342/17,836) were core genes (those found in ≥98% of chromosomes). The phylogeny (midpoint-rooted and restricted to the  $n = 67/377$

**Table 1 | Replicon distribution of  $n = 67$  *bla*<sub>TEM-1</sub> promoter variants**

| Promoter SNV/ Replicon | Chromosome | Plasmid | Total |
|------------------------|------------|---------|-------|
| CG (wildtype)          | 1          | 22      | 23    |
| G175A                  | 2          | 21      | 23    |
| C32T                   | 3          | 5       | 8     |
| C32T, G175A            | 9          | 4       | 13    |
| Total                  | 15         | 52      | 67    |

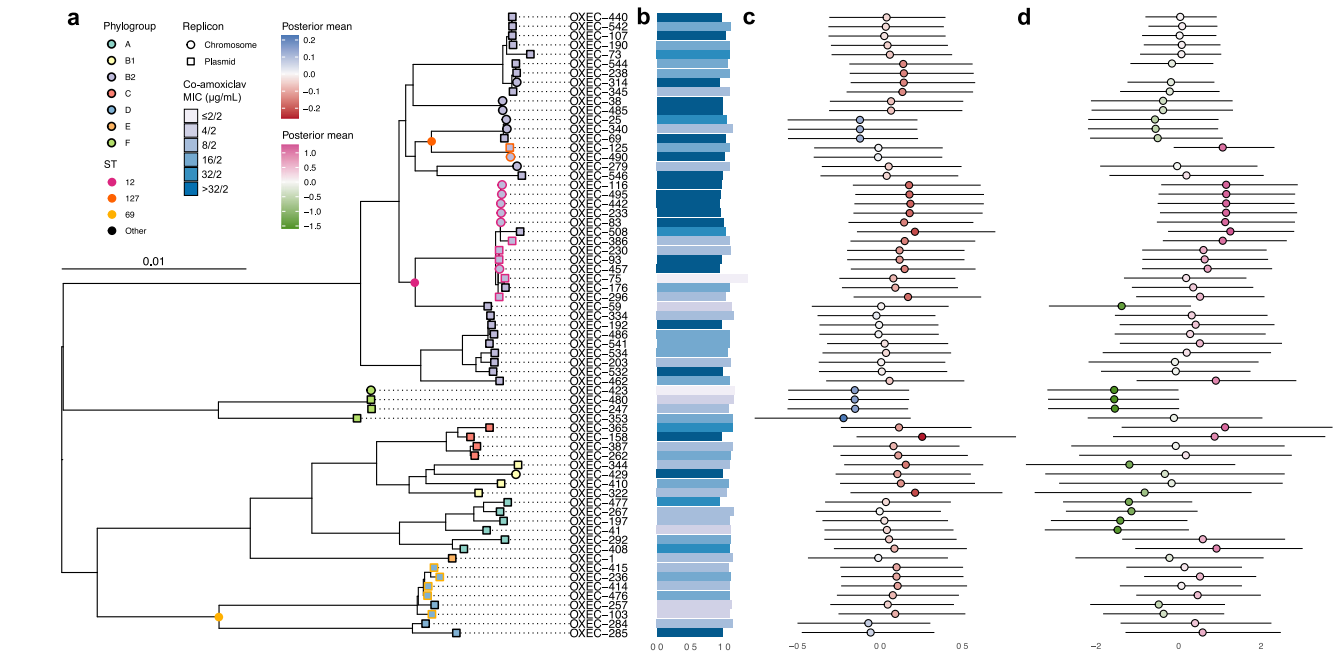
Single-nucleotide variants (SNVs) are for the 32<sup>nd</sup> and 175<sup>th</sup> positions by Sutcliffe numbering<sup>26</sup>.

isolates in the expression analysis) is given in Fig. 2a. Using  $b = 1000$  ultrafast bootstraps, all phylogroup node supports were 100%, and more generally, 76.7% (287/374) of internal node supports were 100%, and 87.4% (327/374) were at least 95% (see “Methods”). Moreover, the Robinson-Foulds distance between the ML tree and consensus tree was 4, indicating nearly identical topology.

Briefly, the expression linear mixed model employed Markov Chain Monte Carlo (MCMC) to estimate parameters. The response variable *bla*<sub>TEM-1</sub> ΔCt (normalised and 95th percentile truncated) was related to the fixed effects (i) *bla*<sub>TEM-1</sub> cell copy number (normalised), (ii) presence of the C32T promoter mutation, (iii) presence of the G175A mutation, and (iv) their interaction. Random effects were incorporated to account for qPCR replicates and

phylogenetic relationships between isolates. See Supplementary Information for model specification, outputs and diagnostics.

In decreasing order of effect size, C32T, G175A, and a one unit increase in contig copy number all increased expression (decreased ΔCt; Table 2). There was no additional effect of G175A if C32T was also present (−1.69 < −1.71). After accounting for these covariates, we still identified a contribution from isolate phylogeny. The posterior



**Fig. 2 | Intrinsic expression of *bla*<sub>TEM-1</sub> shapes co-amoxiclav MIC across the *E. coli* phylogeny.** **a** A midpoint-rooted core gene phylogeny of *E. coli* chromosomes from the 67 isolates with expression quantified. Tips are coloured by phylogroup. ST12, ST127 and ST69 are highlighted with tip outlines and circles at their crown nodes. Tip shape distinguishes location of *bla*<sub>TEM-1</sub> on chromosomes (circles) and plasmids (squares). **b** Bar length records mean *bla*<sub>TEM-1</sub> ΔCt for each tip (log<sub>10</sub>-scaled; lower values denote higher *bla*<sub>TEM-1</sub> expression). Bar colour records isolate co-amoxiclav MIC for each tip. **c** Posterior means (coloured circles) and 95% HPD

intervals (horizontal lines) for phylogenetic effect on negative *bla*<sub>TEM-1</sub> ΔCt for each tip (multiplied by −1 for ease of comparison). Red indicates above average expression and blue indicates below average expression. **d** Posterior means (coloured circles) and 95% HPD intervals (horizontal lines) for phylogenetic effect on co-amoxiclav MIC for each tip. Pink indicates above average MIC and green indicates below average MIC. Note that isolates OXEC-238 and OXEC-490 were excluded from the MIC model (see exclusion criteria).

**Table 2 | Parameter estimates for *bla*<sub>TEM-1</sub> delta cycle threshold (ΔCt) genotype-phenotype model**

| Variable                                     | Beta coefficient posterior mean | 95% HPD (l, u) | p <sub>MCMC</sub> |
|--|---------------------------------|----------------|-------------------|
| Intercept                                    | 0.30                            | 0.03, 0.59     | 0.0245            |
| C32T*  | −1.71                           | −2.08, −1.34   | <1e-05            |
| G175A*                                       | −0.31                           | −0.57, −0.04   | 0.0202            |
| C32T*:G175A* interaction                     | 0.33                            | −0.14, 0.81    | 0.1779            |
| <i>bla</i> <sub>TEM-1</sub> cell copy number | −0.12                           | −0.24, 0.01    | 0.0679            |

All values were taken from chain 1. Lower (l) and upper (u) values are given for the 95% highest posterior density (HPD) intervals. p<sub>MCMC</sub> is the posterior probability that the coefficient was positive (defined in the Supplementary Information). The effect of both C32T\* and G175A\* is −1.71−0.31+0.33=−1.69.

