

# Discordance between different bioinformatic methods for identifying resistance genes from short-read genomic data, with a focus on *Escherichia coli*

Timothy J. Davies<sup>1,2,\*</sup>, Jeremy Swann<sup>1,2</sup>, Anna E. Sheppard<sup>1,2</sup>, Hayleigh Pickford<sup>1,2</sup>, Samuel Lipworth<sup>1,2</sup>, Manal AbuOun<sup>3</sup>, Matthew J. Ellington<sup>2,4</sup>, Philip W. Fowler<sup>1</sup>, Susan Hopkins<sup>2,4</sup>, Katie L. Hopkins<sup>2,5</sup>, Derrick W. Crook<sup>1,2,6</sup>, Timothy E. A. Peto<sup>1,2,6</sup>, Muna F. Anjum<sup>3</sup>, A. Sarah Walker<sup>1,2,†</sup> and Nicole Stoesser<sup>1,2,6,\*†</sup>

## Abstract

Several bioinformatics genotyping algorithms are now commonly used to characterize antimicrobial resistance (AMR) gene profiles in whole-genome sequencing (WGS) data, with a view to understanding AMR epidemiology and developing resistance prediction workflows using WGS in clinical settings. Accurately evaluating AMR in Enterobacterales, particularly *Escherichia coli*, is of major importance, because this is a common pathogen. However, robust comparisons of different genotyping approaches on relevant simulated and large real-life WGS datasets are lacking. Here, we used both simulated datasets and a large set of real *E. coli* WGS data ( $n=1818$  isolates) to systematically investigate genotyping methods in greater detail. Simulated constructs and real sequences were processed using four different bioinformatic programs (ABRicate, ARIBA, KmerResistance and SRST2, run with the ResFinder database) and their outputs compared. For simulation tests where 3079 AMR gene variants were inserted into random sequence constructs, KmerResistance was correct for 3076 (99.9%) simulations, ABRicate for 3054 (99.2%), ARIBA for 2783 (90.4%) and SRST2 for 2108 (68.5%). For simulation tests where two closely related gene variants were inserted into random sequence constructs, KmerResistance identified the correct alleles in 35 338/46 318 (76.3%) simulations, ABRicate identified them in 11 842/46 318 (25.6%) simulations, ARIBA identified them in 1 679/46 318 (3.6%) simulations and SRST2 identified them in 2 000/46 318 (4.3%) simulations. In real data, across all methods, 1 392/1 818 (76%) isolates had discrepant allele calls for at least 1 gene. In addition to highlighting areas for improvement in challenging scenarios, (e.g. identification of AMR genes at  $<10\times$  coverage, identifying multiple closely related AMR genes present in the same sample), our evaluations identified some more systematic errors that could be readily soluble, such as repeated misclassification (i.e. naming) of genes as shorter variants of the same gene present within the reference resistance gene database. Such naming errors accounted for at least 2 530/4 321 (59%) of the discrepancies seen in real data. Moreover, many of the remaining discrepancies were likely 'artefactual', with reporting of cut-off differences accounting for at least 1 430/4 321 (33%) discrepant. Whilst we found that comparing outputs generated by running multiple algorithms on the same dataset could identify and resolve these algorithmic artefacts, the results of our evaluations emphasize the need for developing new and more robust genotyping algorithms to further improve accuracy and performance.

Received 27 June 2023; Accepted 21 November 2023; Published 15 December 2023

**Author affiliations:** <sup>1</sup>Nuffield Department of Medicine, Oxford University, Oxford, UK; <sup>2</sup>National Institute for Health Research (NIHR) Health Protection Research Unit on Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford, Oxford, UK; <sup>3</sup>Bacteriology, Animal and Plant Health Agency, Surrey, UK; <sup>4</sup>Antimicrobial Resistance and Healthcare Associated Infections (AMRHA) Division, UK Health Security Agency, London, UK; <sup>5</sup>HCAI, Fungal, AMR, AMU and Sepsis Division, UK Health Security Agency, London, UK; <sup>6</sup>Oxford University Hospitals NHS Foundation Trust, Oxford, UK.

**\*Correspondence:** Nicole Stoesser, nicole.stoesser@ndm.ox.ac.uk; Timothy J. Davies, davies.t.j@live.co.uk

**Keywords:** antimicrobial resistance genotyping; *Escherichia coli*; genomics; resistance prediction .

**Abbreviations:** AMR, antimicrobial resistance; APHA, animal and plant health agency; AST, antimicrobial sensitivity test; DNA, deoxyribonucleic acid; MLST, multilocus sequence type; UKHSA, UK health security agency; WGS, whole genome sequencing.

†These authors contributed equally to this work

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary file, three supplementary tables and nine supplementary figures are available with the online version of this article.

001151 © 2023 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

### Impact Statement

Whole-genome sequencing is widely used for studying the epidemiology of antimicrobial resistance (AMR) genes in bacteria; however, there is some concern that outputs are highly dependent on the bioinformatics methods used. This work evaluates these concerns in detail by comparing four different commonly used AMR gene typing methods using large simulated and real datasets. The results highlight performance issues for most methods in at least one of several simulated and real-life scenarios. However most discrepancies between methods were due to differential labelling of the same sequences related to the assumptions made regarding the underlying structure of the reference resistance gene database (i.e. that resistance genes can be easily classified in well-defined groups). This study advances our understanding of discrepancies between the outputs of different AMR typing algorithms, with relevance for historic and future work using these algorithms. Some of the discrepancies can be resolved by choosing methods with fewer assumptions about the reference AMR gene database and manually resolving outputs generated using multiple programs. However, ideally new and better methods are needed.

## DATA SUMMARY

Sequencing data are available at the following NCBI BioProject accession numbers: PRJNA779173, PRJNA540750, PRJNA604975, PRJEB26317. For further information see <https://doi.org/10.1093/bioinformatics/btr708>.

## INTRODUCTION

Whole-genome sequencing (WGS) has become a major tool for characterizing the epidemiology of bacterial antimicrobial resistance (AMR) genes, representing a potentially highly discriminatory, non-targeted approach with significant advantages over other more targeted molecular techniques [1]. In addition, WGS-based antibiotic susceptibility prediction has been successfully implemented as part of diagnostic and treatment workflows for *Mycobacterium tuberculosis* [2] and more recently *Salmonella* species [3]. Accurate WGS-based profiling of complete AMR gene content and prediction of susceptibility phenotypes would represent an attractive option for other commonly encountered clinical bacterial pathogens, such as Enterobacterales, including *Escherichia coli*. However, many of these pathogens, together with the antimicrobials commonly used to treat them, have proved more challenging, with methods designed for this yet to meet the standards required to be used in clinical practice [4–6].

