

Whole Genome Sequencing and Resistance in Escherichia coli



Timothy J Davies

Linacre College, Oxford University

Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2019

Abstract

Thesis title: Whole Genome Sequencing and Resistance in Escherichia coli

Student name: Timothy J Davies

College: Linacre College, Oxford University

Degree: Doctor of Philosophy

Submitted: Trinity Term 2019

Rising antimicrobial resistance is an increasing problem worldwide. *Escherichia coli* presents a particular challenge as it spreads resistance rapidly via mobile genetic elements. Combatting this requires fast, detailed identification of resistance to prevent its spread and inform treatment. Culture-based antimicrobial susceptibility testing (AST) is limited by the time required for bacterial growth and does not directly identify mechanisms of resistance. However, many molecular techniques require specific targets, limiting their routine use in clinical microbiology. Whole genome sequencing (WGS) offers an alternative, solving many of these issues. However, evidence of WGS's ability to identify clinically significant resistance in *E. coli* is limited. This thesis investigates this in detail.

Using methods similar to those from "proof-of-principle" studies, beta-lactam resistance was predicted from WGS data in several sets of *E. coli*, including a large set of unselected bloodstream infection isolates. While agreement between WGS-predicted and observed AST was high for many antibiotics, for some, including several commonly used broad spectrum antibiotics, there was significant discrepancy. Focussing on amoxicillin-clavulanate resistance due to the extent of disparity and its clinical significance, the causes of discrepancy were investigated. Instead of the discrepancy being due to the bioinformatic approach, incomplete knowledge of resistance mechanisms or phenotyping errors as is commonly believed, results suggested oversimplifications of both phenotype and genotype were the major cause. Future work is needed to investigate WGS-based prediction for other antibiotic and antimicrobial combinations.

This work demonstrates the feasibility of WGS-based prediction for a phenotype widely regarded as one of the most complex. However, it also highlights many barriers WGS-based prediction will need to overcome prior to clinical implementation. These include not only technical problems such as how to best curate and use genetic data for resistance prediction, but also non-technical problems such as the need to challenge long held assumptions regarding the nature of antimicrobial susceptibility testing.

Table of Contents

Chapter 1: Preface	9
1.1. Acknowledgments.....	9
1.2. Funding.....	10
1.3. Declaration of contribution to thesis	11
1.4. Attributions.....	11
1.5. Publications	14
1.6. Abbreviations.....	16
Chapter 2: Introduction	21
2.1. Improving diagnostics to combat antimicrobial resistance	21
2.2. Background	23
2.2.1. <i>Escherichia coli</i> : an archetypal Gram-negative organism	23
2.2.2. Beta-lactams	25
2.2.3. Beta-lactam resistance in <i>E. coli</i> and other Gram-negatives	31
2.2.4. Antimicrobial susceptibility testing	40
2.3. WGS-based prediction of AST phenotypes.....	48
2.3.1. Approaches to predicting susceptibility using WGS	48
2.3.2. Rule-based resistance prediction	50
2.3.3. Non-rule-based resistance prediction	53
2.3.4. Specific issues surrounding beta-lactam resistance prediction in <i>E. coli</i>	55
2.3.5. Published comparisons of WGS-based and observed phenotypes	58
2.4. Thesis outline.....	61
Chapter 3: Description of data and methods.....	63
3.1. Introduction	63
3.2. Obtaining datasets.....	63
3.2.1. OUH dataset	63
3.2.2. PHE dataset.....	64
3.2.3. APHA dataset	65
3.3. Routine processing.....	65
3.3.1. Sequencing	65
3.3.2. Phenotyping.....	69
3.3.3. Analysis.....	74
3.4. Dataset description.....	74
3.4.1. Relatedness of isolates	74
3.4.2. Beta-lactam resistance genes	79
3.4.3. Isolate phenotypes.....	80
3.4.4. Strengths and limitations of datasets.....	86
Chapter 4: Large-scale testing of the performance of WGS-based resistance prediction in <i>Escherichia coli</i>	89
4.1. Introduction	89
4.2. Methods.....	90
4.2.1 Baseline data and resistance prediction.....	90
4.2.2. Investigating potential genotypic causes of discrepancy	93
4.2.3. Investigating potential prediction algorithm causes of discrepancy	96
4.2.4. Understanding associations between phenotype and genotype.....	97

4.3. Results	97
4.3.1. Baseline prediction.....	97
4.3.2. Potential genotypic causes of poor predicted versus observed phenotype agreement	100
4.3.3. Prediction algorithm causes of discrepancy	110
4.3.4. Evaluating associations between genetic elements and phenotype.....	114
4.4. Discussion	116
<i>Chapter 5: Pitfalls of identifying the resistance genotype from WGS data</i>	<i>124</i>
5.1. Introduction	124
5.1.1. Background.....	125
5.2. Methods.....	125
5.2.1. Comparing resistance genotypes in real data.....	125
5.2.2. Investigating discrepant using simulations	133
5.2.3. Assessing combined approaches on real world data	144
5.3. Results	146
5.3.1. Comparing resistance genotypes.....	146
5.3.2. Characterising genotypic discrepancies.....	148
5.3.3. Understanding discrepancies using simulations	159
5.3.4. Impact of bioinformatic method on beta-lactam resistance prediction.....	162
5.4. Discussion	165
<i>Chapter 6: Moving to a more realistic model of beta-lactam resistance.....</i>	<i>171</i>
6.1. Introduction	171
6.1.1. Background and rationale.....	171
6.2. Methods.....	175
6.2.1. Sample baseline selection, antimicrobial susceptibility, and sequencing data.....	175
6.2.2. Resistance genotyping.....	176
6.2.3. Resistance prediction from genotype	182
6.3. Results	184
6.3.1. Samples and phenotypes.....	184
6.3.2. Examining features of the resistance genotype.....	184
6.3.3. Evaluating the impact of the extended features on resistance prediction.....	198
6.4. Discussion	200
<i>Chapter 7: The problem of binary interpretation of a continuous phenotype</i>	<i>207</i>
7.1. Introduction	207
7.1.1. The amoxicillin-clavulanate phenotype.....	207
7.2. Methods.....	214
7.2.1. Sample selection	214
7.2.2. Phenotyping.....	217
7.2.3. Comparison of phenotypes.....	219
7.2.4. Comparison with genotype.....	219
7.3. Results	230
7.3.1. Samples	230
7.3.2. Evaluating phenotyping methods	231
7.3.3. Comparing agar dilution phenotypes with WGS-based predictions.....	235
7.3.3. Discrepancies between WGS-predicted and observed phenotypes	238
7.3.4. Modelling MIC as a function of resistance features.....	239
7.3.5. Predictions of MIC in an independent validation set	243
7.4. Discussion	244

Chapter 8: Discussion	252
8.1. Introduction	252
8.2. Summary of findings	253
8.2.1. Predicting beta-lactam antibiograms using gene-presence absence alone.....	253
8.2.2. The choice of WGS-based method.....	253
8.2.3. Accounting for alternate mechanisms of beta-lactam resistance.....	254
8.2.4. Phenotype method dependency of WGS versus AST comparisons.....	255
8.2.5. Quantitative approaches	255
8.3. Interpretation	256
8.3.1. WGS-based resistance prediction and assumptions about resistance.....	256
8.3.2. Models of amoxicillin-clavulanate resistance	259
8.4. Implications	261
8.4.1. Importance of findings for clinical practice	262
8.4.2. Importance of findings for different bacteria, antimicrobials and settings.....	262
8.4.3. Importance of findings for applicability of WGS-based resistance prediction.....	263
8.4.4. Importance of findings for resistance databases	265
8.4.5. Importance of findings for beta-lactam resistance epidemiology	265
8.5. Strengths and limitations	267
8.5.1 Strengths	267
8.5.2. Limitations	268
8.6. Suggested future work.....	271
8.6.1. Moving WGS into clinical practice.....	271
8.6.2. Extending WGS's utility	273
8.6.3. Beta-lactam resistance epidemiology	276
References	278
Appendices	310

List of Figures

FIGURE 1: PAN-GENOME OF <i>E. COLI</i>	25
FIGURE 2: NORMAL CELL WALL CONSTRUCTION OF A) <i>STAPHYLOCOCCUS AUREUS</i> , B) <i>E. COLI</i>	27
FIGURE 3: THE MIMICRY OF BETA-LACTAM ANTIBIOTICS TO D-ALANYL-D-ALANINE (D-ALA-D-ALA)	28
FIGURE 4: STRUCTURE OF BETA-LACTAM CLASSES	29
FIGURE 5: THE KYOTO ENCYCLOPEDIA OF GENES AND GENOMES (KEGG) PENICILLINS REFERENCE PATHWAY	30
FIGURE 6: INCREASE IN THE NUMBER OF NATURALLY OCCURRING BETA-LACTAMASES BY A) MOLECULAR CLASS, B) MAJOR FUNCTIONAL GROUPS	32
FIGURE 7: MOLECULAR AND STRUCTURAL CLASSES OF BETA-LACTAMASES, TOGETHER WITH REPRESENTATIVE MEMBERS AND SUBSTRATE PROFILES	33
FIGURE 8: ESBL <i>E. COLI</i> CARRIAGE RATES IN THE COMMUNITY, ACCORDING TO THEIR GEOGRAPHICAL AND TEMPORAL DISTRIBUTION	35
FIGURE 9: GLOBAL DISTRIBUTION OF CARBAPENEMASES IN ENTEROBACTERIACEAE, BY COUNTRY AND REGION.	36
FIGURE 10: DIFFERENCES BETWEEN GRAM-POSITIVE AND GRAM-NEGATIVE PERMEABILITY BARRIERS	38
FIGURE 11: CATEGORISATION OF WGS-BASED RESISTANCE PREDICTION APPROACHES	49
FIGURE 12: RULE-BASED RESISTANCE PREDICTION	52
FIGURE 13: HISTOGRAMS OF SEQUENCING QC METRICS	67
FIGURE 14: AGAR DILUTION METHOD	70
FIGURE 15: MIC DISTRIBUTIONS AND CLINICAL BREAKPOINTS FOR TESTED BETA-LACTAMS (N=1818 ISOLATES)	73
FIGURE 16: MASH DISTANCES OF ISOLATES	76
FIGURE 17: MLSTS IN EACH DATASET	77
FIGURE 18: RELATEDNESS OF ISOLATES WITHIN DATASETS	78
FIGURE 19: TEN MOST COMMON BETA-LACTAM ANTIBIOGRAMS FOR EACH DATASET	84
FIGURE 20: STRUCTURAL SIMILARITY OF 3RD GENERATION CEPHALOSPORINS CEFTRIAXONE AND CEFTAZIDIME	85
FIGURE 21: ASSOCIATION BETWEEN CORE-GENOME SNP DISTANCE AND MASH SIMILARITY FOR HIGHLY RELATED ISOLATES	88
FIGURE 22: BETA-LACTAMASE GROUPING AND CONSEQUENT RESISTANCE PREDICTION	92
FIGURE 23: A) SENSITIVITY AND B) SPECIFICITY OF WGS-BASED RESISTANCE PREDICTION BY ISOLATE SET	101
FIGURE 24: SEARCH ISSUES CAUSED BY POORLY CURATED SEQUENCE	103
FIGURE 25: EXAMPLES OF "BUBBLES" IN ASSEMBLY GRAPHS FOR BLATEM AND BLACMY GENES	108
FIGURE 26: COMPARING RESISTANCE PREDICTIONS FOR DIFFERENT ASSEMBLERS AND CUT-OFFS	109
FIGURE 27: IMPACT OF KNOWN POTENTIAL ALGORITHM PROBLEMS ON RESISTANCE PREDICTION	112
FIGURE 28: ASSOCIATION BETWEEN BETA-LACTAMASES AND PHENOTYPES	115
FIGURE 29: PROCESSING REQUIRED FOR WGS-BASED METHODS VERSUS TRADITIONAL AST	119
FIGURE 30: OVERVIEW OF SRST2 APPROACH	129
FIGURE 31: OVERVIEW OF THE ARIBA APPROACH.	130
FIGURE 32: SIMULATION PLAN	134
FIGURE 33: CHOOSING ASSEMBLER FOR SIMULATIONS USING ABRICATE A) PERFORMANCE WITH SINGLE-KMER VELVET ASSEMBLIES VERSUS VELVETOPTIMISER B) PERFORMANCE WITH VELVETOPTIMISER VERSUS SPADES AND SHOVILL	136
FIGURE 34: IDENTIFYING ELEMENTS FROM THE RESFINDER DATABASE IN ISOLATION	139
FIGURE 35: ASSOCIATION BETWEEN JACCARD INDEX FROM 31-MERS VS A) 11-MERS B) 17-MERS AND C) 21-MERS	141
FIGURE 36: COMPARING SIMILARITY METRICS A) JACCARD INDEX, OVERALL AND COMPARED WITH B) NEEDLEMAN WUNCH AND C) BLASTN PERCENTAGE IDENTITY	142
FIGURE 37: BETA-LACTAMASE DEPTH OF COVERAGE FROM REAL (NON-SIMULATED) SAMPLES	143
FIGURE 38: METHODS OF COMBINING PREDICTED PHENOTYPES ACROSS THE 4 BIOINFORMATIC TOOLS	145
FIGURE 39: CONCORDANCE OF RESISTANCE GENOTYPING METHODS BY ANTIBIOTIC CLASS	147
FIGURE 40: FREQUENCY OF UNSUPPORTED GENOTYPES (SUGGESTED BY A SINGLE METHOD ALONE) BY ANTIBIOTIC AND BIOINFORMATIC METHOD	148
FIGURE 41: THE IMPACT OF THRESHOLD ON DATABASE CLUSTERING	151
FIGURE 42: PAIRWISE ALIGNMENT OF TWO GENES IDENTIFIED BY SRST2 IN THE SAME ISOLATE	152
FIGURE 43: KMERRESISTANCE REPORTED BETA-LACTAMASE GENE DEPTH	154