\* Numbering by Sutcliffe<sup>26</sup>.

for contribution of variance from phylogeny demonstrated a long right tail (mean = 0.07; 95% highest posterior density, HPD = [0.00, 0.21]; see “Methods”), suggestive of heterogeneity in phylogenetic signal, where deeper splits between major lineages may explain disproportionately large differences in expression. For qPCR replicates, the contribution of variance exhibited minimal skew (mean = 0.15; 95% HPD = [0.08, 0.23]). To investigate this further, we computed the posterior mean and 95% HPD credible interval for each tip in the phylogeny (Fig. 2c). Compared to the average across the *E. coli* phylogeny, some phylogroups (B1, C) and STs (12) were associated with increased *bla*<sub>TEM-1</sub> expression, whilst some phylogroups (E, F) and STs (372) were associated with decreased *bla*<sub>TEM-1</sub> expression.

***ampC* gene variation is highly concordant with *E. coli* lineage**  
In many Gram-negative species, chromosomal *ampC* is regulated by the transcriptional activator AmpR and is inducible in the presence of beta-lactams. However, *E. coli* lacks *ampR* and therefore expresses *ampC* constitutively; overproduction depends on mutations in the promoter or attenuator regions. At the time of writing, the beta-lactamase database (BLDB) contains *n* = 4915 non-synonymous variants of the gene<sup>28</sup>.

To quantify how well *ampC* variants agree with phylogroup and ST, we calculated the homogeneity (*h*) and completeness (*c*); both range from 0 to 1; see “Methods”). Briefly, *h* = 1 means that a phylogroup or ST contains a single *ampC* variant. Conversely, *c* = 1 means that all instances of an *ampC* variant fall within the same phylogroup or ST. For phylogroups, we found *h* = 0.489 and *c* = 0.964, and for STs (excluding 38/377, which were unassigned), *h* = 0.938 and *c* = 0.877. Overall, this suggests that phylogroups tend to contain distinct *ampC* variants, which are generally ST-specific, and overall, that *E. coli* phylogeny is a suitable proxy for *ampC* variation.

Whilst many *E. coli* *ampC* variants present a narrow spectrum of hydrolytic activity, some can potentially hydrolyse third-generation cephalosporins following mutations in the promoter sequence. To explore promoter variation, we aligned all *n* = 377 *ampC* promoter sequences. Mutations outside positions −42 to +37 (according to Jaurin numbering<sup>44</sup>) were disregarded based on existing characterisations<sup>29,30</sup>. In total, *n* = 12 *ampC* promoter SNVs were identified, with variation dominated by the *E. coli* K12 wildtype at 47% (177/377). Table 3 documents all *n* = 11 mutations identified. A given *ampC* variant associated almost uniquely with an *ampC* promoter variant, yet *ampC* promoter variants were associated with multiple *ampC* variants (*h* = 0.483 and *c* = 0.941).



**Table 3 | Variation in *n* = 377 *ampC* promoters**

| Region                         | Position | <i>E. coli</i> K12 ( <i>n</i> ) | Mutation ( <i>n</i> ) |
|--------------------------------|----------|---------------------------------|-----------------------|
| Spacer                         | −28      | G (243)                         | A (134)               |
|                                | −18      | G (329)                         | A (48)                |
| Between −10 box and attenuator | −1       | C (329)                         | T (48)                |
|                                | +11      | T (371)                         | − (6)                 |
| Attenuator                     | +17      | C (316)                         | T (61)                |
|                                | +22      | C (367)                         | T (10)                |
|                                | +26      | T (367)                         | G (10)                |
|                                | +27      | A (367)                         | T (10)                |
|                                | +30      | G (376)                         | A (1)                 |
|                                | +32      | G (364)                         | A (13)                |
|                                | +37      | G (376)                         | T (1)                 |

Positions are according to Jaurin numbering<sup>14</sup>.

### *E. coli* phylogeny drives co-amoxiclav resistance through expression

We next investigated whether the *E. coli* lineages with intrinsically higher *bla*<sub>TEM-1</sub> expression also had intrinsically higher co-amoxiclav MICs. This would be consistent with lineage differences in regulatory regions and epigenetic interactions driving increased resistance.

We employed an MCMC to estimate parameters in an ordinal mixed model. The response variable isolate co-amoxiclav MIC (μg/mL; levels ≤2/2, 4/22, 8/22, 16/2, 32/2, >32/2) was predicted by the fixed effects (i) *bla*<sub>TEM-1</sub> cell copy number (normalised and 95th percentile truncated), (ii) *bla*<sub>TEM-1</sub> genome copy number (>1 vs. 1), (iii) non-wildtype *bla*<sub>TEM-1</sub> promoter SNVs, and (iv) non-wildtype *ampC* promoter SNVs. For the model, we only used isolates for which every *bla*<sub>TEM-1</sub> gene was linked to a promoter, and all the promoters were the same variant. We then filtered out isolates with *bla*<sub>TEM-1</sub> promoter and *ampC* promoter variants that appeared less than 10 times. This left *n* = 292/377 isolates. Full model specification, convergence diagnostics, and outputs are given in Supplementary Information.

In decreasing order of effect size, the presence of C32T and G175A in the *bla*<sub>TEM-1</sub> promoter, the presence of just C32T, *bla*<sub>TEM-1</sub> cell copy number and *bla*<sub>TEM-1</sub> genome copy number all increased co-amoxiclav MIC (Table 4); the remaining effects were compatible with chance. As with the expression model, the posterior distribution for contribution of variance from phylogeny demonstrated a long right tail (mean = 2.80; 95% HPD = [0.72, 5.16]). The phylogeny for the *n* = 292 isolates with co-amoxiclav MIC tip effects is given in Fig. S6.

We found that phylogenetic tip effects alone were sufficient to shift category membership across the EUCAST resistance breakpoint (> 8/2 μg/mL). Fixing all other covariates at the midpoint of their latent categories, 19% (55/292) of isolates had a greater than 50% posterior probability of being shifted from the susceptible to the resistant category due to their phylogenetic effect alone. Of these, most were represented by phylogroup B2 and D at 64% (35/55) and 20% (11/55), respectively. The most represented sequence types were ST12, ST127 and ST69 at 27% (15/55), 20% (11/55) and 18% (10/55), respectively. The isolates in these sequence types also used in the expression model are highlighted in Fig. 2d.