Several key components are required for WGS-based AMR genotyping and predictions of susceptibility phenotype, including a robust AMR gene reference catalogue linking each genetic mechanism/sequence with a given phenotype, and accurate AMR gene identification and classification algorithms. Several catalogues and bioinformatics algorithms are now available [7–13], but only limited comparative evaluation of their outputs has been undertaken. The genetic mechanisms underpinning AMR in Enterobacterales and some other bacteria (e.g. *Pseudomonas aeruginosa*) are much more complex than those in *M. tuberculosis*, and whilst some studies suggest that WGS-based genotyping holds promise for AMR gene characterization and the prediction of antimicrobial susceptibility for several different Enterobacterales species [14–16], the limited reproducibility and reliability of such methods in a blinded, head-to-head analysis across nine bioinformatics teams has recently been highlighted [17]. However, this study was small ( $n=10$  sequencing datasets,  $n=7$  isolates), encountered a limited set of typing discrepancies and used highly selective samples, meaning the impact of these issues on larger, real-world datasets remains unclear.

We therefore used simulations and three large, independent and diverse *E. coli* sequencing datasets to investigate the robustness and reproducibility of four widely used WGS-based AMR genotyping methods (ABRicate, ARIBA, KmerResistance and SRST2) at scale, investigating any encountered discrepancies.

## METHODS

### AMR gene identification methods

We evaluated the impact of different bioinformatics tools using the same AMR gene catalogue, namely the ResFinder database (v.29/10/2019). To be included, bioinformatics tools had to: (i) have publicly available code, (ii) run on local computing architecture without major modification, (iii) accept different AMR gene databases to ensure broad and long-term typing usability and (iv) have a command line interface that could enable batch processing of large numbers of samples (Table S1, available in the online version of this article).

We identified four publicly available bioinformatic tools that met these criteria and used distinct AMR gene identification approaches: ABRicate [18] v0.8.11 (which searches for AMR genes in assemblies using BLASTN [19] v2.2.31+), SRST2 [11] v0.2.0 (which maps reads directly onto the formatted AMR gene database using Bowtie 2 [20] v2.2.9), ARIBA [10] v2.11.1 (which combines these two approaches, first mapping reads to the AMR gene database using minimap, and then creating local assemblies of the mapped reads using Fermi-lite) and KmerResistance [12] v2.0 (which analyses shared k-mers between the

query sequences and reference sequences in the AMR gene database) (Fig. S1). To mimic broad usability, each program was run using default parameters. For ABRicate, assemblies were first produced using SPAdes [21] v3.12.0 run with default parameters.

### Simulated data: single and multiple allele identification, and low coverage scenarios

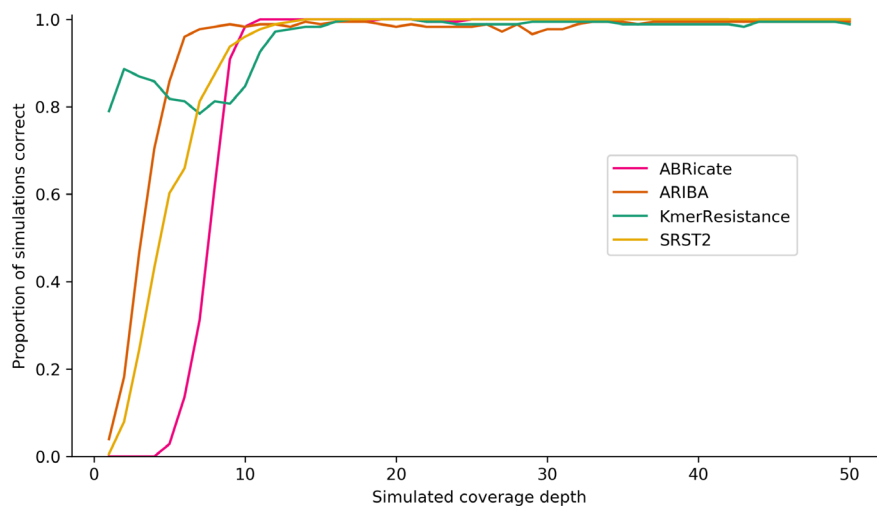
Prior to evaluating real data, we considered the accuracy of each method in identifying known AMR gene alleles ‘inserted’ into simulated flanking sequence constructs. For this, each AMR gene variant in the ResFinder database ( $n=3079$ ) was flanked by 1 kb of random sequence (using Numpy v1.16.4 [22] and combined using BioPython [23] v1.74) and reads simulated at  $40\times$  coverage using ART (details and rationale in Supplementary Methods, Figs 1 and S2) [24]. Other ART parameters were: error profile=‘HISEQ2500’, mean DNA fragment length (standard deviation)=480 bp (150 bp) and read length=151 bp. As we were interested to see how each of the tools performed when given correct and relatively simple sequencing data, we assessed the ‘recovery’ of each construct (i.e. how well we could recreate constructs from assembling from simulated reads) using QUAST [25] v4.3. This confirmed that all assemblies were a single contig covering  $>90\%$  of the original construct, and there were no errors (single-nucleotide polymorphisms, insertions or deletions) within the coding regions of the resistance genes. Each bioinformatic method was then tested to see if it could correctly identify the AMR gene variant, using default parameters. We repeated this 10 times to assess variability across repeats.

We also considered two *a priori* scenarios that are thought to affect AMR genotyping [26], namely a *multiple-allele* scenario in which multiple closely genetically related alleles (see below) of a given AMR gene were present, and a *low-quality* scenario reflected by low sequencing coverage. For the multiple-allele scenario we excluded target AMR gene variants that were incorrectly identified individually by any method (see Results), and then calculated pairwise nucleotide similarity between all remaining AMR gene variants. To do this, each remaining AMR gene variant was split into 31-mers, which were then compared with 31-mer sets from every other non-excluded AMR gene variant using pairwise Jaccard’s similarity indices. AMR gene variant pairs were defined as similar if they shared any 31-mer, resulting in a total of 46318 possible similar AMR gene variant pairs (Figs S3–S5).

For the low-coverage scenario, reads were simulated from 176  $bla_{TEM}$  gene-containing constructs at coverage depths ranging from 1–50 $\times$  using ART ( $n=176\times 50=8800$  simulations), reflecting total  $bla_{TEM}$  diversity present in the ResFinder database at the time of simulation (aside from those incorrectly identified in the single-allele scenario). Each construct contained a random perfect reference  $bla_{TEM}$  variant flanked by 1 kb of random sequence on each side produced using Numpy/BioPython as above. Simulated reads were then processed by each genotyping method using default settings and the identified variants were compared with the known  $bla_{TEM}$  variants present in each construct. The measure of performance for this scenario was the proportion of  $bla_{TEM}$  variants correctly identified by each method at each coverage level.