FIGURE 44: READS ASSIGNED TO SECONDARY INCOMPLETE BLA _{TEM} ALLELE ONLY BY KMERRESISTANCE	156
FIGURE 45: SIMULATING THE IMPACT OF COVERAGE ON RESISTANCE ALLELE IDENTIFICATION	160
FIGURE 46: SIMULATION SCORE BY PAIR SIMILARITY FOR ABRICATE	161
FIGURE 47: SENSITIVITY AND SPECIFICITY OF COMBINATION APPROACHES FOR PREDICTING AMPICILLIN RESISTANCE	163
FIGURE 48: SENSITIVITY AND SPECIFICITY OF COMBINATION APPROACHES FOR PREDICTING AMOXICILLIN-CLAVULANATE RESISTANCE	164
FIGURE 49: CHEMICAL STRUCTURE OF A) BENZYL PENICILLIN B) AMOXICILLIN AND C) POTASSIUM CLAVULANATE	171
FIGURE 50: IMPACT OF QUANTITY OF TEM-1 BETA-LACTAMASE ON LEVELS OF RESISTANCE IN 36 E. COLI	174
FIGURE 51: LENGTH ADJUSTED DEPTH OF COVERAGE OF TRANSMISSIBLE BETA-LACTAMASES FROM ARIBA	178
FIGURE 52: THE LARTIGUE P3 PROMOTER ²¹⁸ , AND SEQUENCE DIFFERENCES OF OTHER NAMED BLA _{TEM} PROMOTERS	181
FIGURE 53: AMPC PROMOTER REGION AND SEQUENCE MUTATIONS THOUGHT TO CAUSE CONSTITUTIVE AMPC HYPER-EXPRESSION	182
FIGURE 54: BETA-LACTAMASE COPY NUMBERS	187
FIGURE 55: ASSOCIATION BETWEEN COPY NUMBER AND AMOXICILLIN-CLAVULANATE MIC	188
FIGURE 56: COPY NUMBERS IN AMOXICILLIN-CLAVULANATE RESISTANT AND SUSCEPTIBLE ISOLATES	189
FIGURE 57: ROC CURVE COMPARING DNA COPY NUMBER WITH RESISTANCE PREDICTION	190
FIGURE 58: EXTENDED ALGORITHM RESISTANCE PREDICTION BY AMOXICILLIN-CLAVULANATE MIC	200
FIGURE 59: E. COLI MIC DISTRIBUTION FOR AMOXICILLIN-CLAVULANATE	206
FIGURE 60: CHECKERBOARD ASSAYS FOR COMBINATION ANTIBIOTICS	209
FIGURE 61: TYPICAL MICS TESTED FOR COMBINATION ANTIBIOTICS	210
FIGURE 62: MIC DISTRIBUTION A) AMOXICILLIN AND B) AMOXICILLIN-CLAVULANATE.	212
FIGURE 63: AMOXICILLIN-CLAVULANATE MIC DISTRIBUTION FOR ESBL-PRODUCING E. COLI TESTED BY DIFFERENT METHODS	214
FIGURE 64: SAMPLING FRAME FOR ISOLATES SELECTED FOR DETAILED AGAR DILUTION PHENOTYPING.	216
FIGURE 65: STRATIFIED RANDOM SAMPLING STRATEGY FOR DETAILED AGAR DILUTION PHENOTYPING	217
FIGURE 66: INHERENT CLUSTERING OF PHENOTYPIC TESTS	221
FIGURE 67: QQ PLOTS CHECKING RANDOM EFFECT AND RESIDUAL NORMALITY.	223
FIGURE 68: SIZE OF BATCH ERRORS WITH 95% CONFIDENCE INTERVAL	224
FIGURE 69: COMPARISON OF MICS FROM DIFFERENT PHENOTYPIC METHODS	232
FIGURE 70: RESISTANCE CATEGORY SWITCHING ON REPEAT TESTING	234
FIGURE 71: PROPORTION WGS PREDICTED RESISTANT (EXTENDED ALGORITHM) BY EUCAST-BASED AND CLSI-BASE AGAR DILUTION MIC	236
FIGURE 72 VARIABILITY OF RESISTANCE CLASSIFICATION OF ISOLATES CONTAINING ONLY LOW* COPY NUMBER BLA _{TEM-1}	237
FIGURE 73: CHANGES IN DOUBLING DILUTION MIC INDEPENDENTLY ASSOCIATED WITH EACH FEATURE/TESTING METHOD	242
FIGURE 74: MODEL BASED MIC PREDICTION FOR NON-SUBSAMPLE ISOLATES (N=704)	244
FIGURE 75: DISTRIBUTION OF MICS OF E. COLI TO TRIMETHOPRIM.	247
FIGURE 76: DISTRIBUTION OF MICS OF E. COLI TO CEFTOLOZANE-TAZOBACTAM	248
FIGURE 77: ASSUMPTION-BASED INFERENCE FROM GENOTYPE TO OUTCOME	257
FIGURE 78: MIC DISTRIBUTIONS UNDER A "SINGLE-MECHANISM" MODEL.	258
FIGURE 79: MICS IN MULTI-GENE RESISTANCE	260

List of Tables

TABLE 1: SUMMARY OF CLSI CATEGORY DEFINITIONS	45
TABLE 2: SUMMARY OF EUCAST CATEGORY DEFINITIONS	46
TABLE 3: COMPARISON OF ISO AND FDA STANDARDS FOR EVALUATION OF THE PERFORMANCE OF ANTIMICROBIAL SUSCEPTIBILITY TESTING DEVICES	47
TABLE 4 ADVANTAGES/DISADVANTAGES OF ASSEMBLY AND MAPPING BASED APPROACHES	51
TABLE 5:SUMMARY OF PERFORMANCE OF WGS-BASED BETA-LACTAM RESISTANCE PREDICTION FOR E. COLI	59
TABLE 6: ISOLATE SEQUENCING QUALITY CONTROL	68
TABLE 7: EUCAST CLINICAL BREAKPOINTS VS ECOFFS	72
TABLE 8: BETA-LACTAMASE GENES IDENTIFIED ACROSS DATASETS	80
TABLE 9: ISOLATE BETA-LACTAM PHENOTYPES	82
TABLE 10: BETA-LACTAMASES BY DATABASE	94
TABLE 11: TESTED ABRICATE PARAMETERS	95
TABLE 12:INITIAL PREDICTION RESULTS	99
TABLE 13: CHANGES IN VME/ME RATES WITH METHOD CHANGES	113
TABLE 14: COMMONLY USED METHODS FOR SEARCHING WGS DATA FOR RESISTANCE SEQUENCES	127
TABLE 15: ART PARAMETERS	143
TABLE 16: CAUSES OF DISCREPANCY OF BETA-LACTAM GENOTYPES	149
TABLE 17: ISOLATES CONTAINING LOW COVERAGE/PARTIAL GENES VARIABLY IDENTIFIED BY DIFFERENT METHODS	155
TABLE 18: CAUSES OF DISCREPANCIES FOR NON-BETA-LACTAM GENOTYPES	158
TABLE 19: MULTIPLE ALLELE SIMULATION PERFORMANCE	161
TABLE 20: AST BETA-LACTAM ANTIBIOGRAM	184
TABLE 21: BETA-LACTAMASES IDENTIFIED	186
TABLE 22: BLATEM PROMOTERS, LARTIGUE TYPES AND AMOXICILLIN-CLAVULANATE RESISTANCE	191
TABLE 23: AMPC PROMOTER MUTATIONS AND AMOXICILLIN-CLAVULANATE RESISTANCE	194
TABLE 24: OMPF DISRUPTION MECHANISM, CONCOMITANT BETA-LACTAM RESISTANCE FEATURES AND BETA- LACTAM ANTIBIOGRAM	196
TABLE 25: OMPC DISRUPTION MECHANISM, CONCOMITANT BETA-LACTAM RESISTANCE FEATURES AND BETA- LACTAM ANTIBIOGRAM	197
TABLE 26: POSITIVE PREDICTIVE VALUE OF INDIVIDUAL GENETIC FEATURES FOR PREDICTING AMOXICILLIN- CLAVULANATE RESISTANCE	198
TABLE 27: CATEGORIZATION OF BETA-LACTAMASES	225
TABLE 28: MODELLING METHODS FOR EACH GENETIC FEATURE	227
TABLE 29: COMPARISON OF AGAR DILUTION AND FULL OUH DATASETS	231
TABLE 30: SIMULATED COMPARISONS OF EACH TEST AGAINST ITSELF AS A REFERENCE	235
TABLE 31: UNADJUSTED EFFECT OF RESISTANCE FEATURES ON AGAR DILUTION LOG ₂ MIC	240
TABLE 32: INDEPENDENT PREDICTORS OF AGAR DILUTION LOG ₂ (MIC)	241
TABLE 33: ESTIMATED VARIATION IN MIC ACCORDING TO DIFFERENT SOURCES FROM THE RANDOM EFFECTS MODEL	243

Chapter 1: Preface

1.1. Acknowledgments

As with many scientific endeavours, the work in this thesis would not have been possible without the support and guidance of my supervisors, colleagues and peers at MMM. In particular I would like to thank my supervisors, Sarah Walker, Tim Peto, Muna Anjum, Nicole Stoesser and Neil Woodford not only for their mentorship but also their seemingly endless patience when managing my doubts and fears as I learn the basics of research. Outside of my supervisory team, several individuals have been instrumental in making my work possible: Derrick Crook for his wise microbiological advice, Anna Sheppard for the key role she has played helping me understand the fundamentals bacterial biology and bioinformatics, Jeremy Swann for always answering my coding “SOS” calls, and both Manal AbuOun and Dai Griffiths for looking after me in the laboratory. I would also like to profoundly thank my family, especially my wife Laura and daughter Isabel who have been incredibly understanding, or at least as understanding as a 18 month old gets in Isabel’s case, about my work. They have always given unending support throughout my DPhil, keeping me going in everything I do.

1.2. Funding

This DPhil 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) [grant HPRU-2012-10041].

1.3. Declaration of contribution to thesis

I, Timothy Davies, designed and conducted all the analyses presented in this thesis, with the support of my supervisors and colleagues.

Specific assistance provided by others in relation to work in this thesis is outlined below.

1.4. Attributions

Chapter 3 – “Description of data and methods”

Isolate collections used for the thesis were supplied by Oxford University Hospital NHS Foundation Trust (OUH) microbiology laboratory, Public Health England’s (PHE) Antimicrobial Resistance and Healthcare Associated Infections reference unit, and the UK Animal and Plant Health Agency’s (APHA) Bacteriology department. Collection and sequencing of isolates from the OUH microbiology laboratory was undertaken by the Crook/Peto laboratory staff. Sequencing of PHE and APHA isolates was also performed by the Crook/Peto laboratory staff. I assisted with isolate sequencing and collection. The Sequencing Hub at the Wellcome Trust Centre for Human Genetics undertook library preparation and sequencing on the Illumina HiSeq. All raw read data were routinely processed through the Crook/Peto research group’s bioinformatic pipeline to obtain mapped consensus sequences and *de novo* assemblies. I selected and applied appropriate sequencing quality control (QC) thresholds. Phenotyping of isolates from the OUH microbiology laboratory was performed by NHS staff. Phenotyping from PHE was performed by David Griffiths and Nicole Stoesser. Phenotyping of APHA isolates was performed by Manal AbuOun and Emma Stubberfield. Phuong Quan extracted minimum inhibitory concentration (MIC) results for the OUH and PHE isolates from the Infections in Oxfordshire Research Database. I performed the resistance genotyping and comparative analyses of

isolates. Computing resources were provided and maintained by the Crook/Peto informatic team and the Wellcome Trust Centre for Human Genetics.

Chapter 4 – “Large-scale testing of the performance of WGS-based resistance prediction in Escherichia coli”

I researched the area and designed the study. I selected the bioinformatic method and resistance databases. I designed the resistance prediction algorithm with advice from Nicole Stoesser and Hang Phan. Together with Jeremy Swann I created a resistance genotyping pipeline (using <https://www.nextflow.io/>) to enable easy and rapid scaling. I performed non-pipeline re-assembly of isolates. I performed the analysis of resistance prediction algorithm performance.

Chapter 5 – “Pitfalls of identifying the resistance genotype from WGS data”

I researched the area and designed the study. I selected the bioinformatic methods for testing. Together with Jeremy Swann I created resistance genotyping pipelines for each of the bioinformatic methods. I ran all programs and analysed the results. I designed the simulations with advice from Martin Hunt and Anna Sheppard. I created the combined resistance predictions and analysed their performance.

Chapter 6 – “Moving to a more realistic model of beta-lactam resistance”

I researched the area, designed the study, and identified the additional resistance elements for search with advice of Nicole Stoesser. I wrote the scripts enabling the interrogation of sequence data for extra elements. Anna Sheppard advised me on detecting porin gene loss of function from sequencing data. I performed the receiver operator curve analysis and

constructed the resistance prediction algorithm. I analysed and compared the performance of resistance prediction using the extended set of features.