Lastly, we developed a combined model to test whether the phylogenetic influence on *bla*<sub>TEM-1</sub> expression causally varies co-amoxiclav MIC (see Supplementary Information). Here, we only used the predictors identified as significant from previous expression and MIC models (*bla*<sub>TEM-1</sub> cell and genome copy numbers, and *bla*<sub>TEM-1</sub> promoter SNV). Briefly, the model estimates a parameter that scales the phylogenetic and non-phylogenetic random effects from expression to MIC. Under causality, the scaling parameter should be constant across all random effect terms. We found the scaling parameter had a

**Table 4 | Parameter estimates for co-amoxiclav minimum inhibitory concentration (MIC) genotype-phenotype model**

| Variable   |  | Beta coefficient | 95% HPD                 | <i>p</i> <sub>MCMC</sub> |
|--|--|------------------|-------------------------|--------------------------|
|  |  | posterior mean   | ( <i>l</i> , <i>u</i> ) |                          |
| Intercept  |  | 3.92             | 2.92, 4.85              | <1e-05                   |
| <i>bla</i> <sub>TEM-1</sub> cell copy number               |  | 2.07             | 1.38, 2.78              | <1e-05                   |
| <i>bla</i> <sub>TEM-1</sub> genome copy number (> 1 vs. 1) |  | 0.97             | 0.02, 1.90              | 0.0414                   |
| <i>bla</i> <sub>TEM-1</sub> promoter SNV vs. wildtype      | G175A <sup>a</sup>                       | 0.17             | −0.35, 0.69             | 0.5313                   |
|  | C32T <sup>a</sup>                        | 6.06             | 4.15, 8.08              | <1e-05                   |
|  | C32T <sup>a</sup> and G175A <sup>a</sup> | 5.87             | 3.89, 7.87              | <1e-05                   |
| <i>ampC</i> promoter SNV vs. wildtype                      | G-28A <sup>b</sup> and C17T <sup>b</sup> | 0.44             | −0.63, 1.48             | 0.4122                   |
|  | G-28A <sup>b</sup>                       | 0.87             | −0.89, 2.58             | 0.3039                   |
|  | G-18A <sup>b</sup> and C-1T <sup>b</sup> | 0.83             | −1.37, 3.11             | 0.4402                   |

All values were taken from chain 1. For single-nucleotide variant (SNV) vs. wildtype, *E. coli* K12 was used as the wildtype baseline. Lower (*l*) and upper (*u*) values are given for the 95% highest posterior density (HPD) intervals. *p*<sub>MCMC</sub> is the posterior probability that the coefficient was positive (defined in the Supplementary Information).

<sup>a</sup> Numbering by Sutcliffe<sup>26</sup>.

<sup>b</sup> Numbering by Jaurin<sup>14</sup>.

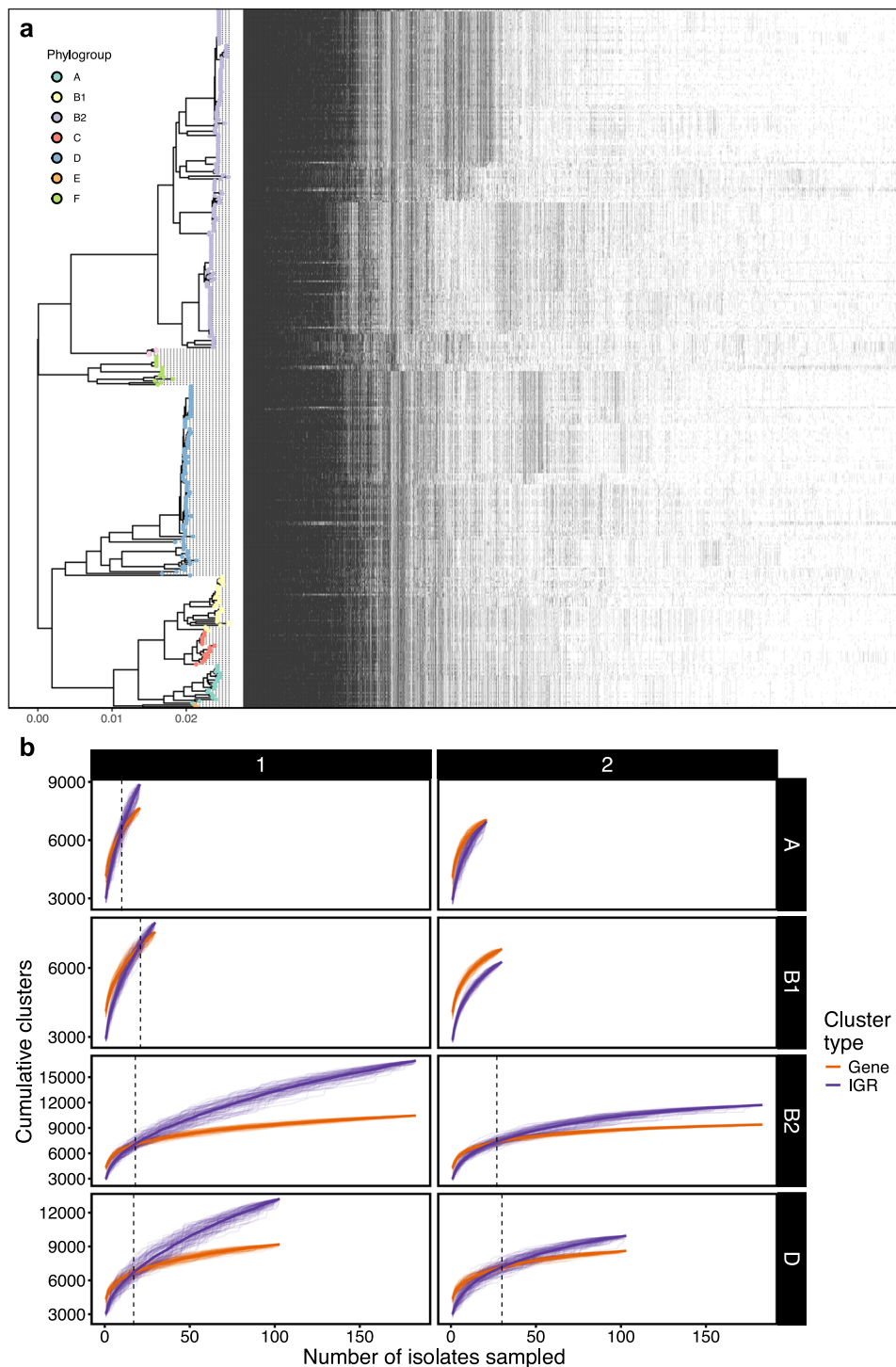
posterior mean = −1.13 (95% HPD = [−0.72, −1.49]; *p*<sub>MCMC</sub> = 0.002; values taken from chain 1). This supports a direct and substantial influence of expression on MIC, mediated by phylogenetic relationships.