### Real data: isolate selection

To evaluate performance on real data, we then studied a total of 1818 *E. coli* isolates comprising 3 different WGS datasets in order to reflect different strain-level and AMR gene diversity: (i) 984 sequentially collected bloodstream infection isolates at Oxford University Hospitals (OUH) NHS Foundation Trust [27] (‘Oxford dataset’); (ii) 497 animal commensal *E. coli* isolates donated by the UK Animal and Plant Health Agency (APHA) [28] (‘APHA dataset’) and (iii) 337 *E. coli* isolates collected by UK Health Security Agency’s (UKHSA’s) Antimicrobial Resistance and Healthcare Associated Infections (AMRHAI) Reference Unit, which



**Fig. 1.** Proportion of correct genotype calls for single AMR gene variants in simulated constructs by coverage depth and bioinformatics method.

investigates isolates enriched for rare or important resistance genotypes encountered in the UK (sequenced for this study, ‘UKHSA dataset’). The ‘APHA dataset’ included five technical repeats (see Supplementary Material)

Isolates were recultured from frozen stocks stored in nutrient broth plus 10% glycerol at  $-80^{\circ}\text{C}$ . DNA was extracted using the QuickGene DNA Tissue kit S (Kurabo Industries, Japan) as per the manufacturer’s instructions, with an additional mechanical lysis step (FastPrep, MP Biomedicals, USA) immediately following chemical lysis. A combination of standard Illumina and in-house protocols was used to produce multiplexed paired-end libraries, which were sequenced on an Illumina HiSeq 2500, generating 151 bp paired-end reads. High-quality sequences were *de novo*-assembled using Velvet [29] v1.0.18 as previously described [30]. *In silico* Achtman [31] multilocus sequence types (MLSTs) were defined using ARIBA [10] v2.11.1.

While this work does not attempt to predict resistance from WGS data, each isolate had linked antimicrobial susceptibility test (AST) data (summarized in Table S2, Fig. S6), which we have included as the complexity of resistance genotype identification is associated with the phenotype. Isolates had complete AST data available for ampicillin, ceftazidime and one other third-generation cephalosporin (cefotaxime for the animal commensal isolates, ceftriaxone for all others), gentamicin, ciprofloxacin and co-trimoxazole.

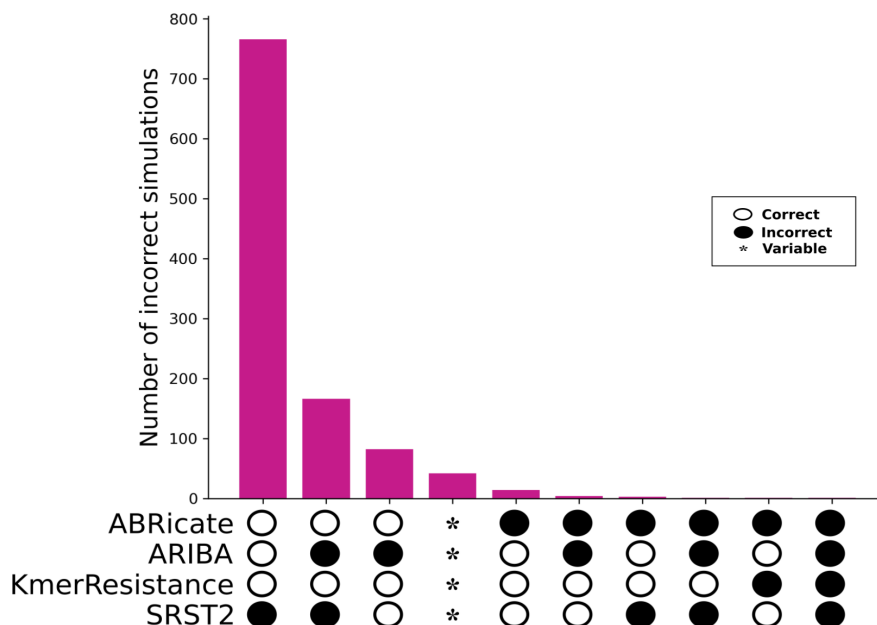
We compared the AMR genotypes reported for each isolate by each method, stratified by the antibiotic class to which resistance was conferred, as specified in the ResFinder database, namely: beta-lactams, aminoglycosides, quinolones, trimethoprim and sulphonamides. Discrepancies were classified according to which of the four bioinformatics methods agreed (Fig. S6). The cause of discrepancy was investigated for all beta-lactam resistance genotypes, because these antibiotics are most commonly used for clinical *E. coli* infections, and then for discrepancy patterns occurring in  $>1.5\%$  ( $n=27$ ) of isolates for the other classes.

## RESULTS

### Simulated scenarios

#### Accurate identification of single AMR gene variants in simulated sequence constructs

For the 3079 AMR gene variants in the ResFinder database, all four genotyping methods correctly identified those inserted into random sequence contexts for all repeats in 1999 (64.9%) cases. ARIBA was the only tool to intermittently correctly identify alleles across repeats [ $n=42/3079$  (1.3%) alleles], being always correct for 2783/3079 (90.4%) alleles. All other tools were consistent across repeats, with KmerResistance being correct for 3076 (99.9%) simulations, ABRicate for 3054 (99.2%) and SRST2 for 2108 (68.5%) (Fig. 2). For SRST2, most errors were due to its approach of pre-clustering reference sequences into sub-families by sequence identity prior to genotyping, thereby essentially excluding *a priori* the possibility of identifying alleles that were not selected as the representative for these sub-family clusters. This error is explained in more detail below, as it also affected genotyping in real isolate sequences.



**Fig. 2.** Identification of known single AMR gene variants in simulated contexts by bioinformatic method. Note only cases where one or more methods were incorrect are shown ( $n=1081$ ). \*, genes were variably correctly identified across 10 repeats.