Chapter 7 – “The problem of binary interpretation of a continuous phenotype”

I researched the area and designed the study. Derrick Crook and Neil Woodford advised me as to the best “gold-standard” phenotyping method. Sarah Walker assisted me in subsample selection. Manal AbuOun and Muna Anjum trained me to perform agar dilution phenotypes. I performed agar dilution phenotyping and performed analysis comparing it with resistance genetics. With the help of Sarah Walker, I constructed the mixed effect model and created prediction for non-agar dilution isolates using model fixed components.

1.5. Pre-prints and Publications

Work in relation to this thesis.

Davies TJ, Stoesser N, Sheppard AE, Abuoun M, Fowler P, Swann J, Quan TP, Griffiths D, Vaughan A, Morgan M, Phan HT, Jeffery KJ, Andersson M, Ellington MJ, Ekelund O, Mathers AJ, Bonomo RA, Woodford N, Crook DW, Peto TEA, Anjum MF, Walker AS. 2019. Reconciling the potentially irreconcilable? Genotypic and phenotypic amoxicillin-clavulanate resistance in *Escherichia coli*. bioRxiv 511402.

Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, Dunn L, Jeffery K, Butler CC, Hayward G, Andersson M, Morgan M, Oakley S, Mason A, Hopkins S, Wyllie DH, Crook DW, Wilcox MH, Johnson AP, Peto TEA, Walker AS. 2018. Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records. *Lancet Infect Dis* 18:1138–1149.

In relation to other work during the thesis.

Rancoita PM V, Cugnata F, Gibertoni Cruz AL, Borroni E, Hoosdally SJ, Walker TM, Grazian C, Davies TJ, Peto TEA, Crook DW, Fowler PW, Cirillo DM, CRyPTIC Consortium for the CRyPTIC Consortium for the CRyPTIC Consortium for the Cr. 2018. Validating a 14-Drug Microtiter Plate Containing Bedaquiline and Delamanid for Large-Scale Research Susceptibility Testing of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* 62:e00344-18.

van Aartsen JJ, Moore CE, Parry CM, Turner P, Phot N, Mao S, Suy K, Davies T, Giess A, Sheppard AE, Peto TEA, Day NPJ, Crook DW, Walker AS, Stoesser N. 2019. Epidemiology of

paediatric gastrointestinal colonisation by extended spectrum cephalosporin-resistant
Escherichia coli and Klebsiella pneumoniae isolates in north-west Cambodia. BMC Microbiol
19:59.

1.6. Abbreviations

AIC	Akaike information criterion
AMC	Amoxicillin-clavulanate
AMP	Ampicillin
<i>ampC</i>	<i>E. coli</i> 's chromosomal <i>ampC</i> beta-lactamase gene. Also used in line with beta-lactamase literature to describe Ambler Class C cephalosporinases
AMR	Antimicrobial resistance
AMRHAU	Public Health England's Antimicrobial Resistance and Healthcare Associated Infections Reference Unit
APHA	UK Animal and Plant Health Agency
AST	Antimicrobial susceptibility testing
ATM	Aztreonam
ATU	Area of technical uncertainty
AUC	Area under the curve (for Receiver operating characteristic analysis)
BD Phoenix	The Becton-Dickenson Phoenix, an automated susceptibility testing system
Bit-score	A scoring system normalised alignment score obtained as part of standard BLAST output
BL/BLI	Combination beta-lactam/beta-lactamase inhibitor antibiotic
<i>bla</i> _{ACC}	Ambler class C beta-lactamase gene

<i>bla</i> _{ACT}	AmpC type beta-lactamase gene
<i>bla</i> _{CARB}	Carbenicillinase beta-lactamase gene
<i>bla</i> _{CMY}	Cephamycinase beta-lactamase gene
<i>bla</i> _{CTX-M}	Cefotaximase beta-lactamase gene
<i>bla</i> _{DHA}	Beta-lactamase gene discovered at Dhahran, Saudi Arabia
<i>bla</i> _{IMI}	Imipenemase beta-lactamase gene
<i>bla</i> _{KPC}	Klebsiella pneumoniae carbapenemase
<i>bla</i> _{LEN}	LEN-1-like beta-lactamase gene
<i>bla</i> _{NDM}	New Delhi metallo-beta-lactamase gene
<i>bla</i> _{OXA}	Oxacillinase beta-lactamase gene
<i>bla</i> _{SHV}	Sulfhydryl variable beta-lactamase gene
BLAST	Basic local alignment search tool
<i>bla</i> _{TEM}	Temoneira beta-lactamase gene
bp	Base pairs
BSAC	British Society for Antimicrobial Chemotherapy
BSI	Bloodstream infection
CA	Categorical agreement
CAZ	Ceftazidime
CDC	U.S. Centers for Disease Control and Prevention
CI	Confidence interval
CLSI	Clinical Laboratory Standards Institute
contig	A continuous sequence of DNA that has been assembled from overlapping cloned DNA fragments

copy number	the number of times a gene is present in an isolate as estimated by relative DNA coverage versus chromosomal genes
COV	Breadth of gene coverage
CPE	Carbapenemase producing Enterobacteriaceae
CRE	Carbapenem resistant Enterobacteriaceae
CRO	Ceftriaxone
CTX	Cefotaxime
CXM	Cefuroxime
depth	Depth of gene coverage
DNA	Deoxyribonucleic acid
EA	Essential agreement
ECOFF	Epidemiological cut-off breakpoint
ERT	Ertapenem
ESBL	Extended spectrum beta-lactamase
ETEC	Enterotoxigenic <i>Escherichia coli</i>
EUCAST	European Committee on Antimicrobial Susceptibility Testing
ExPEC	Extra-intestinal pathogenic <i>Escherichia coli</i>
FDA	US Food and Drug Administration
FEP	Cefepime
HGT	Horizontal gene transfer
IORD	Infectious Disease in Oxfordshire Research Database

IQR	Inter-quartile range
ISO	International Organization for Standardization
kb	Kilobase
k-mer	A nucleotide sequence of length k
L1/2	Labile beta-lactamase enzyme from <i>Stenotrophomonas maltophilia</i>
MASH	Min-wise independent permutations locality sensitive hashing (minHASH)
mD	Minor discrepancy
MD/ME	Major discrepancy/error
MDR	Multi drug resistant
MEM	Meropenem
MIC	Minimum inhibitory concentration
MLST	Multi-locus Sequence Typing
MMM	The Modernizing Medical Microbiology consortium
n50	A weighted median statistic used to describe assembly quality, such that 50% of the assembly is contained in contigs or scaffolds equal to or larger than this value.
NCBI	US National Centre for Biotechnology Information
<i>ompC</i>	Outer membrane porin C
<i>ompF</i>	Outer membrane porin F
ONT	Oxford Nanopore Technology
OUH	Oxford University Hospitals NHS Foundation Trust

P3, P4, P5, PA/PB, PC/PD	Lartigue <i>bla</i> _{TEM} promoter type
PacBio	Pacific Biosciences
PBP	Penicillin binding protein
PHE	Public Health England
PID	Percentage identity
PIP	Piperacillin
PK/PD	Pharmacokinetic/Pharmacodynamic
PTZ	Piperacillin-Tazobactam
QC	Quality control
QQ plot	Quantile-quantile plot
ROC	Receiver operating characteristic
SDD	Susceptible dose dependent
SNP	Single nucleotide polymorphism
VMD/VME	Very major discrepancy/error
WGS	Whole genome sequencing
WHO	World Health Organization
WCHG	Wellcome Centre for Human Genetics

Chapter 2: Introduction

2.1. Improving diagnostics to combat antimicrobial resistance

The discovery of antibiotics at the start of the 20th century is hailed as one of the most important advancements in the history of medicine. In addition to being one of the major reasons behind a 20-fold reduction in deaths attributable to infectious disease ¹, the treatable nature of infection has enabled many other risky medical interventions such as surgery and cancer chemotherapy to become routine. However, with antimicrobial use also came antimicrobial resistance (AMR) ². Much of this predated the widespread human use of antimicrobials ³, and reflects an on-going evolutionary arms race between organisms and their pathogens ⁴. However, widespread human use is thought to have applied significant selection pressure, resulting in a rapid increase in both frequency and diversity of resistance mechanisms. Further, once acquired, many resistance mechanisms persist even when no longer exposed to antimicrobials ⁵. These effects, coupled with increasing mobility of both people and pathogens through globalization, have resulted in resistance occurring worldwide ⁶. This rapid rate of selection and spread, combined with the relatively slow rate of discovery of new antimicrobials ⁷, threaten a “post antibiotic era”, where even simple infections are resistant to all known antibiotics, and can once again become lethal ⁸.

Consequently, AMR is recognised as one of the leading challenges facing modern healthcare by many of the world’s national and international public health bodies, including the World Health Organisation (WHO), U.S. Centers for Disease Control and Prevention (CDC) and Public Health England (PHE). Consequently, each has developed strategic plans to combat AMR (e.g. the UK 5 year AMR action plan⁹), with a key point of them all being the avoidance

of inappropriate prescribing of current antimicrobials ^{10,11}. While difficult to precisely define, the problem of antimicrobial misuse and overuse is believed to be extensive ^{12,13} and solving it complex. While there are behavioural and societal factors surrounding how and why antibiotics are prescribed ^{14,15}, limitations of current diagnostics have been identified as an important cause of why antibiotics are over-prescribed ^{16,17}. Traditional culture-based susceptibility methods take at least several hours ¹⁸, and in some cases (e.g. *Mycobacterium tuberculosis*) several weeks, to identify pathogens and their antimicrobial susceptibilities. Further, many of these tests require a nearby laboratory, staffed with skilled technicians, to even perform them. For many settings, such as the outpatient and emergency departments, these resources and turnaround times are simply not feasible due to both clinical (e.g. the time critical need for antibiotics in sepsis ¹⁹) and practical (e.g. due to the high patient throughput) reasons. The current alternative is to take specimens for culture and treat with broad-spectrum antibiotics in the interim, and then “de-escalate” therapy once diagnostic results are available. However, while de-escalation is a key strategy in almost all antibiotic stewardship guidelines, in practice it is done far less often than it should be ²⁰, resulting in ever increasing use of broad spectrum antimicrobials ²¹.

Faced with these challenges, there has been intense research on faster and procedurally simpler diagnostic technologies ^{22,23}. While there are several promising technologies being researched, such as multiplex PCR, whole genome sequencing (WGS) is one of the most attractive for several reasons: it can reduce times to diagnosis ²⁴, it a single test for all pathogens and resistance mechanisms, it can identify dangerous resistance mechanisms directly (e.g. carbapenem hydrolysing enzymes) ²⁵, and it can give important epidemiological insights to assist with outbreak control ²⁶. Further, early investigations into WGS-based AMR

prediction have been very promising^{27,28} and in *M. tuberculosis* workflows already use WGS for routine clinical diagnosis and AMR prediction^{29,30}. However, for Enterobacteriaceae, which are some of the most critical resistance threats³¹, WGS-based resistance prediction is still at a proof-of-concept stage³².

This thesis investigates the ability of WGS to predict resistance to beta-lactams (the most commonly used class of antibiotics worldwide³³) in *Escherichia coli* (a member of the Enterobacteriaceae family and one of the commonest causes of bloodstream infection (BSI) worldwide³⁴). It starts by assessing whether methods used in proof-of-principle studies are sufficient to predict beta-lactam antibiograms in unselected clinical isolates. On finding a major issue in resistance prediction for one type of very commonly used beta-lactam antibiotic, beta-lactam/beta-lactamase inhibitor combinations, the underlying reasons for discrepancy are explored. Understanding and accounting for the causes of the discrepancy, WGS-based resistance prediction is ultimately proved possible for this antibiotic. While demonstrating feasibility, this thesis also highlights several microbiological and bioinformatic assumptions which will need to be challenged should WGS-based resistance prediction be adopted into clinical practice.

2.2. Background

2.2.1. *Escherichia coli*: an archetypal Gram-negative organism

First identified by Theodor Escherich in 1885³⁵, *E. coli* is a Gram-negative, facultatively anaerobic bacterium commonly found as part of the normal flora of the lower intestines of mammals and birds. *E. coli* is one of the most well-studied organisms because of its wide

availability and the ease with which it can be genetically manipulated. With regards to its clinical importance, strains are broadly classified into one of three groups according to their interaction with humans ³⁶; commensal strains, enteropathogenic strains which cause diarrhoeal illnesses, and extra-intestinal pathogenic (ExPEC) strains which cause infections in a wide array of sites outside of the gastrointestinal tract. *E. coli* is a major cause of infection, with ExPEC being the commonest cause of urinary tract infection (UTI) ³⁷ and bloodstream infection (BSI) ³⁴ worldwide and enterotoxigenic *E. coli* (ETEC), an enteropathogenic strain, being one of the commonest bacterial causes of diarrhoea in low-income countries ³⁸.

A “typical” *E. coli* genome is 5 megabases long and contains 4900 genes ³⁹, with less than half of these being in the core genome ⁴⁰ (Figure 1). The significant proportion of accessory genome content and large and still growing pan-genome (the set of all genes seen to date in the species) give *E. coli* significant diversity, explaining the broad range of clinical syndromes it can cause. Like all Gram-negative organisms, *E. coli* generally has a large circular chromosome, and then, additionally to this, a number of smaller genetic elements such as plasmids, which are often “mobile” (e.g. conjugative plasmids which can move host by horizontal gene transfer (HGT)) and have highly variable, constantly changing structure ⁴¹. While only making up a small proportion of the DNA, mobile genetic elements disproportionally contain much of the resistance conferring genes of *E. coli*. These features make resistance in Gram-negative organisms highly challenging to both understand and control.