### Intergenic architecture and dynamics vary between *E. coli* phylogroups

The results from our modelling suggest that lineage-specific regulatory systems play a role in the expression of *bla*<sub>TEM-1</sub>, which in turn modulates resistance. The intergenic regions (IGR) in bacterial chromosomes contain their regulatory systems, with one study finding that 86% of IGRs in a strain of *E. coli* were transcriptionally active<sup>31</sup>. Like genes, IGRs are subject to genetic flow within bacterial populations. Yet, how closely these dynamics mirror one another remains unclear. Understanding whether IGRs evolve more rapidly than coding sequences can shed light on how bacterial populations adapt and diversify not just at the level of the genes they carry, but in how they control the expression of those genes, ultimately affecting their phenotypic traits.

We first began with a traditional pangenome analysis for all *n* = 377 chromosomes (see “Methods”). Here, we identified *n* = 15,781 gene clusters, with 5.0% (782/15,781) present in all isolates, and singletons (one member) and doubletons (two members) comprising 22.7% (3576/15,781) and 9.0% (1413/15,781), respectively. We then performed a similar analysis, but for IGR clusters (see “Methods”). In total, we identified *n* = 33,345 IGR clusters, of which 1.0% (322/33,345) were present in all isolates, consistent with a small core IGR system. Most IGR clusters represented singletons at 44.7% (14,903/33,345) or doubletons at 11.9% (3977/33,345). We then visualised the top 10,000 IGR clusters against our *E. coli* phylogeny in Fig. 3a. This indicated phylogroup- and sequence type-specific patterns in intergenic sequence.

We next wanted to understand whether IGR diversity outpaced coding-sequence diversity. To do this, we built cluster accumulation curves, randomising genome-addition order 100 times per phylogroup, then plotting the cumulative number of clusters discovered versus the number of genomes sampled. We performed this for all phylogroups with at least 20 members (A, B1, B2 and D) in two separate runs: for all cluster sizes, and clusters with at least two members to



**Fig. 3 | Intergenic region content and dynamics vary across the *E. coli* phylogeny. a** The top 10,000 most common IGR clusters ordered by prevalence (left to right), arranged by midpoint-rooted core gene phylogeny. Tips are coloured by

phylogroup. **b** Gene and IGR cluster accumulation curves for phylogroups A, B1, B2 and D. Dashed vertical lines indicate crossover points of averaged accumulation curves.

account for possible assembly and bioinformatic errors. The averaged curves are presented in Fig. 3b. When considering clusters of all sizes, the IGR curves began below the gene curves, reflecting rapid discovery of moderately common accessory genes, but all overtook them after a period of sampling. This crossover marks the sampling depth at which IGR novelty becomes the dominant source of new information.

Fitting Heaps' law to the mean curves provided a concise measure of openness (see Fig. 3b and Table 5). Briefly,  $\alpha$  describes the rate at

which diversity accumulates as more samples are taken, where a higher  $\alpha$  indicates faster accumulation. We found that  $\alpha$  was consistently higher for IGR curves than for gene curves, and phylogroup A exhibited the highest  $\alpha$ , indicative of the most open IGR system. To ensure these patterns were not driven by unequal sampling, we regressed  $\alpha$  against both phylogroup sample size and cluster type (IGR vs gene). In this model (adjusted  $R^2 = 0.73$ ), sample size had no significant effect on  $\alpha$  ( $\beta = -1.2e-04$ ,  $p$ -value = 0.48), whereas IGRs showed a highly

**Table 5 | Heaps’ law model alpha ( $\alpha$ ) estimates for gene and intergenic region (IGR) cluster analysis**

| Phylogroup | Minimum cluster size | Cluster type | $\alpha$ |
|------------|----------------------|--------------|----------|
| A          | 1                    | IGR          | 0.39     |
|            |                      | Gene         | 0.21     |
|            | 2                    | IGR          | 0.30     |
|            |                      | Gene         | 0.18     |
| B1         | 1                    | IGR          | 0.32     |
|            |                      | Gene         | 0.20     |
|            | 2                    | IGR          | 0.23     |
|            |                      | Gene         | 0.16     |
| B2         | 1                    | IGR          | 0.36     |
|            |                      | Gene         | 0.17     |
|            | 2                    | IGR          | 0.26     |
|            |                      | Gene         | 0.14     |
| D          | 1                    | IGR          | 0.35     |
|            |                      | Gene         | 0.17     |
|            | 2                    | IGR          | 0.27     |
|            |                      | Gene         | 0.15     |

$\alpha$  represents the decay parameter from fitting Heaps’ law model. Contains estimates for *E. coli* isolate phylogroups A ( $n = 21$ ), B1 ( $n = 30$ ), B2 ( $n = 183$ ) and D ( $n = 103$ ).

significant positive shift ( $\beta = 0.137$ ,  $p$ -value =  $2.2e-05$ ). Thus, the faster accumulation of IGR diversity truly reflects biological differences in chromosomal IGR dynamics, not sampling artifacts.

Discussion

In our dataset of clinical *E. coli*, *bla*<sub>TEM-1</sub> was overwhelmingly carried by conjugative plasmids. This means it can spread between bacterial hosts and different genetic backgrounds. We demonstrated that different bacterial hosts intrinsically vary in their ability to express *bla*<sub>TEM-1</sub> when accounting for variation in promoters (C32T and G175A mutations) and contig copy number. Moreover, our findings suggest that some clinically successful lineages (e.g., ST12) are better at expressing *bla*<sub>TEM-1</sub> than less clinically successful lineages (e.g., phylogroup F). With a second model, we also found that different *E. coli* lineages vary intrinsically in co-amoxiclav MIC (accounting for *bla*<sub>TEM-1</sub> genome and cell copies, and *bla*<sub>TEM-1</sub> and *ampC* promoter variants). Again, we observed that some clinically successful lineages (e.g., ST12) had higher resistance than less clinically successful lineages (e.g., phylogroup F). We also quantified that the clinically successful sequence types ST12, ST69 and ST127 had the highest probability of flipping an isolate from the sensitive to resistant category. A third model demonstrated that these two traits were causally linked: *E. coli* phylogeny drives co-amoxiclav resistance through variable expression of *bla*<sub>TEM-1</sub>, and underscores the necessity of fully resolving bacterial genomes to incorporate accurate genetic, genomic, and phylogenetic information in resistance prediction models. Future work could include evaluations of single amino acid substitutions in TEM-1 (that hydrolyse third-generation cephalosporins and carbapenemases<sup>28</sup>), which are typically carried in more complex genetic backgrounds.