**Table 1.** Performance of genotyping methods in evaluating simulated constructs with two related allelic variants. Percentage reported out of a total of 46 318 simulations performed for each method

Genotyping call	No. of calls (%)			
	ABRicate	ARIBA	KmerResistance	SRST2
No correct calls	6355 (14%)	29571 (64%)	1795 (4%)	10298 (22%)
One correct call but additional incorrect calls	21012 (45%)	164 (<1%)	1028 (2%)	184 (<1%)
One correct call, no incorrect calls	415 (1%)	14500 (31%)	8150 (18%)	33319 (72%)
Two correct calls, but additional incorrect calls	6712 (14%)	404 (1%)	7 (<1%)	517 (1%)
Two correct calls, no incorrect calls	11842 (26%)	1679 (4%)	35338 (76%)	2000 (4%)

### Impact of the presence of multiple closely related alleles on genotyping calls

The multiple-allele simulation caused problems for all algorithms. Overall, KmerResistance was the most successful, identifying both correct genes in 35 338/46 318 (76%) scenarios (Table 1). By contrast the next most successful was ABRicate, achieving this in 11 842/46 318 (26%) with the remaining tools, with ARIBA and SRST2 only achieving this in 1679 (4%) and 2000 (4%) scenarios, respectively. In addition, assembly-based algorithms frequently failed to completely assemble genes, with ABRicate and ARIBA reporting fragmented/incomplete alleles for 26 848/58 499 (46%) and 2483/18 830 (13%) of the correct alleles they identified, respectively. SRST2, as expected, found only a single allele in most [33319/46 318 (72%)] cases, as dictated by its clustering parameters. Unsurprisingly all four programs were most likely to make erroneous genotyping calls as the simulated pairs of alleles became more closely related (Fig. S7a, b).

### Impact of sequencing depth on genotyping calls

KmerResistance was able to identify *bla*<sub>TEM</sub> alleles at lower coverage than any of the other methods (Fig. 1). Above 15× depth of coverage for the gene, all methods correctly identified *bla*<sub>TEM</sub> alleles in simulated constructs in >95% of cases (Fig. 1). All methods were able to identify all of the *bla*<sub>TEM</sub> alleles correctly at least once, but examples existed for all methods where the allele was correctly identified at low coverage, but then mis-classified at higher coverage. In general, ABRicate and SRST2, while requiring greater sequencing depth to correctly identify *bla*<sub>TEM</sub> alleles initially, were more accurate at higher coverage depths, making erroneous calls for only 1/176 (0.6%) and 0/176 (0%) of *bla*<sub>TEM</sub> alleles at depths >20×. In contrast, for >20× coverage ARIBA and KmerResistance made erroneous allele calls for 23/176 (13%) and 6/176 (3%) *bla*<sub>TEM</sub> variants respectively. Above 40× coverage ABRicate was incorrect for one (0.6%), ARIBA for four (2%), KmerResistance for one (0.6%) and SRST2 for zero (0%) simulated *bla*<sub>TEM</sub> alleles.

## Real data

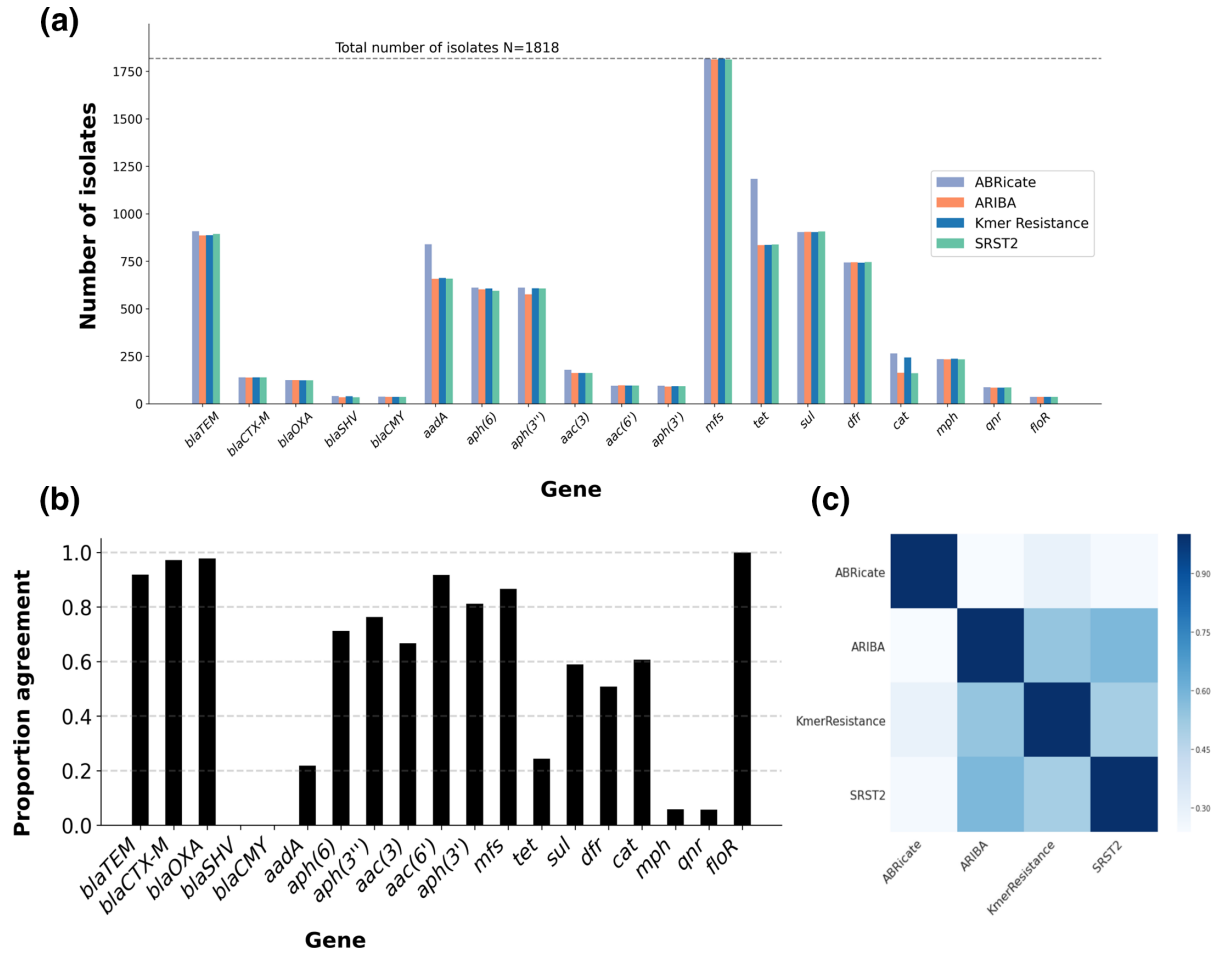
### *E. coli* isolate diversity, antimicrobial susceptibility phenotypes and antimicrobial resistance genotypes

The 1818 isolates were diverse, representing >260 MLSTs, which were differentially distributed among the datasets. For example, although ST131 was the most common [207/1818 (11%) isolates], this was largely due to the fact it was by far the most common in the UKHSA dataset [74/337 (22%) isolates]. In the Oxford dataset, it was only the second most common MLST [123/984 (13%) isolates] after ST73 [161/984 (16%) isolates] and it was rare in the APHA isolates [10/497 isolates (2%)].