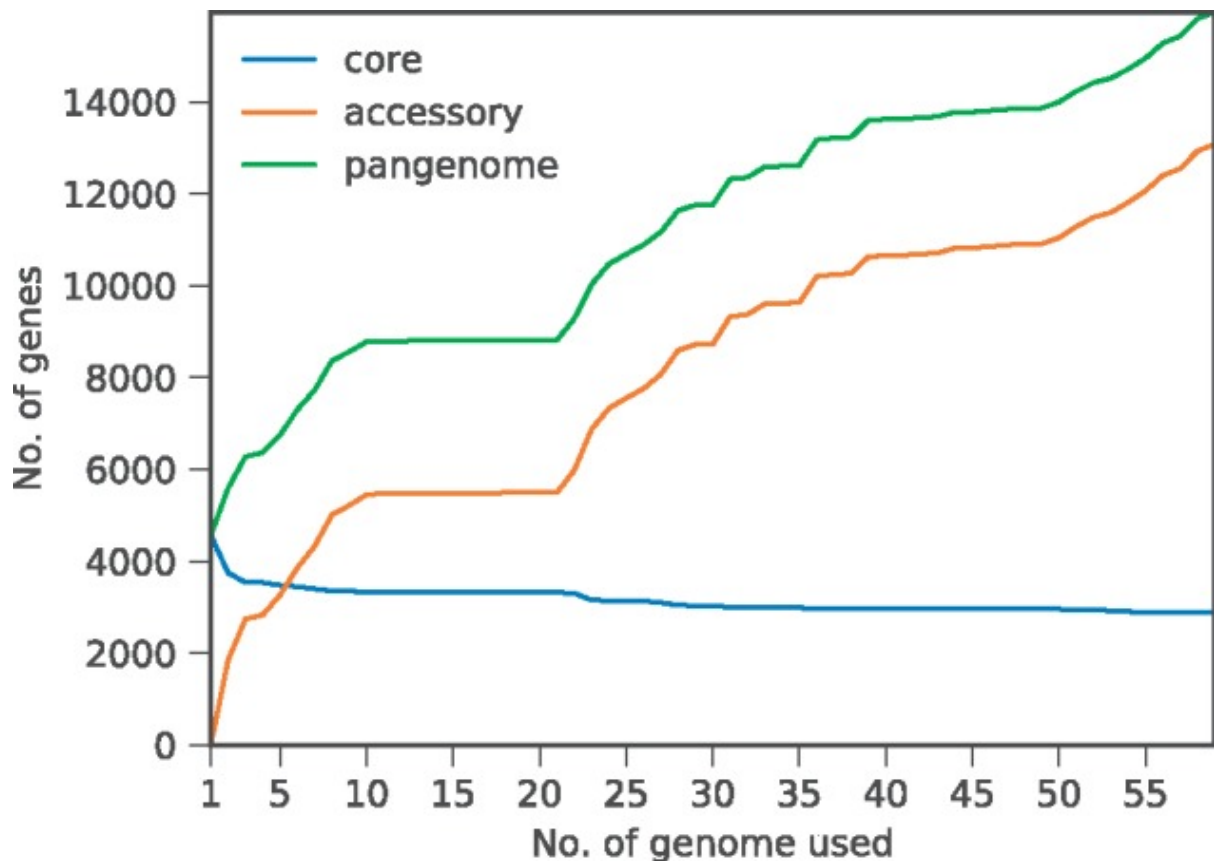


Figure 1: Pan-genome of *E. coli*

Note: Reproduced from Her et al 2018 ⁴⁰.

2.2.2. Beta-lactams

As “descendants” of the first ever naturally derived antibiotic penicillin, beta-lactams are one of the oldest and most well-known antibiotic classes. While penicillin was originally discovered by accident by Alexander Fleming in 1928 ⁴², it was not until 1940 that the potential impact of this discovery was realised, when, based on work by Howard Florey and Ernst Chain, sufficient quantities of it could be purified for use in human trials ^{43,44}.

Following the success of penicillin, intensive search for new antibiotics, both natural and synthetic, during the “golden age” of antibiotic discovery (1950-1970s) yielded many other new major classes of antibiotics, including four other classes of beta-lactams ⁴⁵. However, despite initial optimism that the war on infectious diseases was won ⁴⁶, within 10 years

plasmid-borne resistance mechanisms to all beta-lactam classes had been identified ⁴⁷. In spite of rising resistance, beta-lactams are still the most commonly used antibiotics worldwide because of their broad spectrum activity and low toxicity profile ³³.

Method of action

Beta-lactams target the mechanisms for making bacterial cell walls, on which many bacteria, including *E. coli*, are highly dependent to maintain their structure. A normal bacterial cell wall is a large structure made of cross-linked peptido-glycan chains and thought to be constructed in three phases ⁴⁸ (Figure 2). During the first stage, the N-Acetyl Muramic acid penta-peptide is created in a stepwise fashion in cellular cytoplasm. These penta-peptides are then transported to the cell membrane for linking with the N-Acetyl Glucosamine peptides, creating the peptido-glycan backbone. These peptido-glycan molecules are then finally attached to existing cell wall peptido-glycan chains in a two stage process of molecule cross linking using transpeptidase and then removal of the terminal D-alanine of the opposite strand.

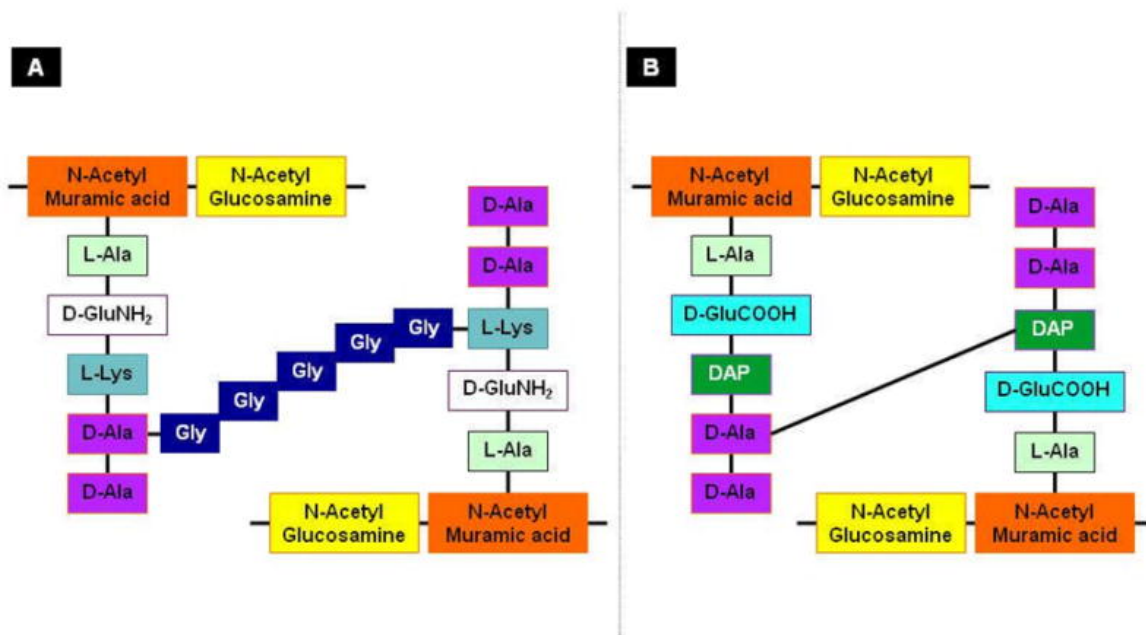


Figure 2: Normal cell wall construction of A) *Staphylococcus aureus*, B) *E. coli*

Note: reproduced from Kong et al, 2010 ⁴⁸.

Beta-lactams, which are structurally similar to D-alanine (Figure 3), are thought to act by inhibiting the proteins (the penicillin binding proteins (PBPs)) which catalyse these last two crosslinking and D-alanine removal steps ^{49,50}. Of note, within any given organism, there are several different, organism specific PBPs, each of which has their own role in peptidoglycan biosynthesis and their own beta-lactam inhibition profiles ⁵¹. For example, the first generation cephalosporin cephalothin has greater affinity for *E. coli* PBP1a than *E. coli* PBP1b, whereas cephaloridine has strong affinity for both, leading to increased cephalothin susceptibility but not cephaloridine susceptibility of PBP1b-deficient *E. coli* mutants ⁵². The chain of events leading to cell death was originally thought to be that PBP inhibition disrupted the balance between cell wall synthesis and hydrolysis, leading to loss of cell wall integrity and subsequent lysis (27). This view has more recently come into question with studies suggesting there may be lysis-independent pathways able to cause cell death ⁵⁴.

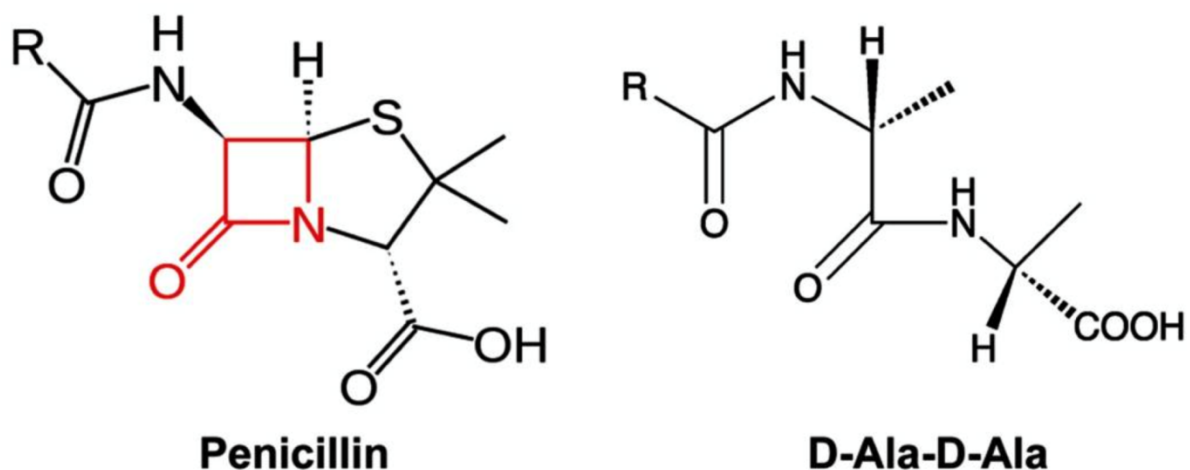


Figure 3: The mimicry of beta-lactam antibiotics to d-alanyl-d-alanine (d-Ala-d-Ala)

Note: The four-member lactam ring in penicillin highlighted in red. Reproduced from Zeng et al, 2013 ⁵⁵.

Classes of beta-lactams

While all beta-lactams contain a variant of the beta-lactam ring involved in the mechanism detailed above, the individual antibiotics can be grouped into five classes (the penicillins, cephalosporins (and cephamycins), carbapenems, monobactams and clavams) depending on their base structure (Figure 4), which can then be iteratively modified to produce new antibiotics (Figure 5). The interactions between cell wall structure, PBPs and beta-lactam resistance mechanisms then define their “spectrum” of activity. Generally, the basic members of each class are narrow spectrum (i.e. only effective against a small number of pathogens and mechanisms), and become progressively more “broad” as the antibiotics are increasingly modified to evade resistance mechanisms - although this is not universally the case ⁵⁶. Of note, for any beta-lactam antibiotic, there are multiple naturally encountered beta-lactamases which can hydrolyse it ⁵⁷.

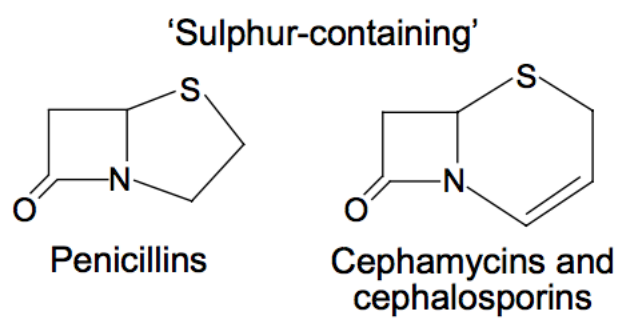
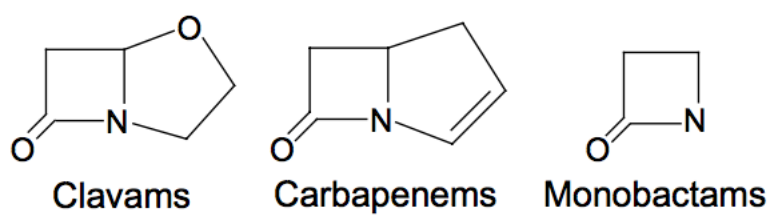


Figure 4: Structure of beta-lactam classes

Note: adapted from McGowan et al ⁵⁸.