This study has limitations. Firstly, it is possible that, due to fragmented plasmid assemblies, some isolates identified as having multiple copies of *bla*<sub>TEM-1</sub> on multiple plasmids instead had multiple copies on the same plasmid. Nonetheless, our expression analysis only considered isolates with a single copy of *bla*<sub>TEM-1</sub> in the genome, mitigating this concern. Secondly, we only examined expression for a subsample of our isolates due to resource limitations. Thirdly, whilst there is not an agreed upon standard reference gene for quantifying beta-lactamase expression<sup>32–34</sup>, previous work has shown 16S to be

stable<sup>33</sup>. Crucially, our delta Ct values were consistent within isolates. Fourthly, we only observed two potentially relevant porin mutations (a premature stop codon in *OmpC* on OXEC-40’s chromosome and in *OmpF* on OXEC-423’s chromosome), limiting our ability to investigate their effects on phenotype. We also identified no relevant efflux pumps beyond AcrEF-TolC (found in every isolate), and it was out of the scope of the study to explore their functionality. Fifthly, we found that tip effects did not strictly align with phylogenetic structure. We used a single fixed phylogeny and a single global phylogenetic-variance parameter, which does not capture uncertainty in tree topology or clade-specific evolutionary rates. Consequently, tips within densely sampled clades can be over-shrunk toward their mean, while tips in sparse clades can be under-shrunk. Future work should incorporate phylogenetic uncertainty (e.g., sampling from a tree posterior) and explore multi-variance or partitioned phylogenetic models to assess the impact of tree balance on tip-effect estimation. Sixthly, modelling phylogenetic effect terms for all the plasmid replicons in the dataset would be computationally prohibitive and statistically unstable. Whilst future studies could focus on curating datasets with less plasmid diversity, this would not be reflective of true clinical bacterial populations. Penultimately, automated susceptibility testing methods, like the BD Phoenix™ used here, may not agree completely with reference methods; yet previous work has shown strong agreement with the EUCAST agar dilution method<sup>35</sup>. Lastly, plasmid copy number is not static. Moreover, in the presence of antibiotics, it has been demonstrated that resistance gene-carrying plasmids can increase their copy number to increase the chance of survival<sup>36</sup>. Our point estimates of plasmid copy number were derived from genome assemblies sequenced in the absence of antibiotics, which likely represent a lower bound. Nonetheless, we found strong signal to suggest the import of plasmid copy number on resistance, even if under our sensitivity testing, plasmid copy number potentially increased within isolates.

We posit that lineage-specific regulatory systems in *E. coli*, shaped by horizontal gene flow, may house the key modulators of *bla*<sub>TEM-1</sub> expression. Although comprehensive dissection of these elements (for example, through ChIP-seq) extends beyond the scope of this study, our findings suggest that future efforts directed at mapping trans-acting factors, small RNAs, differential DNA methylation and nucleoid-associated protein binding across phylogroups will be essential. Additionally, examining other resistance genes and their expression patterns in a similar phylogenetic framework could provide a broader understanding and prediction of resistance mechanisms across different bacterial species and antibiotics. Our study demonstrates the some clinically successful lineages are better at expressing *bla*<sub>TEM-1</sub>, and have a higher probability of flipping from sensitive to resistant. We speculate that, since *bla*<sub>TEM-1</sub> is both widespread in clinical bacterial populations and its spread is plasmid mediated, being able to better control the expression of *bla*<sub>TEM-1</sub> when acquired is a selective advantage for clinical isolates.

Methods

Ethics oversight

The use of genotypic and phenotypic data from these isolates is covered by ethical permissions (London–Queen Square Research Ethics Committee, REC ref. 17/LO/1420). Isolates were a subset of those evaluated in a previous study<sup>37</sup>, for which data linkage with patient data/antibiotic susceptibility test data was enabled through the Infections in Oxfordshire Research Database (IORD). This database has generic Research Ethics Committee, Health Research Authority and Confidentiality Advisory Group approvals (19/SC/0403, 19/CAG/0144) which facilitate the pseudo-anonymised linkage of routinely collected NHS electronic healthcare record data from the Oxford University Hospitals NHS Foundation Trust Clinical Systems Data Warehouse and research data (e.g., sequencing data) from the Modernising Microbiology and Big Infection Diagnostics Theme of the Oxford NIHR



Biomedical Research Centre, Oxford. IORD links records by a specific, random, number ensuring that no patient-identifiable information is shared with researchers using this resource.

### Isolate selection

We considered  $n = 548$  candidate *E. coli* bacteraemia isolates cultured from patients presenting to Oxford University Hospitals NHS Foundation Trust between 2013 and 2018, and selected from a larger study of systematically sequenced isolates based on screening their short-read only assemblies with NCBIAMRFinder (v. 3.11.2) for *bla*<sub>TEM-1</sub>, and the absence of other beta-lactamases<sup>38</sup>.

### DNA extraction and sequencing

Sub-cultures of isolate stocks, stored at  $-80^{\circ}\text{C}$  in 10% glycerol nutrient broth, were grown on Columbia blood agar (CBA) overnight at  $37^{\circ}\text{C}$ . DNA was extracted using the EasyMag system (bioMerieux) and quantified using the Broad Range DNA Qubit kit (Thermo Fisher Scientific, UK). DNA extracts were multiplexed as 24 samples per sequencing run using the Oxford Nanopore Technologies (ONT) Rapid Barcoding kit (SQK-RBK110.96) according to the manufacturer's protocol. Sequencing was performed on a GridION using version FLO-MINI06 R9.4.1 flow cells with MinkNOW software (v. 21.11.7) and basecalled using Guppy (v. 3.84). Short-read sequencing was performed on the Illumina HiSeq 4000, pooling 192 isolates per lane, generating 150 bp paired end-reads<sup>37</sup>.

### Dataset curation and genome assembly

Full details are given in Supplementary Information. Briefly, short- and long-read quality control used fastp (v. 0.23.4) and filtlong (v. 0.2.1), respectively<sup>39,40</sup>. We also used Rasusa (v. 0.7.1) on  $n = 3/548$  long-read sets due to memory constraints<sup>41</sup>. Genome assembly used Flye (v. 2.9.2-b1786) with bwa (v. 0.7.17-r1188) and Polypolish (v. 0.5.0), and Unicycler (v. 0.5.0), using SPAdes (v. 3.15.5), miniasm (v. 0.3-r179) and Racon (v. 1.5.0)<sup>42–48</sup>. Plasmid contig validation used Mash screen (v. 2.3) with PLSDB (v. 2023\_06\_23\_v2)<sup>49,50</sup>. All assemblies were annotated with NCBIAMRFinder (v. 3.11.26 and database v. 2023-11-15.1)<sup>38</sup>. Alongside, we validated the presence of *bla*<sub>TEM-1</sub> using tblastn (v. 2.15.0+) with the NCBI Reference Gene Catalog TEM-1 RefSeq protein WP\_000027057.1 and 100% amino acid identity<sup>51</sup>. Following genome assembly, we removed  $n = 171/548$  isolates, either because (i) the chromosome did not circularise (116/171), (ii) it carried a non-*bla*<sub>TEM-1</sub> *bla*<sub>TEM</sub> variant and/or an additional acquired beta-lactamase (54/171), or (iii) the chromosome was too short consistent with misassembly ( $\sim 3.5\text{Mbp}$ ; 1/171). This left a final dataset of  $n = 377$  isolates.