Correspondingly, the set also contained a broad range of resistance genes, but the exact number was dependent on the method of search. With there being no definitive way to identify the ‘truth’ of the underlying sequence from short-read data alone, we have *a priori* assumed that no one method is more authoritative than the others. However, it is helpful to pick and report the results from one method as a baseline, to aid describing the dataset in a clear and succinct way. For legibility, we have included results as reported by ABRicate, as this is the most conceptually simple and interrogatable approach (see Table S3 for the results from other methods). The most common AMR-associated sequence identified was *mdfA*. This is known to be universal in *E. coli*, and correspondingly was identified in all 1818 isolates in the dataset. There were no other ubiquitous AMR genes; however, several were common across datasets, with *bla*<sub>TEM</sub>, *aadA*, *sul*, *tet* and *dfr* genes occurring in >40% of the isolates. As expected, more UKHSA isolates contained extended-spectrum beta-lactamase (54/337 vs 94/1481) and carbapenemase (18/337 vs 1/1481) genes ( $P<0.001$ ). Aside from *bla*<sub>TEM</sub>, other beta-lactamases were rare among the APHA dataset. Outside of beta-lactam-associated AMR genes, the Oxford dataset had the lowest proportion of other AMR genes for all the different gene families encountered in this study.

### Genotyping discrepancies

Altogether, 10487 different genes ( $n=15588$  different alleles) were identified in the 1818 isolates by the 4 methods; 1392/1818 (76%) isolates had discrepancies across the 4 bioinformatics methods for at least 1 gene. At the gene level, aside from for the *tet*, *aadA*



**Fig. 3.** Gene identification concordance vs allele identification concordance. (a) The number of isolates containing at least one allele of the name gene families (x-axis) stratified by method. (b) The proportion of times a given gene was identified concordantly by all four methods. (c) Pairwise agreement between the different methods across all isolates.

and *cat* genes, the performance of the bioinformatic tools was similar (Fig. 3a), with tools reporting each gene in approximately the same proportion of isolates (within  $\pm 2\%$ ). With regard to the three outliers, ABRicate reported *tet* and *aadA* genes in 19 and 10% more isolates, respectively, than the other three tools, and ABRicate and KmerResistance reported *cat* genes in 5% more isolates than ARIBA and SRST2. By contrast, the alleles reported by each tool were often discrepant, with alleles of some genes (e.g. *bla<sub>SHV</sub>*, *bla<sub>CMY</sub>*) consistently being differentially reported (Fig. 3b). Consequently, pairwise agreement between any two different tools was  $<59\%$  ( $n=1065$  isolates, Fig. 3c). While unsupported genotype reports (i.e. where the output of one tool was not supported by any other) were common for all tools (Fig. 4), KmerResistance reported fewer unsupported genotypes than the other three tools ( $P<0.001$ ).

### Causes of genotyping discrepancy

At least 2530/4321 (59%) of allele-level discrepancies were due to programs naming the same underlying sequence differently (allele assignment-related differences). We identified three major causes of differences through investigation of discrepantly reported genes: (i) difficulty distinguishing between optimal matches among alleles with nested sequences ( $n=1737$  genes); (ii) spurious identification of additional alleles due to reads being multiply mapped to distant variants of the same allelic family ( $n=547$  genes); and (iii) tools choosing different optimal matches based on DNA sequence alignment when the database only contains one sequence per protein ( $n=197$ ) (Fig. 5). These issues occurred alone in 1944/2530 (77%) discrepantly reported genes and/or in combination in 586/2530 (23%) cases. In isolation, these errors typically caused only a single method to be discordant, but when combined they resulted in more complex patterns of discrepancy and could make all four methods disagree with one another. In addition to allele assignment, ABRicate's more relaxed requirement for complete gene coverage (which aims to mitigate assembly errors) caused at least 1430/4321 (33%) allele-level discrepancies. Discrepancies less easily classified as (but likely related to) allele assignment/cut-offs did occur, but only affected 381/10487 (4%) of reported genotypes.

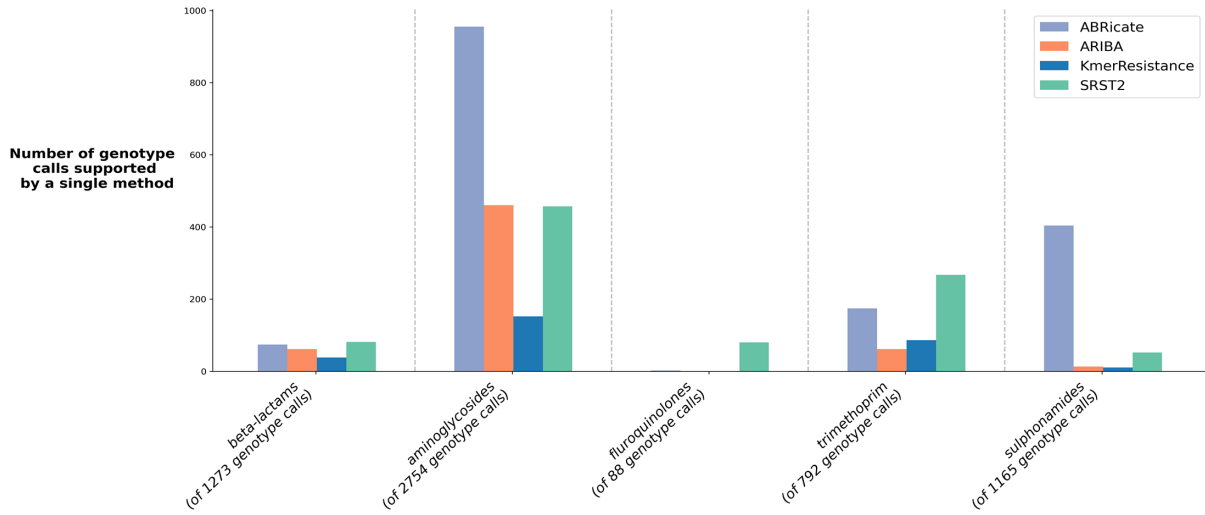


Fig. 4. Genotype calls produced by a single method only, stratified by antibiotic class.