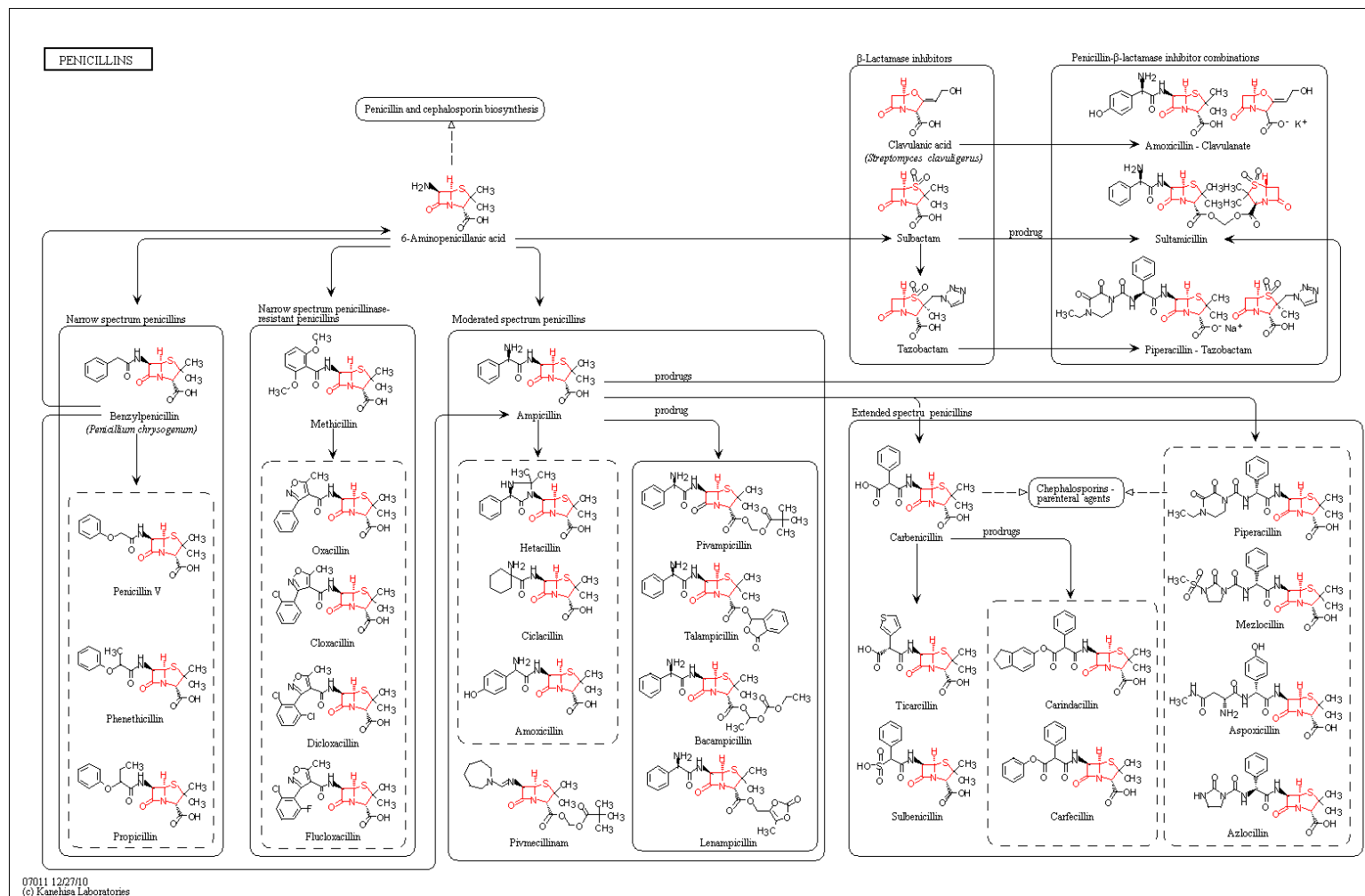


Figure 5: The Kyoto Encyclopedia of Genes and Genomes (KEGG) penicillins reference pathway

Note: available at https://www.genome.jp/kegg-bin/show_pathway?map07011⁵⁹.

2.2.3. Beta-lactam resistance in *E. coli* and other Gram-negatives

Generally, for any antibiotic, there are three methods by which an individual bacteria can have resistance: modification of the antibiotic (e.g. beta-lactamases), restricting access of the antibiotic to its target site (e.g. achieved either due to individual cell mechanisms such as efflux and permeability changes or the factors limiting access to the environment surrounding the cell such as biofilms) and modifying the target site (PBPs) to evade antibiotic action ⁶⁰. In *E. coli*, antibiotic modification by beta-lactamases is the most common mechanism, and, because of the ability of genes to spread via horizontal gene transfer, considered the most clinically important type of beta-lactam resistance ⁵⁷.

Beta-lactamases: background and nomenclature

Beta-lactamases are the name for a large group of well-studied, diverse enzymes that protect bacteria by hydrolysing compounds that contain the beta-lactam ring. Thought to be derived from PBPs to protect bacteria from naturally occurring beta-lactams, many of these enzymes are ancient ^{3,61}, as evidenced by their presence in remote locations with no human contact. Depending on the enzyme, they can either be restricted to the bacterial chromosome or found on transmissible plasmids. Since the discovery of the first beta-lactamase ⁴³ (the chromosomally encoded *E. coli ampC* beta-lactamase), more than 2800 more unique, naturally occurring, beta-lactamase enzymes have been identified, with this number exponentially increasing over time ⁴⁷ (Figure 6).

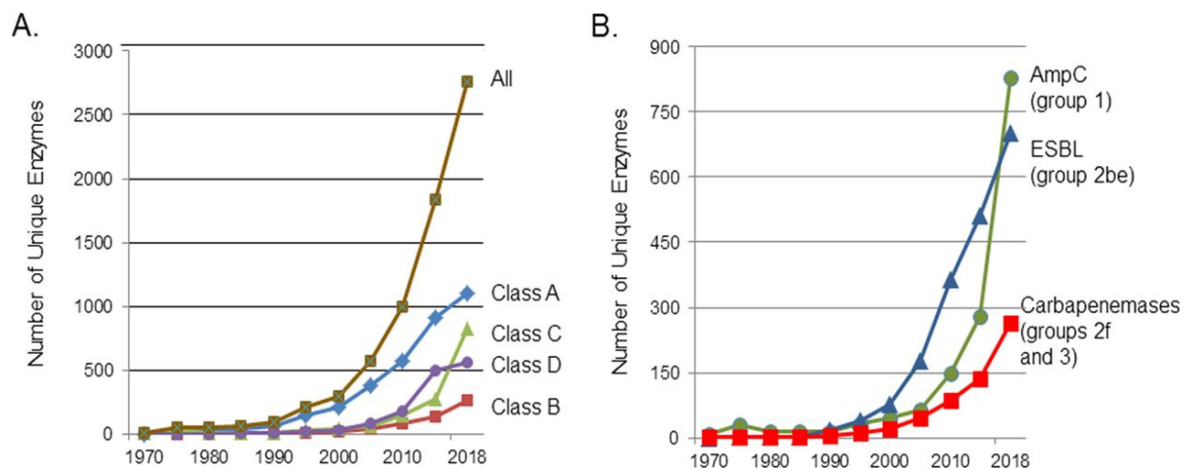


Figure 6: Increase in the number of naturally occurring beta-lactamases by A) molecular class, B) major functional groups

Note: Reproduced from Bush 2018 ⁴⁷.

While naming of these enzymes has been somewhat erratic and prone to error ⁶², their classification schemes are more uniformly accepted, with two being in common use as of July 2019 ⁶³. The first, based on molecular structure, is the Ambler classification system ⁶⁴, which groups these enzymes into one of four structural classes (Figure 7). Class A (e.g. the *bla*_{TEM}, *bla*_{SHV}, *bla*_{CTX-M} families), C (e.g. AmpC and the *bla*_{CMY} family), and D (the *bla*_{OXA} family) enzymes have an active site serine, and class B (e.g. the *bla*_{NDM} family) are metallo-beta-lactamases, which are structurally very distinct from the other classes and possess either one or two zinc ions in their active sites ⁶⁵. The second classification, which is highly correlated with the structural classification, is the functional classification proposed by Bush, Jacoby and Medeiros ⁶⁶ detailed in Figure 7. Essentially, this partitions beta-lactamases into three groups, cephalosporinases (group 1, Ambler class C enzymes), penicillinases (group 2, Ambler classes A and D enzymes) and carbapenemases (group 3, Ambler class B enzymes) depending on the “preferred” substrate of the enzyme, although beta-lactamases are not solely limited to hydrolysing their preferred substrate. For example, several enzymes which hydrolyse third generation cephalosporins belong to class 2 (e.g.

*bla*_{TEM-3}, *bla*_{CTX-M} enzymes) as they efficiently hydrolyse both broad spectrum penicillins and cephalosporins and are often highly structurally related to existing class 2 enzymes.

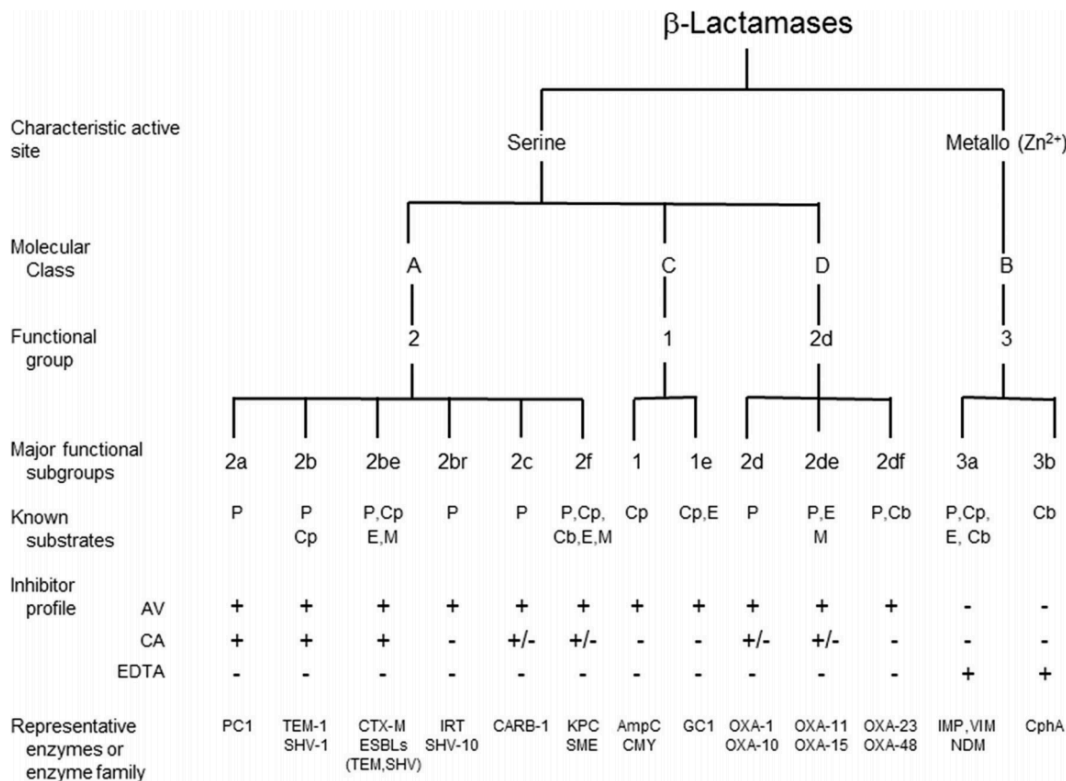


Figure 7: Molecular and functional relationships among beta-lactamases. AV, avibactam; CA, clavulanic acid; Cb, carbapenem; Cp, cephalosporin; E, expanded-spectrum cephalosporin; M, monobactam; P, penicillin. Reproduced from Bush et al 2018 ⁴⁷.

Beta-lactamases: intrinsic and extrinsic

Many Gram-negative species, including *E. coli*, have chromosomally encoded beta-lactamases in addition to horizontally acquired beta-lactamases, which give them innate resistance to some antibiotics ⁶⁷. Important examples include the inducible *ampC* cephalosporinase of *Enterobacter* spp. and *Pseudomonas aeruginosa*, and *Klebsiella pneumoniae*'s *bla*_{SHV-1}. It is important to distinguish these from horizontally acquired mechanisms for two reasons. Firstly, mechanisms capable of horizontal transfer are often considered more threatening due to their ability to pass between organisms and cause

complex “outbreaks” of resistance ^{41,68}. Secondly, these chromosomal mechanisms behave in a different way to their horizontally acquired counterparts, due to different modes of expression. They can be either constitutively expressed at a low (causing little to no resistance) or high level (in a similar fashion to horizontally acquired beta-lactamases), or their expression can be induced by, among other things, the presence of certain antibiotics ^{67,69}. The extent of induction is antibiotic dependent and can lead to complex phenotyping issues, where, for a specific beta-lactam, an isolate is susceptible when tested solely for that beta-lactam, but once exposed to a second inducing antibiotic can become resistant. *E. coli* is a relatively straightforward case, lacking an *ampR* repressor gene and only constitutively expressing its chromosomal *ampC* at a low level, causing little impact on beta-lactam resistance ⁷⁰. Of note however, mutations in the promoter region of this gene are known to drastically increase constitutive expression, leading to production of significant amounts of cephalosporinase ⁷¹.

Beta-lactamases: notable enzymes

Among the beta-lactamases, two groups of enzymes have been classed as key AMR threats by the CDC and the WHO: those which hydrolyse third generation cephalosporins (Extended Spectrum Beta-Lactamases, ESBLs), and those which hydrolyse carbapenems.