### Antibiotic susceptibility testing

Antibiotic susceptibility testing was performed using the BD Phoenix™ system in accordance with the manufacturers' instructions, generating MICs for co-amoxiclav.

### Generation of cDNA template

RNA extraction and DNase treatment were performed on replicates of each isolate ( $n = 3$  biological/ $n = 3$  technical) as described previously<sup>52</sup>. RNA was quantified post DNase treatment using Broad Range RNA Qubit kit (Thermo Fisher Scientific, UK), normalised to 1  $\mu\text{g}$  and reverse transcribed to cDNA using SuperScript IV VILO (Thermo Fisher Scientific, UK) under the following conditions:  $25^{\circ}\text{C}$  for 10 min,  $42^{\circ}\text{C}$  for 60 min and  $85^{\circ}\text{C}$  for 5 min.

### qPCR quantification of *bla*<sub>TEM-1</sub> expression

*bla*<sub>TEM-1</sub> expression was quantified in a selection of isolates: initially  $n = 35$  isolates in triplicate, referred to as batch 1; then a further  $n = 48$  isolates in duplicate, referred to as batch 2. Batch 1 were randomly selected by MIC ( $n = 2$  MIC  $\leq 2/2$ ,  $n = 5$  MIC  $4/2$ ,  $n = 9$  MIC  $8/2$ ,  $n = 10$

MIC  $16/2$ ,  $n = 4$  MIC  $32/2$ ,  $n = 5$  MIC  $> 32/2$ ). Batch 2 was enhanced for specific *bla*<sub>TEM-1</sub> promoter mutations, selecting all isolates with a single *bla*<sub>TEM-1</sub> gene with C32T (with or without a G146A mutation) that had not already been tested, and then randomly selecting from other wildtype and G146A, single *bla*<sub>TEM-1</sub> gene isolates. For all qPCR reactions, *E. coli* cDNA was normalised to 1 ng and amplified in a duplex qPCR reaction targeting *bla*<sub>TEM-1</sub> and 16S. qPCR standard curves were prepared for both *bla*<sub>TEM-1</sub> (Genbank Accession: DQ221255.1) and 16S (Genbank Accession: LC747145.1) sequences cloned into pMX vectors (Thermo Fisher Scientific, UK). Tenfold dilutions of linearised plasmids ( $1\text{--}1 \times 10^7$  copies/reaction) were used as a standard curve for each experiment. Both curves were linear in the range tested (16S:  $R^2 > 0.991$ ; TEM-1:  $R^2 > 0.91$ ). The slopes of the standard curves for 16S and *bla*<sub>TEM-1</sub> were  $-3.607$  and  $-3.522$ , respectively. qPCR was performed using a custom 20  $\mu\text{l}$  TaqMan gene expression assay consisting of TaqMan™ Multiplex Master Mix, TaqMan unlabelled primers and a TaqMan probe with dye label (FAM for TEM-1 and VIC for 16S) carried out on the QuantStudio5™ real-time PCR system (Thermo Fisher Scientific, UK). Cycling conditions were  $95^{\circ}\text{C}$  for 20 s, followed by 40 cycles of  $95^{\circ}\text{C}$  for 3 s and  $60^{\circ}\text{C}$  for 30 s, with Mustang purple as the passive reference. For batch 1, triplicate samples were analysed and standardized against 16S rRNA gene expression. Triplicate reactions for each isolate demonstrated good reproducibility for batch 1 (Fig. S7). Of note, for isolate OXEC-75, TEM-1 expression was very low-level, and 5 reactions (1 technical replicate for biological replicate 1, 1 technical replicate for biological replicate 2, and all 3 technical replicates for biological replicate 3) failed to amplify any product. Due to resource constraints, we reduced replicates for batch 2 ( $n = 1$  biological/ $n = 2$  technical; Fig. S8). To reduce model complexity, we omitted some batch 1 isolates ( $n = 16/35$ ) which carried more than one copy of *bla*<sub>TEM-1</sub> in the genome, leaving a total of  $n = 67$  isolates.  $\Delta\text{Ct}$  values were calculated by subtracting mean 16S Ct from mean TEM-1 Ct.

### Assembly annotations

We annotated the chromosomes using Prokka (v. 1.14.6) with default parameters except `--centre X --compliant` (see `annotate.sh`)<sup>53</sup>. Abricate (v. 1.0.1) was used with default parameters and the PlasmidFinder database (v. 2023-Nov-4) to annotate for plasmid replicons<sup>54,55</sup>. Plasmid mobilities were predicted using MOB-suite's MOBytyper (v. 3.1.4) with default parameters<sup>55</sup>. Briefly, a plasmid was labelled as putatively conjugative if it had both a relaxase and mating pair formation (MPF) complex, mobilisable if it had either a relaxase or an origin of transfer (*oriT*) but no MPF, and non-mobilisable if it had no relaxase and *oriT*. Lastly, we assigned sequence types (STs) and phylogroups to our *E. coli* chromosomes using mlst (v. 2.23.0) with default parameters and EzClermont (v. 0.7.0) with default parameters, respectively<sup>24,56</sup>. We used blastn (v. 2.15.0+) with a custom database of known *bla*<sub>TEM-1</sub> promoters<sup>13,25,57</sup>. Due to the high similarity between the *P3*, *Pa/Pb*, *P4* and *P5* reference sequences, we chose the top hit in each position.

### SNV analysis

We first determined the sets of sequences we wanted to align: (i) *bla*<sub>TEM-1</sub> ( $n = 451$ ; some genomes carried multiple copies), (ii) *bla*<sub>TEM-1</sub> promoters ( $n = 409$ ), (iii) *ampC* ( $n = 377$ ) and (iv) *ampC* promoters ( $n = 377$ ). For *bla*<sub>TEM-1</sub> and *ampC*, we extracted the relevant sequences using the coordinate and strand information from the NCBIAMR-Finder output (see `extractGene.py`). For the *bla*<sub>TEM-1</sub> promoters, we used coordinate and strand information from the earlier blastn results. For the *ampC* promoters, we took the sequence 200 bp upstream of the *ampC* gene then manually excised the  $-42$  to  $+37$  region in AliView<sup>58</sup>. Sets of sequences were aligned using MAFFT (v. 7.520) with default parameters except `--auto`<sup>59</sup>. Variable sites were examined using snp-sites (v. 2.5.1) with default parameters and in `-v` mode<sup>60</sup>.



## Contig copy number

We used BWA (v. 0.7.17-r1188) to map the quality-controlled short-reads to each contig, then SAMtools (v. 1.18) for subsequent processing (see `copyNumber.sh`)<sup>43,61</sup>. For each contig, we calculated the mean depth over its length, then within each assembly, normalised by the mean depth of the chromosome.