**Allele assignment-related discrepancies**

The most common type of allele assignment-related discrepancy ( $n=1737$  genes) was the result of tools struggling to choose optimal matches where the database contained nested sequences. One such example of this ( $n=24$ ) was caused by the sequences for two different *dfrA7* alleles in the October 2019 Resfinder database, *dfrA7\_1\_AB161450* and *dfrA7\_5\_AJ419170*. The shorter of the two (*dfrA7\_1\_AB161450*, 474 bp long) aligns almost perfectly (percentage identity=99%, 1 single-nucleotide gap) with the first 473 bases of *dfrA7\_5\_AJ419170*. ARIBA, KmerResistance and SRST2, which look for the best identity sequence matches,

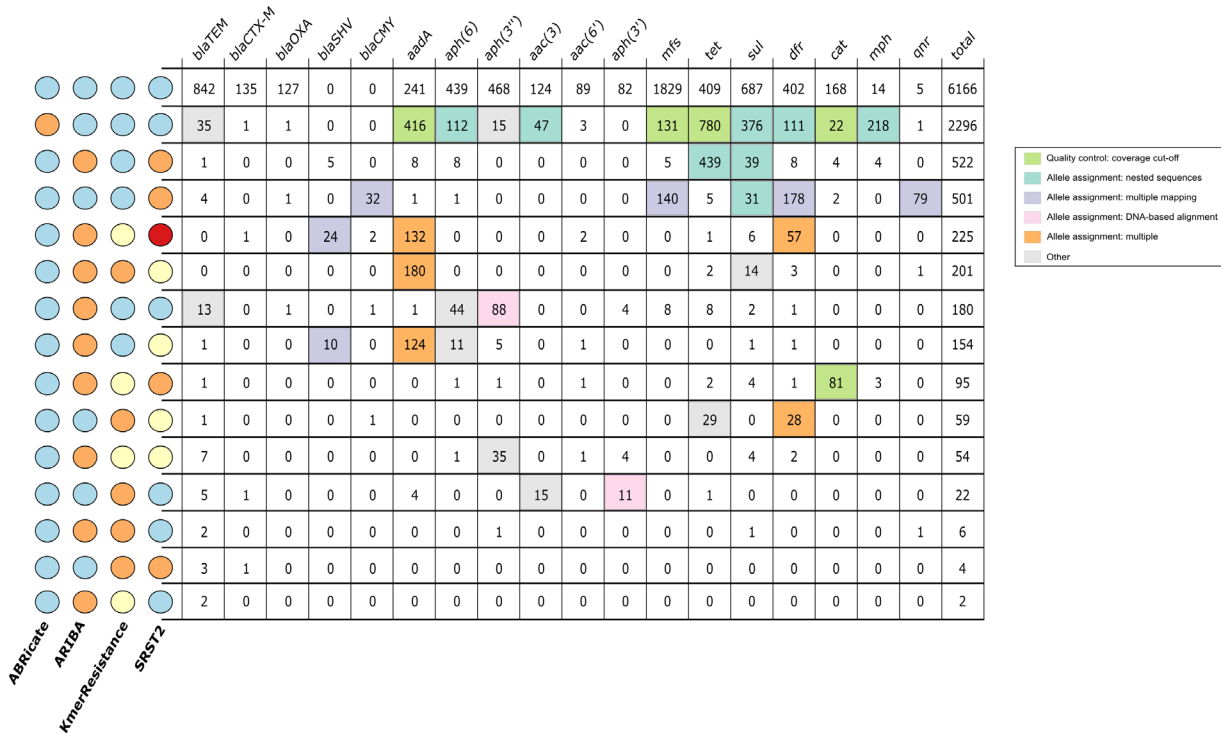


Fig. 5. Genotyping agreement across all four bioinformatics algorithms, stratified by gene. Colours on the left indicate which methods agreed with one another, with circles with the same colour indicating agreement. Colours in the main panel of the figure were used to identify the cause of the discrepancy, as denoted in the figure key. Cells (in the figure) were coloured if >90% of isolates were caused by a given discrepancy. Cells with <10 isolates were not investigated.

all report that the sample contains a perfect match for *dfrA7\_1\_AB161450*. By contrast, ABRicate, which uses BLAST to identify optimal sequences, reports that the sample contains a near perfect match to *dfrA7\_5\_AJ419170*, as with this being a longer match it is more statistically significant. Similar errors occurred for several other genes, including *sul*, *tet*, *aph(6)* and *aac(3)*.

The second most common allele assignment-related discrepancy ( $n=547$  genes) represented tools reporting multiple alleles due to reads mapping to two or more distant variants of the same allelic family. An example observed was ARIBA and SRST2 reporting multiple *bla<sub>SHV</sub>* alleles. In this instance, ARIBA and SRST2 identified a primary perfect allele and a second allele with a lower quality match. These multiple matches, however, were likely spurious, with <10 reads mapping individually to each allele, no clear heterozygosity observed in read pileups, and no fragmentation in assembly graphs. This is the result of how mapping methods identify optimal matches. Both ARIBA and SRST2 map reads to each sequence in the database, and then compare 'closely related' sequences to decide which mapping is optimal. Defining closely related, however, is not straightforward (Fig. S8). Reads mapping to more than one set of closely related sequences can result in tools finding multiple gene variants when the isolate only had one gene original

The final common allele assignment-related discrepancy ( $n=197$  genes) was due to allele reporting based on which sequence in the database had the optimal DNA alignment with the target resistance gene. Although resistance gene nomenclature is largely based on protein sequence, resistance gene databases mostly only catalogue one nucleotide sequence linked to an associated protein sequence. Variant alleles with synonymous mutations fail to perfectly match any element, and may have an alternative optimal DNA match. We observed this on nine occasions where ABRicate, KmerResistance and SRST2 identified imperfect nucleotide-level matches to *aph(3'')-Ib\_2\_AF024602* and ARIBA identified an imperfect match to *aph(3'')-Ib\_4\_AF313472*. However, the sequence they were matching to in the SPAdes and ARIBA assembly was a 100% identity and coverage protein match to *aph(3'')-Ib\_5\_AF321551*.

### Non-allele assignment-related discrepancies

In addition to allele assignment-related discrepancies that were caused by bioinformatics algorithms, genotyping calls were also affected by partial/low coverage of AMR gene targets and assembly fragmentation, consistent with the results from simulations. For some of these, such as the 1430 cut-off-related discrepancies occurring for *tet*, *mfs*, *aadA* and *cat* genes, each program identified the same section of sequence, making it clear that the different programs had different thresholds for reporting, but other situations were less clear. To investigate this in detail, we examined beta-lactamase matches that were either partial/low coverage or occurred across fragmented assemblies.