Carbapenemase producing Enterobacteriaceae (CPE) and Enterobacteriaceae producing ESBLs have been classed by the CDC as urgent and serious threats respectively ³¹. Since initially found in the 1980s, these mechanisms have spread around the world ^{72,73} (Figure 8-9), and outbreaks are generally refractory to traditional methods of control ^{74,75}. In the UK, recent work suggests between 5-10% of individuals are colonised with ESBL-producing Enterobacteriaceae ⁷⁶, and while the prevalence of CPE carriage is unknown, numbers of

these isolates referred to the PHE reference laboratories have continued to rise year on year, being identified in approximately 3000 isolates in 2017 ⁷⁷. Infections with strains carrying these enzymes are resistant to many first-line empiric antibiotic regimens, are seen as more difficult to treat, and are commonly associated with higher mortality and longer hospital stays ^{74,78,79}.

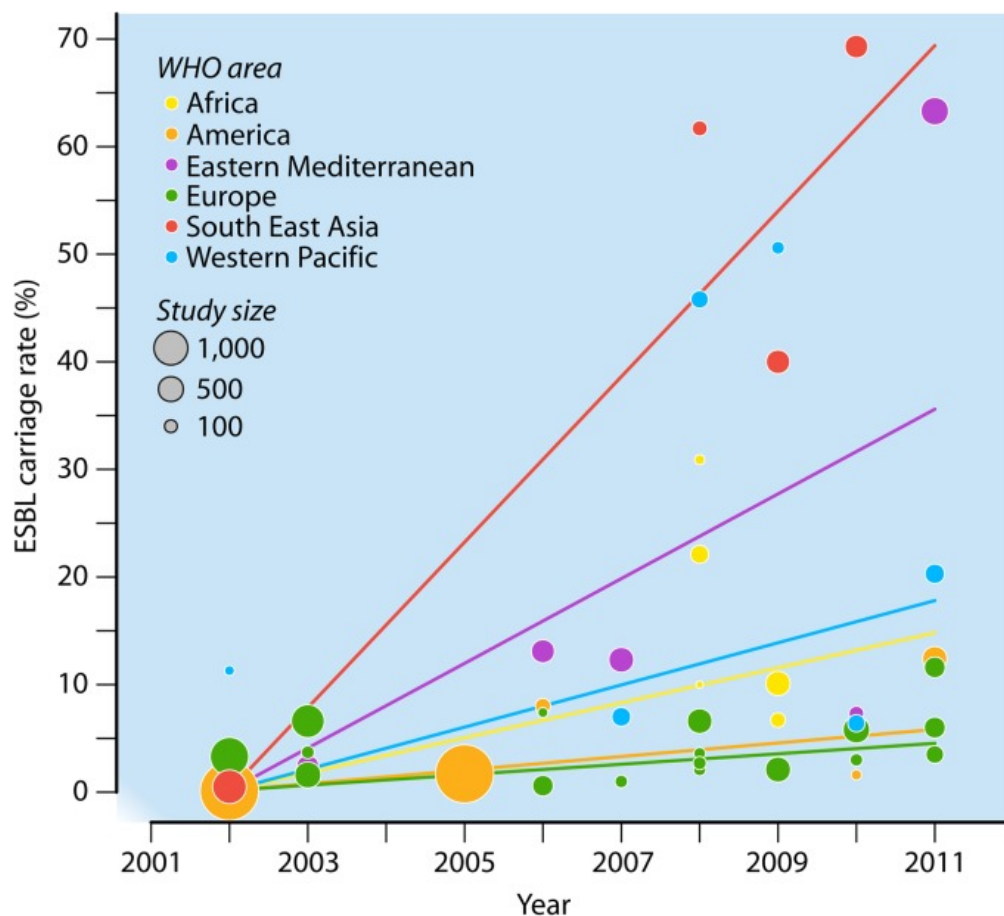


Figure 8: ESBL *E. coli* carriage rates in the community, according to their geographical and temporal distribution

Note: Each bubble area is proportional to the size of the corresponding study. The lines represent the evolution of ESBL *E. coli* carriage rates over time for each geographical area, as established by a weighted linear regression model using the values reported in the literature from 2002 to 2011. Reproduced from Woerther et al 2013 ⁸⁰.

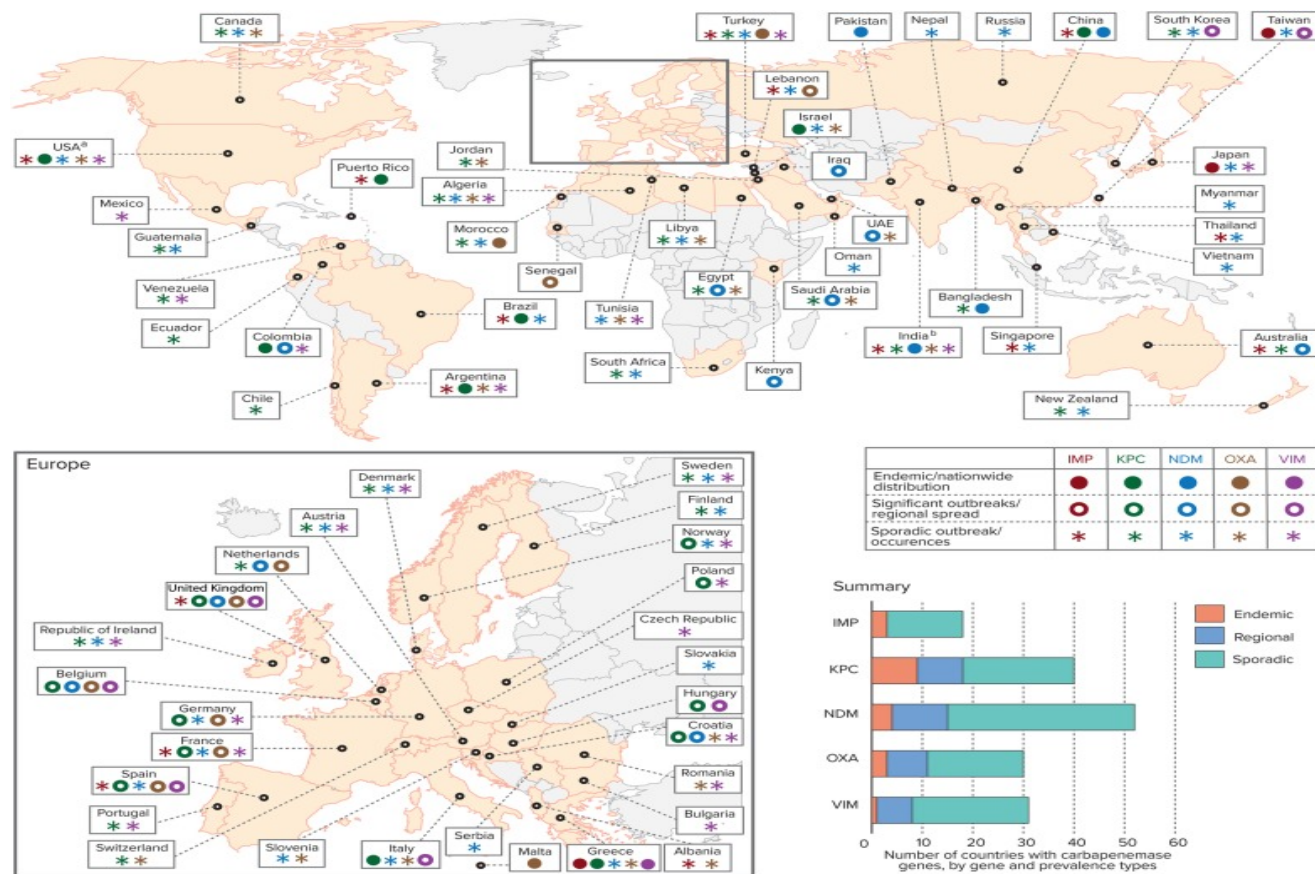


Figure 9: Global distribution of carbapenemases in Enterobacteriaceae, by country and region.

Note: Reproduced from Logan et al 2017 ⁷².

Beta-lactamases: traditional pharmaceutical responses to the beta-lactamase challenge

There are two main strategies pharmaceutical companies have used to counter resistance due to beta-lactamases: modifying antibiotic design so that they are no longer efficiently hydrolysed by beta-lactamase enzymes, and combining active antibiotics with drugs which inhibit the hydrolytic activity of beta-lactamases ⁸¹. Until recently, the first approach has generally been favoured; however, increasing resistance and slow development of antibiotics has left few single antibiotic therapeutic regimens available for some infections ⁸². For example, *Stenotrophomonas maltophilia*, an opportunistic pathogen which is multi-drug resistant due to multiple intrinsic resistance mechanisms, has two intrinsic inducible beta-lactamases, a metallo-beta-lactamase (L1) and a cephalosporinase (L2) ⁸³. No single beta-lactam antibiotic is able to circumvent both mechanisms, and, given its other resistance mechanisms, can leave no single therapeutic agent options at all ⁸⁴. While this is an extreme case, even single beta-lactamases such as *bla*_{KPC} can cause problems, with singular treatments only recently becoming available and still not in widespread use ⁸⁵. This has led to renewed interest in the second approach, concomitantly administering inhibitor drugs with beta-lactam antibiotics ⁸⁶, with many of the new agents in development for treating multi-drug resistant (MDR) Gram-negative organisms being of this type ⁸². Simplistically, this approach is thought of as the inhibitor preventing destruction of the beta-lactam, restoring its natural activity ⁸⁷. In reality effects are slightly less simple, with the distinction between antibiotic and inhibitor not always being clear. For example, traditional inhibitors are clavams (beta-lactams) and have some antibacterial activity of their own ⁸⁸, and conversely carbapenems can act as beta-lactamase inhibitors ⁶⁵.

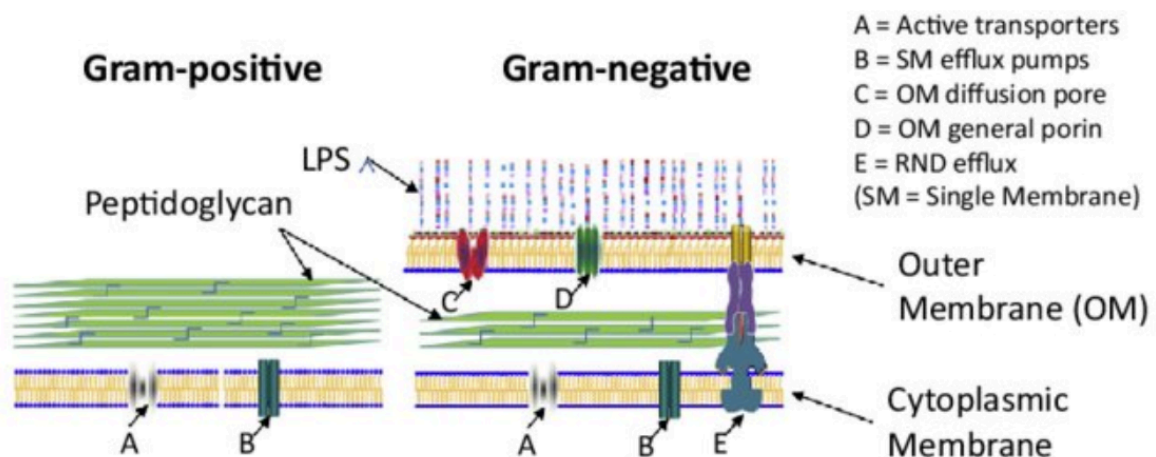


Figure 10: Differences between Gram-positive and Gram-negative permeability barriers

Note: Gram-positive and Gram-negative envelopes: A = active transporter; B = single membrane efflux pumps; C = outer membrane facilitated diffusion pore; D = outer membrane general porin; E = Resistance-Nodulation-Cell division(RND) efflux pump spanning both outer and cytoplasmic membranes. All five A-E (transporters, efflux pumps, diffusion pore and porin) are present in Gram-negative bacterial strains whereas only active transporter (A) and single membrane efflux pump (B) are present in Gram-positive bacterial strains. Reproduced from Singh et al. 2017 ⁸⁹.

For beta-lactams to operate, they need to reach their target site at the bacterial cell wall in sufficient quantities. In Gram-positive organisms, this is relatively simple, as the wall is directly inside the bacterial capsule and is only protected by the extracellular beta-lactamases that they create ⁹⁰. In Gram-negatives, the wall is contained in the periplasmic space between the cell outer and inner membranes (Figure 10). This produces a significant extra challenge for two reasons. Firstly, entry and exit of this space is tightly controlled by an array of efflux pumps (which move substances across the cell membrane by active transport) and porins (pores which allow the passive movement of substances across the cell membrane), which can prevent meaningful access to antibiotic targets. Secondly, Gram-negatives do not extrude their beta-lactamases into the extracellular space, but instead concentrate them in the periplasmic space (between the inner and outer membranes). This

feature can add significant complexity by adding a quantitative aspect to the interaction between antibiotic molecules, PBPs and beta-lactamases ⁹⁰. When combined with innate beta-lactamases, poor permeability and efficient antibiotic efflux gives rise to some of the most challenging multi-resistant organisms encountered ⁹¹.