## Chromosomal core gene phylogeny

Building the chromosomal phylogeny involved four main steps: annotating the chromosomes, identifying the core genes, aligning them and building a phylogeny. Initially, all the chromosomes carried a copy of *ampC*, meaning it was a core gene and would be included in the phylogeny. Since we wanted to manually verify EzClermont phylogroup classifications with the phylogeny and then compare phylogroups to the distribution of *ampC* gene variants, we excised the *ampC* sequence from all the chromosomes beforehand to avoid confounding our analysis (see `removeGene.py`). To identify the core genes (those with  $\geq 98\%$  frequency in the sample), we used Panaroo (v. 1.4.2) with default parameters except `--clean-mode sensitive --aligner mafft -a core --core_threshold 0.98`<sup>62</sup>. Panaroo also aligned our core genes using MAFFT (v. 7.520; see `runPanaroo.sh`)<sup>59</sup>. Lastly, we built the core gene maximum-likelihood phylogeny using IQ-Tree (v. 2.3.0) with default parameters except `-m GTR+F+I+R4 -keep-ident -B 1000 -mem 10 G using -s core_gene_alignment_filtered.aln from Panaroo` (see `runIQTREE.sh`)<sup>63</sup>. The substitution model used was general time reversible (GTR) using empirical base frequencies from the alignment (F), allowing for invariant sites (I) and variable rates of substitution (R4).

## Intergenic region analysis

For the gene cluster analysis, we used Panaroo (v. 1.5.1) with default parameters except `--clean-mode sensitive --merge paralogs`<sup>62</sup>. Running Panaroo with paralog merging is necessary for Piggy, and reduced the overall pangenome size from the previous run. For the intergenic region cluster analysis, we used Piggy (v. 1.5) with default parameters<sup>64</sup>.

## Statistical analysis and visualisation

All statistical analysis was performed in R (v. 4.4.0) using RStudio (v. 2024.04.2+764)<sup>65,66</sup>. We implemented MCMC generalised linear mixed models using the MCMCglmm library in R<sup>67</sup>. Model specifications, convergence diagnostics, parameter estimations and outputs are reported in Supplementary Information; see `modelExpression.R`, `modelMIC.R` and `modelCombined.R`, to reproduce the *bla*<sub>TEM-1</sub> expression, co-amoxiclav MIC and causal models, respectively. Homogeneity and completeness are defined in Rosenberg, A. and Hirschberg, J (2007) and were also implemented in R<sup>68</sup>. A 95% highest posterior density (HPD) credible interval finds the closest points (*a* and *b*) for which  $F(b) - F(a) = 0.95$ , where *F* is the empirical density of the posterior. Figures were plotted with the ggplot2 library<sup>69</sup>. See `buildResults.R` and `igr_analysis.R` to reproduce all statistics and figures in the manuscript.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Metadata for all *n* = 377 genomes included in the final analysis is given in Supplementary Data File 1. Metadata for all *n* = 451 *bla*<sub>TEM-1</sub> annotations identified in these genomes is given in Supplementary Data File 2. qPCR expression data for all replicates is given in Supplementary Data File 3. NCBI accessions for short- and long-read sets and assemblies are given in Supplementary Data File 4. All Supplementary Data Files are also stably archived at <https://doi.org/10.5281/zenodo.16731807>.

## Code availability

All scripts referenced in the Methods can be found in the GitHub repository <https://github.com/wtmatlock/tem>, which is stably archived at <https://doi.org/10.5281/zenodo.16731807>.

## References

- Datta, N. & Kontomichalou, P. Penicillinase synthesis controlled by infectious R factors in Enterobacteriaceae. *Nature* **208**, 239–241 (1965).
- Partridge, S. R. & Hall, R. M. Evolution of transposons containing *bla*<sub>TEM</sub> genes. *Antimicrob. Agents Chemother.* **49**, 1267–1268 (2005).
- Shaw, L. P. & Neher, R. A. Visualizing and quantifying structural diversity around mobile resistance genes. *Microb. Genom.* **9**, 001168 (2023).
- Bailey, J. K., Pinyon, J. L., Anantham, S. & Hall, R. M. Distribution of the *bla*<sub>TEM</sub> gene and *bla*<sub>TEM</sub>-containing transposons in commensal *Escherichia coli*. *J. Antimicrob. Chemother.* **66**, 745–751 (2011).
- Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
- Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **52**, D33 (2023).
- Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
- Sanderson, T. A web tool for exploring the usage of medicines in hospitals in England. *Wellcome Open Res.* **9**, 147 (2024).
- EUCAST: clinical breakpoints and dosing of antibiotics. [https://www.eucast.org/clinical\\_breakpoints](https://www.eucast.org/clinical_breakpoints).
- Yoon, C. H. et al. Mortality risks associated with empirical antibiotic activity in *Escherichia coli* bacteraemia: an analysis of electronic health records. *J. Antimicrob. Chemother.* **77**, 2536–2545 (2022).
- Martinez, J. L. et al. Resistance to beta-lactam/clavulanate. *Lancet* **330**, 1473 (1987).
- Reguera, J. A., Baquero, F., Perez-Diaz, J. C. & Martinez, J. L. Synergistic effect of dosage and bacterial inoculum in TEM-1 mediated antibiotic resistance. *Eur. J. Clin. Microbiol. Infect. Dis.* **7**, 778–779 (1988).
- Lartigue, M. F., Leflon-Guibout, V., Poirer, L., Nordmann, P. & Nicolas-Chanoine, M. H. Promoters *P*<sub>3</sub>, *P*<sub>4</sub>, *P*<sub>5</sub> upstream from *bla*<sub>TEM</sub> genes and their relationship to  $\beta$ -lactam resistance. *Antimicrob. Agents Chemother.* **46**, 4035–4037 (2002).
- Jaurin, B., Grundström, T. & Normark, S. Sequence elements determining *ampC* promoter strength in *E. coli*. *EMBO J.* **1**, 875–881 (1982).
- Siasat, P. A. & Blair, J. M. A. Microbial primer: multidrug efflux pumps. *Microbiology* **169**, 001370 (2023).
- Vital, M. et al. Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification. *ISME J.* **9**, 1130–1140 (2015).
- McNally, A. et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* **12**, e1006280 (2016).
- Card, K. J., Thomas, M. D., Graves, J. L., Barrick, J. E. & Lenski, R. E. Genomic evolution of antibiotic resistance is contingent on genetic background following a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **118**, e2016886118 (2021).
- Wong, A. Epistasis and the evolution of antimicrobial resistance. *Front. Microbiol.* **8**, 246 (2017).
- Wang, X., Yu, D. & Chen, L. Antimicrobial resistance and mechanisms of epigenetic regulation. *Front. Cell. Infect. Microbiol.* **13**, 1199646 (2023).
- Dunn, S., Carrilero, L., Brockhurst, M. & McNally, A. Limited and strain-specific transcriptional and growth responses to acquisition