Partial/low-coverage beta-lactamase genes were discrepantly found in 39 isolates (Fig. S9), particularly affecting *bla<sub>TEM</sub>*-like gene calls (29/39 cases). KmerResistance reported the presence of a beta-lactamase gene in all 39 of these discrepant cases, with calls supported to a varying degree by the other algorithms. However, in all but four cases, KmerResistance reported that the depth of the gene was <5 $\times$ . For the four cases where the gene was present at >5 $\times$  depth as called by KmerResistance, three (present at depth >100 $\times$ ) were omitted from ARIBA reports, as ARIBA assemblies contained mis-sense mutations and the final one (present at depth 17 $\times$ ) also failed to assemble for ABRicate.

Assembly fragmentation affected ABRicate and ARIBA beta-lactam resistance gene calls in 24 cases, with 16 of these likely to be due to the presence of multiple closely related beta-lactamase alleles affecting assembly integrity. The possibility of heterozygous alleles was indicated by the ARIBA flag 'variants\_suggest\_collapsed\_repeat', and the SRST2 'minor allele frequency value' was high (>20%). KmerResistance reported 2 related alleles in 12/16 cases, 1 with high depth, percentage identity and coverage, and 1 much less accurately. This likely reflects KmerResistance's winner-takes-all strategy, where matching unique k-mers to reference alleles are counted, and the reference allele with the most matches is then also assigned all reads with non-unique kmer-matches. This then leaves only reads with unique k-mers matching any closely related secondary allele, resulting in poor depth and coverage metrics.

## DISCUSSION

We evaluated the impact of bioinformatics approaches to AMR genotyping in *E. coli* for four commonly used methods and a widely used AMR gene database (ResFinder). Using >50000 simulations and comparing >1800 sequences sampled across human and animal reservoirs, thereby capturing common and rare AMR genotypes, we highlight that whilst currently available, widely used genotyping methods are useful, their outputs should be carefully considered in light of our findings. Commonly postulated causes of discrepancy, such as low-quality sequencing data, appeared to play little role. Instead, discrepancies were primarily artefactual, occurring because of different approaches in representing the complexity of the reference AMR gene database. Inconsistent labelling of gene variants will also affect the reliability of any catalogue-based methods for phenotypic prediction from WGS-based AMR genotypes. Specifically, predicting phenotype based on the presence of specific allelic variants will be problematic without a reliable method of identification.

Our work agrees with previous findings by Doyle *et al.* on a small and selective dataset [17]; however, we utilized large simulated and real-life datasets to identify these significant genotyping discrepancies between methods, and also characterized the underlying reasons for these discrepancies. We found that most discrepancies were largely due to allele assignment-related differences, i.e. each



method identified the same consensus sequence but then named them differently. Further, many of these discrepancies are caused by implicit and frequently incorrect assumptions about database structure and AMR gene diversity, namely: that AMR genes can be classified in well-defined families using genetic identity, that different approaches to deciding best-matching alleles are equivalent and that isolates will usually not harbour highly genetically related variants of the same AMR gene. Of note, our nomenclature, ‘allele assignment-related discrepancies’, suggests that it is primarily the algorithms that are at fault. However, if database structure was simpler, for example containing fewer variants and avoiding those that are subsequences of one another, there would be fewer ‘allele assignment-related discrepancies’. The assumptions resistance genotypers employ would be less detrimental if database structure was less complicated. However, this highlights a disconnect between the aim of many resistance databases, to catalogue known resistance mechanisms, the naming of which has often evolved organically over the years, and resistance genotypers, to reliably identify resistance gene variants. Sufficiently detailed and standardized curation across resistance databases remains a challenge [32], in part because the nomenclature and AMR gene family structure relevant to Enterobacterales are complicated and there are many variants. For example, highly diverse genotypes (and sometimes phenotypes) are assigned similar family names (e.g. *bla*<sub>CTX-M</sub>, *bla*<sub>OXA</sub>) and solitary single-nucleotide polymorphisms in some cases lead to different resistance phenotypes [e.g. *bla*<sub>TEM-1</sub> (GenBank: AY458016.1) – beta-lactamase inhibitor susceptible, i.e. susceptible to amoxicillin-clavulanate, *bla*<sub>TEM-30</sub> (GenBank: AJ437107.1) – beta-lactamase inhibitor resistant, i.e. resistant to amoxicillin-clavulanate). Further, given that these databases will need to adapt to catalogue more methods of resistance, particularly for the most complex pathogens [33], restricting their scope is not an attractive alternative. Given this, it is not surprising that we found methods that make fewer assumptions (e.g. KmerResistance) to be more robust. Based on our findings, accurate resistance genotyping may require the use of multiple different methods to cross-check results, and a clear understanding of the specific assumptions underlying the methods used, before conclusions about allele presence are drawn. Eventually, the increasingly widespread use of long-read sequencing and the development of new algorithms that cope better with underlying AMR gene diversity in these organisms may make things easier. The ever-improving assemblies obtainable from long-read or combined short and long read data will make the underlying sequence data simpler to interrogate. While it may not resolve naming differences due to the complexity of AMR nomenclature, it may make them substantially easier to identify.

One of the key strengths of this analysis was its combined use of both simulations and real-world data. By using simulations, we were able to benchmark methods against a known truth, which is impossible to do with real-world data. Previous studies using only real-world data have attempted to overcome the absence of complete knowledge of the underlying genotype by using phenotypic data as a reference standard; however genotype–phenotype correlations remain poorly defined [14, 27] By subsequently using a large sequencing dataset of isolates obtained across niches, we were then able to assess the extent of discrepancies in real life, replicating the problems observed in simulated data.

A limitation of this work is that we chose not to evaluate the impact of database choice, and this will represent future work. Currently, as has been highlighted previously [34], there are discrepancies between the AMR databases in common use, with each having a slightly different scope and in some cases differential names for different AMR gene variants [e.g. *strA* vs *aph(6)-Ia* or *aphD* and *strB* versus *aph(6)-Id*]. Comparing databases would have therefore added significant further complexity whilst limiting the generalizability of findings. A further limitation stemming from our fixed choice of database is that we have not analysed any methods where the bioinformatic method and database are intertwined (e.g. ResFinder/PointFinder or RGI). As the interaction between tool and database was the cause of many issues, it is possible that methods that are database-specific will perform better. However, the drawbacks of these combined resources are their inflexibility, again limiting generalizability. A further limitation was that these genotyping algorithms were compared using an older version of the ResFinder database – the most up to date when this work was originally planned. Since this time, 83 changes have been documented on the Resfinder GitHub repository (73 additions, 6 deletions, 4 modifications). The additional sequences were rare in our real-world isolates, accounting for only 54 (<1%) of the genes called. We opted not to reperform the analysis due to its manual nature and because most of the discrepancies relate to underlying principles behind the algorithms rather than the specific implementation. Finally, we have focused our evaluation on *E. coli*, but it is likely that these issues will also more widely affect AMR genotyping, particularly of similar species with complex genotypes.