While the theory of how alterations in permeability and efflux can cause resistance is fairly straightforward, there is only very limited understanding of the exact impact of specific permeability or efflux changes on permeability to individual antibiotics because these effects are challenging to study ⁹². Cell permeability and efflux are not constant, but instead dynamic systems which respond to external stimuli and are regulated by complex species-specific mechanisms. Further, antibiotics often may enter the periplasmic space through multiple routes (including simple diffusion across the outer membrane) and hence disruption to any individual route is often compensated for by alternatives, resulting in only modest changes in resistance. Given these complexities, and the relatively coarse nature of traditional antimicrobial susceptibility testing, only drastic changes, such as complete porin loss of function (e.g. caused by events such as the insertion of sequence into a gene leading to mis-sense mutations) can be reasonably detected as having an effect ⁹³.

In *E. coli*, permeability is mainly regulated by two “non-selective” porins, OmpF and OmpC, which are co-regulated so that reduced expression of one will lead to increased expression of the other and vice versa ⁹². Efflux is primarily mediated by the AcrAB-TolC efflux system which is capable of extruding a wide range of substrates ⁹⁴. Beta-lactams present at physiologically achievable doses are normally able to overcome these mechanisms in ‘wild-type’ *E. coli*. However, in isolates where permeability is compromised (e.g. by mutations

causing porin gene loss of function) resistance to beta-lactams is accentuated⁹⁵. This effect can even give false ESBL-like and CPE-like antibiograms^{25,96}.

Target site modification

Like permeability and efflux, PBPs are species-specific. While PBP modification plays a major role in beta-lactam resistance in Gram-positive organisms (such as the PBP2a produced by the horizontally acquired *mecA* gene in *Staphylococcus aureus*), its role in Gram-negative resistance is less prominent. Altered PBPs have been implicated in clinical *Acinetobacter baumannii* and *P. aeruginosa* isolates⁹⁷, and induced in laboratory *E. coli* strains⁹⁸, however, altered PBPs in clinical *E. coli* isolates are exceptionally rare, with none being known until relatively recently^{99,100}.

2.2.4. Antimicrobial susceptibility testing

In simple terms, antimicrobial susceptibility testing (AST) is a standardized laboratory procedure used to measure the susceptibility of an organism to a given antimicrobial. This definition however hides significant complexity due to ‘susceptibility’ and ‘resistance’ themselves being difficult to define¹⁰¹. Generally it refers to a two stage process by which firstly the minimal concentration of antibiotic (MIC) required to prevent microbial growth *in vitro* is quantified and then secondly this concentration is compared to a set of breakpoints which determine an isolate’s susceptibility (it’s “phenotype”).

Susceptibility testing methods

AST methods originated at the same time as the early development of antibiotics, being used as a visualisation technique for the biological phenomenon. Indeed, Fleming is often credited (although incorrectly ¹⁰²) as being the inventor of the “agar diffusion” technique where he observed the extent of inhibition caused by penicillin on *S. aureus* colonies on agar media ¹⁰³. Since these early techniques there have been multiple advances improving quality control, standardization, speed and automation; however the basic techniques remain the same ^{18,22}. These come in two forms, dilution and diffusion based tests. Dilution based tests measure the minimum concentration of antibiotic required inhibit bacterial growth (the MIC) by attempting to grow the bacteria in ever increasing concentrations of the antibiotic. Diffusion based tests instead aim to measure the area of bacterial growth inhibition around an antibiotic source ¹⁰⁴. While tests in clinical use today are deemed acceptably close to equivalent ^{105,106}, there are subtle differences with each test having their own error profile ^{107–109}. Originally only agar dilution, and now additionally non-automated broth dilution/microdilution, was seen as “gold standard” phenotypes, but these are not in widespread clinical use due to their labour and resource-intensive nature. In practice which test is used is dependent on the needs of the testing laboratory ¹¹⁰. For example, clinical laboratories often require automated susceptibility testing systems such as the BD Phoenix (Beckton, Dickenson and Company, USA) or Vitek 2 (BioMérieux, France) machines capable of processing large numbers of samples at any time of day, whereas many research laboratories may favour less resource intensive techniques which are simpler to administer such as disk diffusion and E-test gradient strips (BioMérieux, France).

Breakpoints; categorising and interpreting AST results

Like AST methods, there are multiple ways in which isolates can be categorised according to their AST results, depending on what the interpretation aims to describe ¹¹¹. Two common methods of categorising isolates use “wild-type” and “clinical” breakpoints. The “wild-type” breakpoint (analogous to the European Committee for Antimicrobial Susceptibility Testing’s (EUCAST) epidemiological cut-off value (ECOFF)) is used to describe the maximum MIC seen to a given antimicrobial among the wild-type population of the organism (containing no acquired/mutational resistance features). This breakpoint is only based on AST data and genetics of the tested microbial species and implies that isolates with MICs above this breakpoint have some additional resistance feature compared to wild-type population of the same species. While this breakpoint is a relatively simple concept, defining which genetic motifs are sufficient to cause an isolate to be non-wild-type can itself be highly complex (e.g. mutations leading to minor permeability differences which only cause small increases in resistance are not easily categorised), meaning ECOFFs are not defined for many antibiotics. In contrast, the “clinical” breakpoint is a breakpoint which clinical laboratories use to give meaningful information to health professionals on how to treat their infected patients. In this definition, an isolate being resistant to an antimicrobial is linked to a “higher chance of treatment failure” for the patient infected with this isolate if treated with said antimicrobial. This definition, in addition to accounting for the MIC/resistance mechanisms present in the isolate, is commonly based on antimicrobial pharmacokinetic/pharmacodynamics (PK/PD) data and clinical outcome data. In general, breakpoints and their meaning can also differ depending on many other factors such as testing method, characteristics of local isolates, and, in clinical isolates, body site of infection ¹¹². In practice, most breakpoints used in clinical practice worldwide are set by

either EUCAST (Europe) or Clinical Laboratory Standards Institute (CLSI). While breakpoints from these two organisations are often very similar, there are occasional and subtle differences, both in exact meaning of susceptibility categories and the antibiotic concentrations to define them (Table 1-2). For example, while resistant and susceptible have similar meanings for the likelihood of therapeutic success for both sets of breakpoints, CLSI infers this from by comparing isolate MIC relative to usual drug concentrations achieved at the site of infection, whereas EUCAST additionally uses clinical outcome data to define breakpoints resulting in a slightly different definition. Further, for non-resistant non-susceptible isolates, the definitions between the two differ even further. Both have concepts representing isolates which are likely susceptible but require increased concentrations of antibiotics to treat them (EUCAST – susceptible, increased exposure, CLSI – susceptible, dose dependent) but EUCAST has no equivalent of CLSI’s Intermediate category. This category accounts for isolates where therapeutic response is uncertain, either because of innate isolate/infection characteristics or due to test uncertainty. EUCAST by contrast only have “areas of technical uncertainty” which only represents isolates with uncertain phenotypes due to technical difficulty measuring them, having no category of isolates that have uncertain interpretations for non-technical reasons. The nuances of differences between breakpoints are often poorly recognised, with isolate susceptibility being somewhat carelessly classified from a single test without specifying which method and breakpoints were used.

Evaluating novel AST methods

Given the considerable variation in breakpoints and AST methods, both the US Food and Drug Administration (FDA) and the International Organization for Standardization (ISO) have

developed standards to assess the accuracy and reproducibility of the results produced by any particular AST method ^{106,113}. While there are some minor differences between these guidelines, these generally revolve around two concepts; categorical agreement, the percentage of isolates where susceptibility interpretation agrees with the reference method, and essential agreement, the proportion of isolates where measured MIC is within one doubling dilution of the MIC measured by the reference method (Table 3). The definitions of “acceptable” by both are fairly similar. However, while these targets need to be met for each organism-antibiotic combination, there is no such requirement for testing the performance of AST methods for specific resistance mechanisms. Consequently, while AST methods may work well on the samples tested at initial accreditation, due to highly variable local patterns of resistance and the rapid rise of novel mechanisms, there is no guarantee they will perform as well on samples in the clinical setting ^{107,114}.

Table 1: Summary of CLSI category definitions

Category	Classification letter	Interpretation
Susceptible	S	Isolates with MICs at or below this threshold are inhibited by the usually achievable concentrations of antimicrobial agent when the dosage recommended to treat the site of infection is used, resulting in likely clinical efficacy of the antimicrobial agent.
Susceptible-dose dependent	SDD	Isolate susceptibility depends on the dosing regimen that is used in the patient. To achieve levels that are likely to be clinically effective against isolates with AST results in the SDD category it is necessary to use dosing regimens that result in higher drug exposure than that achieved with the dose that was used to establish the susceptible breakpoint. For this category, these dosing regimens need to be supported by the literature/widely used/approved, else the isolate is classified as intermediate.
Intermediate	I	Isolates with MICs that approach usually attainable blood/tissue antibiotic levels where clinical response rates may be lower than for susceptible isolates. This category includes isolates where increased antibiotic exposure may treat the infection, but there are no dosing regimens satisfying the SDD definition above. This category also includes a technical buffer zone to prevent technical difficulties from causing major discrepancies in interpretation.
Resistant	R	Isolates with MICs above the resistant breakpoint are not inhibited by the usually achieved antibiotic concentrations in blood/tissue and/or MICs fall in a range where specific mechanisms are likely and clinical efficacy of the agent against the isolate has not been reliably shown in treatment studies
Non-susceptible	NS	A category for isolates with MICs above the susceptibility breakpoint in antibiotics where only a susceptible breakpoint is defined, due to limitations of evidence (e.g. where resistance is very rare, and although there is evidence of treatment success is available in isolates with low MICs, there is little evidence of treatment failure in isolates with higher MICs).

Note: Adapted from ¹¹⁵.

Table 2: Summary of EUCAST category definitions

Category	Classification letter	Interpretation
Susceptible, standard dosing regimen	S	Isolates where there is a high likelihood of therapeutic success when using a standard dosing regimen of the antimicrobial agent.
Susceptible, increased exposure	I	Isolates where there is a high likelihood of therapeutic success when using an antimicrobial agent but only when the isolate is exposed to concentrations of the agent greater than those achieved using a standard dosing regimen. Exposure to the agent can be increased by adjusting the dosing regimen or by its concentration at the site of infection.
Area of Technical Uncertainty	ATU	Isolates where there is uncertain therapeutic success when using an antimicrobial agent. Uncertainty is caused through technical difficulties in performing AST. This zone acts as a buffer to prevent these difficulties from causing major discrepancies in susceptibility categorisation
Resistant	R	Isolates where there is a high likelihood of therapeutic failure even when there is increased exposure to the antimicrobial agent

Note: Adapted from ¹¹⁶

Table 3: Comparison of ISO and FDA standards for evaluation of the performance of antimicrobial susceptibility testing devices

Metric	Concept	Equation	FDA limit	ISO limit
Accuracy				
Essential agreement (EA)	Tested device MIC being within +/- 1 doubling dilution of reference device MIC	$\frac{N_{EA} \times 100}{N}$	> 89.9% ⁺	> 90%
Categorical agreement (CA)	Tested device susceptibility interpretation agreeing with reference susceptibility interpretation	$\frac{N_{CA} \times 100}{N}$	> 89.9%*	> 90%
Very major discrepancy/Error (VMD/VME)	Tested device S: Reference device R	$\frac{N_{VMD} \times 100}{N_{RREF}}$	UCI ≤7.5% [†] LCI ≤1.5%	≤3%*
Major discrepancy/Error (MD/ME)	Tested device R: Reference device S	$\frac{N_{MD} \times 100}{N_{SREF}}$	≤3%	≤3%
Minor discrepancy (mD)	Tested device R/S: Reference device I Tested device I: Reference device R/S	$\frac{N_{mD} \times 100}{N}$	-	-
Quality control (QC)	MIC within range for QC organisms (EUCAST/CLSI references) for a given proportion of tests	NA	≥95%	≥95%
Reproducibility	Repeat MIC within +/- 1 doubling dilution of model for a given proportion of tests	NA	≥95%	≥95%

+ : For the FDA guidelines, the target for EA is to be 95% confident that the true EA is >89.9% (i.e. the lower 95% confidence interval bound for the EA is > 89.9%)

*: For both of these cut-offs, the FDA/ISO set out situations in which worse results than these would be acceptable. The FDA will tolerate a lower CA in the context of a high EA and most errors being minor discrepancies. The ISO tolerate a potentially higher VMD rate again if EA is high in the context of a rare resistance phenotype

†: For the FDA guidelines, the target for VME is to be 95% confident that the true VME is ≤7.5%, and for 1.5% to be greater than the 95% lower confidence interval bound.

N_{RREF} and N_{SREF} are the number of isolates resistant and susceptible by the reference device respectively.

2.3. WGS-based prediction of AST phenotypes

WGS has multiple advantages over traditional culture-based laboratory tests for diagnosing infections and predicting their resistance. It is a relatively straightforward single assay which could theoretically avoid the need for culture¹¹⁷, is capable of diagnosing wide array of infections³² and can provide extensive additional epidemiological information about infections¹¹⁸. Further, WGS can definitively identify many of the “most dangerous” resistance mechanisms which are capable of horizontal transfer, whereas traditional culture can only suggest their presence²⁵. However, whether WGS can replace traditional methods is still unclear^{32,119}. In traditional culture-based AST, the action of an antibiotic on an isolate is directly observed *in vitro*, whereas WGS only determines an isolate’s DNA. Hence, for WGS to replace traditional AST, isolate resistance must be highly predictable from DNA features alone.