- of a multidrug resistance plasmid in genetically diverse *Escherichia coli* lineages. *mSystems* **6**, 10–1128 (2021).
22. Alonso-del Valle, A. et al. Antimicrobial resistance level and conjugation permissiveness shape plasmid distribution in clinical enterobacteria. *Proc. Natl. Acad. Sci. USA* **120**, e2314135120 (2023).
  23. Goussard, S. & Courvalin, P. Updated sequence information for TEM  $\beta$ -lactamase genes. *Antimicrob. Agents Chemother.* **43**, 367–370 (1999).
  24. Waters, N. R., Abram, F., Brennan, F., Holmes, A. & Pritchard, L. Easy phylotyping of *Escherichia coli* via the EzClermont web app and command-line tool. *Access Microbiol.* **2**, e000143 (2020).
  25. Robin, F. et al. Evolution of TEM-type enzymes: biochemical and genetic characterization of two new complex mutant TEM enzymes, TEM-151 and TEM-152, from a single patient. *Antimicrob. Agents Chemother.* **51**, 1304–1309 (2007).
  26. Sutcliffe, J. G. Nucleotide sequence of the ampicillin resistance gene of *Escherichia coli* plasmid pBR322. *Proc. Natl. Acad. Sci.* **75**, 3737–3741 (1978).
  27. Shaw, L. P. et al. Niche and local geography shape the pangenome of wastewater-and livestock-associated Enterobacteriaceae. *Sci. Adv.* **7**, eabe3868 (2021).
  28. Naas, T. et al. Beta-lactamase database (BLDB)—structure and function. *J. Enzym. Inhib. Med. Chem.* **32**, 917–919 (2017).
  29. Tracz, D. M. et al. *ampC* gene expression in promoter mutants of cefoxitin-resistant *Escherichia coli* clinical isolates. *FEMS Microbiol. Lett.* **270**, 265–271 (2007).
  30. Caroff, N., Espaze, E., Bérard, I., Richet, H. & Reynaud, A. Mutations in the *ampC* promoter of *Escherichia coli* isolates resistant to oxyminocephalosporins without extended spectrum  $\beta$ -lactamase production. *FEMS Microbiol. Lett.* **173**, 459–465 (1999).
  31. Raghavan, R., Groisman, E. A. & Ochman, H. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res.* **21**, 1487–1497 (2011).
  32. Rodríguez-Villodres, Á et al. Survival of infection with TEM  $\beta$ -lactamase-producing *Escherichia coli* with Pan- $\beta$ -lactam resistance. *J. Infect.* **89**, 106268 (2024).
  33. Singh, T. et al. Transcriptome analysis of beta-lactamase genes in diarrheagenic *Escherichia coli*. *Sci. Rep.* **9**, 3626 (2019).
  34. Kjeldsen, T. S. B. et al. CTX-M-1  $\beta$ -lactamase expression in *Escherichia coli* is dependent on cefotaxime concentration, growth phase and gene location. *J. Antimicrob. Chemother.* **70**, 62–70 (2015).
  35. Davies, T. J. et al. Reconciling the potentially irreconcilable? Genotypic and phenotypic amoxicillin-clavulanate resistance in *Escherichia coli*. *Antimicrob. Agents Chemother.* **64**, 10–1128 (2020).
  36. Hernandez-Beltra, J. C. R. et al. Plasmid-mediated phenotypic noise leads to transient antibiotic resistance in bacteria. *Nat. Commun.* **15**, 2610 (2024).
  37. Lipworth, S. et al. Ten-year longitudinal molecular epidemiology study of *Escherichia coli* and *Klebsiella* species bloodstream infections in Oxfordshire, UK. *Genome Med.* **13**, 144 (2021).
  38. Feldgarden, M. et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12728 (2021).
  39. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
  40. Wick, R. R. Filtlong. <https://github.com/rrwick/Filtlong> (2019).
  41. Hall, M. Rasusa: Randomly subsample sequencing reads to a specified coverage. *J. Open Source Softw.* **7**, 3941 (2022).
  42. Oxford Nanopore Technologies. medaka. <https://github.com/nanoporetech/medaka> (2018).
  43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
  44. Wick, R. R. & Holt, K. E. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput. Biol.* **18**, e1009802 (2022).
  45. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
  46. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
  47. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
  48. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
  49. Ondov, B. D. et al. Mash screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.* **20**, 232 (2019).
  50. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDb: a resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).
  51. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
  52. Rodger, G. et al. Comparison of direct cDNA and PCR-cDNA Nanopore sequencing of *Escherichia coli* isolates. *Microb. Genom.* **10**, 001296 (2024).
  53. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
  54. Carattoli, A. et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
  55. Robertson, J. & Nash, J. H. E. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* **4**, e000206 (2018).
  56. Seemann, T. mlst. <https://github.com/tseemann/mlst> (2017).
  57. Stoesser, N. et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *J. Antimicrob. Chemother.* **68**, 2234–2244 (2013).
  58. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
  59. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
  60. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
  61. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-science* **10**, giab008 (2021).
  62. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
  63. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
  64. Thorpe, H. A., Bayliss, S. C., Sheppard, S. K. & Feil, E. J. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience* **7**, giy015 (2018).
  65. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. <https://www.R-project.org/> (2025).
  66. Posit Team. RStudio: Integrated Development Environment for R. *Posit Software, PBC*. <http://www.posit.co/> (2024).
  67. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
  68. Rosenberg, A. & Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proc. of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. (2007).

69. Gómez-Rubio, V. ggplot2-elegant graphics for data analysis. *J. Stat. Softw.* **77**, 1–3 (2017).

## Acknowledgements

This work was funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915), a partnership between the UK Health Security Agency (UKHSA) and the University of Oxford. It was also supported by the NIHR Oxford Biomedical Research Centre (BRC). The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NIHR, UKHSA or the Department of Health and Social Care. The authors thank Jarrod Hadfield for providing guidance on implementing the MCMCglmm library.

## Author contributions

W.M.: conceptualisation, methodology, software, validation, formal analysis, data curation, writing–original draft, writing–review & editing, visualisation. G.R.: conceptualisation, methodology, validation, investigation, data curation, writing–review & editing. E.P.: methodology, writing–review & editing. M.C.: investigation, writing–review & editing. N.K.: investigation, writing–review & editing. L.B.: resources, writing–review & editing. M.M.: resources, writing–review & editing. S.O.: resources, writing–review & editing. K.L.H.: investigation, writing–review & editing. A.R.: writing–review & editing, project administration. D.K.: writing–review & editing. M.B.A.: writing–review & editing, supervision. A.S.W.: writing–review & editing, supervision, funding acquisition. S.L.: conceptualisation, writing–review & editing, supervision. N.S.: conceptualisation, writing–review & editing, supervision, project administration, funding acquisition.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-63714-6>.

**Correspondence** and requests for materials should be addressed to William Matlock or Nicole Stoesser.

**Peer review information** *Nature Communications* thanks Ruichao Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025