While WGS-based approaches are attractive for both characterizing AMR gene epidemiology and as a subsequent tool for resistance prediction, this work highlights the need for caution when interpreting resistance genotypes reported by even widely used bioinformatics methods. Before WGS-based approaches can be considered to be reliable for use in *E. coli* (and likely other Enterobacterales), particularly for clinical decision-making or replacing phenotypic data to determine epidemiological trends, database standardization, the development of novel genotyping approaches, and improved validation and evaluation will be required.

#### Funding information

The study was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) [NIHR200915]. D.W.C., T.E.A.P., P.W.F. and A.S.W. are supported by the NIHR Oxford Biomedical Research Centre. The report presents independent research funded by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or Public Health England. N.S. is an Oxford Martin Fellow and an NIHR Oxford BRC Senior Fellow. A.S.W. is an NIHR Senior Investigator.

## Acknowledgements

We are grateful to the microbiology laboratory teams at the John Radcliffe Hospital, Oxford, the Animal and Plant Health Agency, and UK Health Security Agency.

## Author contributions

T.J.D., N.S., A.E.S., A.S.W., D.W.C. and T.E.A.P. conceptualized the study. T.D., N.S., A.S.W., A.E.S. and M.F.A. decided the methodology. N.S., A.S.W., M.F.A., A.E.S., D.W.C. and T.E.A.P. supervised the project. N.S., M.A., M.F.A., M.J.E., K.H. and S.H. acquired and curated the data used in this study. T.J.D. and J.S.W. constructed software pipelines to analyse sequencing data using each of the bioinformatic tools. T.J.D. and A.S.W. investigated the data. T.J.D. performed the formal analysis. N.S., A.E.S., S.L., H.P., A.E.S. and T.E.A.P. assisted with interpreting the cause and impact of discrepancies. T.J.D. and N.S. wrote the original draft. T.J.D., N.S., A.E.S., P.W.F., T.E.A.P. and A.S.W. assisted with data visualization. All authors were involved in the review and editing process.

## Conflicts of interest

The authors declare that there are no conflicts of interest.

## References

- Quainoo S, Coolen JPM, van Hijum S, Huynen MA, Melchers WJG, et al. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clin Microbiol Rev* 2017;30:1015–1063.
- Quan TP, Bawa Z, Foster D, Walker T, Del Ojo Elias C, et al. Evaluation of whole-genome sequencing for mycobacterial species identification and drug susceptibility testing in a clinical setting: a large-scale prospective assessment of performance against line probe assays and phenotyping. *J Clin Microbiol* 2017;56:e01480-17.
- Sherry NL, Horan KA, Ballard SA, Gonçalves da Silva A, Gorrie CL, et al. An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. *Nat Commun* 2023;14:1–12.
- ISO 15189:2022. Medical laboratories — Requirements for quality and competence; (n.d.). <https://www.iso.org/standard/76677.html> [accessed 7 November 2023].
- Guidance for Industry and FDA Class II Special Controls Guidance Document: Antimicrobial Susceptibility Test (AST) Systems; 2018. <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocument> [accessed 11 April 2019].
- Verschuuren T, Bosch T, Mascaro V, Willems R, Kluytmans J. External validation of WGS-based antimicrobial susceptibility prediction tools, KOVER-AMR and ResFinder 4.1, for *Escherichia coli* clinical isolates. *Clin Microbiol Infect* 2022;28:1465–1470.
- Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* 2020;75:3491–3500.
- Brian PA, Amogelang RR, Tammy TY, Kara TK, Mégane B, et al. CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2020:D517–D525.
- Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, et al. Using the NCBI AMRFinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of NARMS isolates. *Antimicrob Agents Chemother* 2019:00483-19.
- Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3:e000131.
- Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014;6:90.
- Clausen P, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J Antimicrob Chemother* 2016;71:2484–2488.
- Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, et al. Point-Finder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother* 2017;72:2764–2768.
- Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *J Antimicrob Chemother* 2013;68:2234–2244.
- Shelburne SA, Kim J, Munita JM, Sahasrabhojane P, Shields RK, et al. Whole-genome sequencing accurately identifies resistance to extended-spectrum  $\beta$ -lactams for major gram-negative bacterial pathogens. *Clin Infect Dis* 2017;65:738–745.
- Stubberfield E, AbuOun M, Sayers E, O'Connor HM, Card RM, et al. Use of whole genome sequencing of commensal *Escherichia coli* in pigs for antimicrobial resistance surveillance, United Kingdom, 2018. *Euro Surveill* 2019;24:1900136.
- Doyle RM, O'Sullivan DM, Aller SD, Bruchmann S, Clark T, et al. Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microb Genom* 2020;6:e000335.
- Seemann T. ABRicate; 2020. <https://github.com/tseemann/abricate>. <https://github.com/tseemann/abricate> [accessed 19 October 2023].
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
- van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* 2011;13:22–30.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.
- Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28:593–594.
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
- Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect* 2017;23:2–22.
- Davies TJ, Stoesser N, Sheppard AE, Abuoun M, Fowler P, et al. Reconciling the potentially irreconcilable? Genotypic and phenotypic amoxicillin-clavulanate resistance in *Escherichia coli*. *Antimicrob Agents Chemother* 2020;64:e02026-19.
- AbuOun M, Stubberfield EJ, Duggett NA, Kirchner M, Dormer L, et al. mcr-1 and mcr-2 variant genes identified in *Moraxella* species isolated from pigs in Great Britain from 2014 to 2015. *J Antimicrob Chemother* 2017;72:2745–2749.
- Zerbino DR. Using the velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* 2010;11.
- Stoesser N, Sheppard AE, Peirano G, Anson LW, Pankhurst L, et al. Genomic epidemiology of global *Klebsiella pneumoniae* carbapenemase (KPC)-producing *Escherichia coli*. *Sci Rep* 2017;7:5917.
- Wirth T, Falush D, Lan R, Colles F, Mensa P, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.

32. Papp M, Solymosi N. Review and comparison of antimicrobial resistance gene databases. *Antibiotics* 2022. detection and prediction from *Pseudomonas aeruginosa*. *medRxiv* 2022:22278689.
33. Madden DE, Baird T, Bell SC, McCarthy KL, Price EP, *et al.* Keeping up with the pathogens: improved antimicrobial resistance
34. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. *Ann N Y Acad Sci* 2017;1388:78–91.

**The Microbiology Society is a membership charity and not-for-profit publisher.**

**Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.**

**Find out more and submit your article at [microbiologyresearch.org](https://microbiologyresearch.org)**