2.3.1. Approaches to predicting susceptibility using WGS

Over the last 10 years several different methods have been applied to predict isolate AST results from WGS data. They can be grouped into one of two general approaches depending on the level of *a priori* knowledge required. The first, conceptually the simplest of approaches, searches WGS data for known resistances features, which are then used together with a set of manually defined “rules” to predict resistance^{28,120–122}. In 2015, this “rule-based” approach was the only method in use and it is still one of the most commonly used approaches as of July 2019¹¹⁹. This method however requires both a complete database of resistance features, and that those features can be reliably linked with their expected phenotypes. Other approaches use statistical methods to drop either one^{40,123,124}

or both of these assumptions ^{125–127}. In this introduction, approaches are therefore categorised into “rule-based”, “database-guided” and “agnostic” approaches (Figure 11), with explanations for each given below.

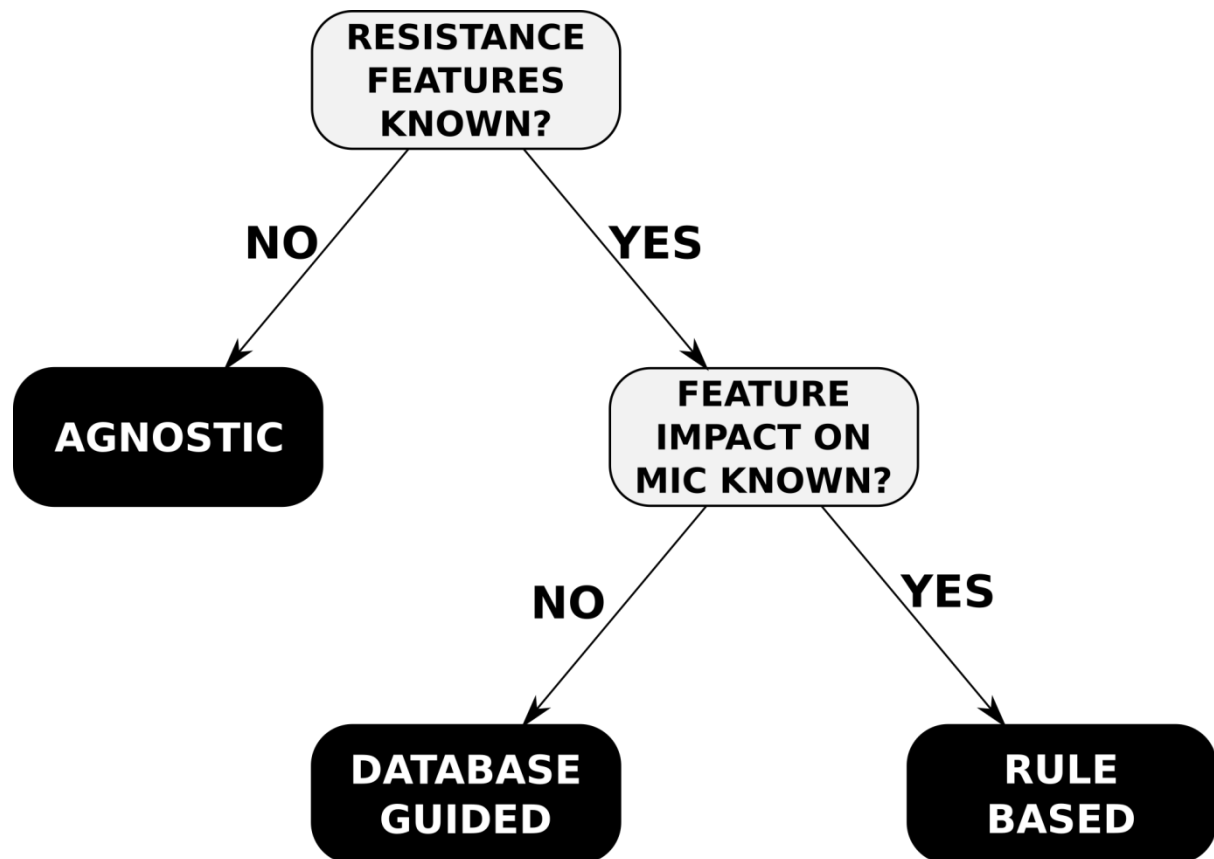


Figure 11: Categorisation of WGS-based resistance prediction approaches

As only the rule-based approach was widely used at the outset of this thesis, an overview of this method is presented here. Work in this thesis then extends from this point, including using a more “database-guided” approach. Non-rule-based approaches are also presented in this introduction to give context, and their benefits/drawbacks are further considered in the discussion.

2.3.2. Rule-based resistance prediction

The basic requirements for using rule-based approach are 1) a database of resistance features, and 2) a method to search WGS data. With no clear “gold standard”, there are several options for each component. Since there are several recent in depth reviews of the methods available ^{128–130}, only selected examples will be explained here to highlight issues within the field. The github repository (https://github.com/arpcard/amr_curation) has also been recently started as a central resource of up to date information on the tools and databases available.

Resistance databases

The underlying resistance database is perhaps the most fundamental component of rule-based resistance prediction. In its most basic form, it is simply a collection of sequences/mutations that have been previously found to be associated with resistance. Thereby, they not only define what elements are found, but also to some extent how they should be interpreted. Maintaining a resistance database, however, is far from straightforward, because of rapidly increasing knowledge of different resistance mechanisms, convoluted classification schemes and complexities of curating ever changing genetic data ¹³¹. In addition to these problems, knowledge of how genotype and phenotype are associated is often incomplete as mechanisms may reside in different host species, have AST tests against only a small numbers of antibiotics and only be tested a small number of times (both in terms of number of isolates and number of repeat tests of the same isolate). Consequently, databases often contain poor quality data and some are no longer maintained at all ¹²⁸ (e.g. the Lahey clinic beta-lactamase database

(<https://www.lahey.org/Studies/>)¹³² ceased assigning allele names/numbers to novel beta-lactamases as of 1st July 2015). Currently there are several databases in widespread use among the genomic community, with the three most popular non-species-specific databases being the Comprehensive Antimicrobial Resistance Database (CARD) database¹³³, the Resfinder/Pointfinder databases²⁸, and the National Centre of Biotechnology Information's (NCBI) Bacterial Antimicrobial Resistance Reference Gene Database (BARRGD)¹³⁴. Each of these databases is generally thought to be reliable aside from minor curation issues, but their different structures and scope prevent easy comparison of their contents.

Sequence-search methods

As of July 2019, several bioinformatic tools have been written to identify AMR features in any of the databases above from short-read sequencing data. The two oldest, simplest and most widely used of these are using BLAST¹³⁵ on sequence assemblies²⁸, and using a mapper¹³⁶ to directly map reads to the resistance database¹³⁷. Each of these methods has perceived strengths and weaknesses (Table 4).

Table 4 Advantages/disadvantages of assembly and mapping based approaches

Method	Advantages	Disadvantages
BLAST premade assemblies (first published 2012 ²⁸)	• Sequencing technology agnostic • Always gives consensus sequence • Allows inspection of surrounding sequence • Enables better assessment of indels	• Assembly errors can result in missing elements • Heterozygotes may be collapsed to give consensus sequence only, but may be important for resistance • Assembly can add significantly to processing time
Mapping (first published 2014 ¹³⁷)	• Better identification of resistance elements • Enables visualization of heterozygotes • Works from raw reads	• Unable to see surrounding elements • Difficult to confirm indel vs gap • Sequencing technology specific • Unable to map to more diverse sequences

Several methods have been created to address these issues, including combination methods¹³⁸, using debrujin graphs²⁷, and using k-mer (nucleotide sub-sequences of length k) sharing¹³⁹. However, due to only limited evidence being available¹⁴⁰, how these methods perform in relation to one another is unknown.

Resistance prediction algorithms

The actual algorithm by which resistance is predicted from genotype data is perhaps the most poorly understood and variable component of the rule-based approach. Approaches thus far have assumed that for each member of the resistance database its presence has a causal effect on AST results which is quantifiable, predictable and generally independent of sequence context^{121,141}. For example, presence of a *bla*_{CTX-M-15} gene in an isolate (which produces an ESBL beta-lactamase) is inferred to lead to resistance to all third generation cephalosporins. Consequently, this algorithm in essence is a look up table, where presence/absence of each element dictates resistance (Figure 12).

```
##### PSEUDOCODE FOR PREDICTION ALGORITHM #####  
  
for each antibiotic:  
    # Initialisation step  
    predicted phenotype := S  
    for each resistance sequence identified in isolate:  
        # update step  
        look up resistance sequence in database  
        if resistance sequence associated with resistance to antibiotic:  
            predicted phenotype := R
```

Figure 12: Rule-based resistance prediction

By its nature, this approach works well for effects associated with either the presence/absence of genes and/or mutations in genes. This is due to the typical certainty with which these feature reference sequences can be identified from sequence data and the relative ease with which they can be catalogued into databases. The limited understanding and apparent variability of some features such as different promoters makes them more difficult to identify and catalogue, proving a challenge for this type of approach.

Strengths and limitations of the rule-based approach

Rule-based approaches rely on the user finding and specifying which genetic features are important for AST prediction (e.g. specifying a rule that states that an isolate would be predicted ertapenem resistant if the *bla*_{OXA-48} gene was identified in its sequencing data). The major strength of this approach is its interpretability. AST predictions are explainable using widely accepted biology, giving confidence in the prediction. Further, given predictions are based on the literature of in-depth biological investigation of mechanisms from all over the globe, many predictions are likely highly generalisable. While in depth knowledge of mechanisms can be a strength, it also causes a major weakness with this approach, as predictions can only be as good as the knowledge on which they are founded. Many genetic features of resistance can be difficult to identify and catalogue. Further, impacts of some features on AST results, particularly new antibiotics, can be unknown.

2.3.3. Non-rule-based resistance prediction

The major difference between rule-based and non-rule-based approaches is that rule-based methods require user made rules of how to determine resistance. Non-rule based approaches circumvent this using statistical models. For the purposes of this thesis, they can

be classified into two categories, those which require identification of resistance sequences (database-guided) and those which do not (agnostic) (Figure 11). Like rule-based approaches, database-guided approaches start with an initial sequence search step. The resistance genotype is then used as input features for models ^{123,142} to infer the phenotype. In contrast, purely agnostic approaches use no databases and instead use the sequence data directly (thus far most often using kmers obtained from genome assemblies) as input ¹²⁵.

Database-guided approaches reduce the number of model features drastically, simplifying computation and interpretation, and enabling modelling of smaller datasets. The trade-off however, is that, like rule-based approaches, resistance can only be attributable to sequences in the resistance database, eliminating the possibility that resistance may be due to features not in the database. By their nature, agnostic approaches allow any genetic sequence to potentially be a feature for resistance prediction. However, these models typically have many millions of input features, require large amounts of data which itself requires significant computational resource and technical expertise to analyse, and are prone to overfitting ¹⁴³.

Strengths and limitations of the non-rule-based approaches

The non-rule-based approaches address many of the weakness of the traditional rule-based approaches. Firstly, they reduce (or eliminate) the reliance on previous literature and the need for resistance mechanisms/phenotypes to be catalogued in databases. Further, as they are simply learned from “real” data as opposed to from *in vitro* experiments with isolated mechanisms, they may be better at predicting impacts of resistance features in real-world contexts. Finally, as they do not rely on the deterministic nature of rule-based prediction,

these methods can be employed without a specialist knowledge of the mechanisms involved. This feature may greatly help their scalability to enable widespread clinical use. The major drawbacks of these approaches are their interpretability and generalisability ¹⁴³. As prediction is paramount for these methods, selected features may not be causally associated with resistance, but instead associated via confounding effects. Given data used for model training is likely biased both by where isolates come from and why they were selected for sequencing, strength of association of these confounding effect features with phenotype may not be reproducibly seen in other datasets; that is, whilst these effects may improve prediction accuracy in the original dataset, these effects may not be generalisable. The other major weakness of these approaches is the large training set size and computational resource required to make good predictive models. However, with the rapidly increasing quantity and quality of linked genetic and phenotype data, this obstacle may prove less of a barrier in the future.

2.3.4. Specific issues surrounding beta-lactam resistance prediction in *E. coli*

Added complexity of beta-lactam resistance and the presence/absence fallacy

Compared to the other antibiotic classes, beta-lactams are arguably one of the more complex. Often other classes only contain between 1-5 antibiotics in widespread clinical use, compared to over 20 beta-lactams ³³. While part of this complexity reflects their unique clinical utility, being both broad-spectrum and having low toxicity, the major reason is that it reflects an ongoing evolutionary and pharmaceutical arms race of beta-lactams versus their increasingly complex resistance mechanisms. Clinically, specific beta-lactams are chosen to circumvent specific resistance mechanisms. For example, *Enterobacter spp* can produce an inducible chromosomal AmpC. Cephalosporins which induce enzyme production are quickly

hydrolysed, but ones that do not can be considered in treatment ¹⁴⁴. Further, small genetic differences (e.g. a lone single nucleotide polymorphism, SNP) can have a drastic change on the beta-lactam resistance profile ¹⁴⁵. To meet clinical need, WGS-based resistance prediction would need to accurately predict susceptibility for multiple closely related drugs. This task is likely more problematic compared to predicting resistance for other antibiotic classes where often only one or two different agents are used in most clinical settings.

Further, the complexity of predicting beta-lactam resistance is often underestimated due to it primarily being due to the presence of specific beta-lactamases. Deceptively, resistance due to the presence/absence of specific genes is thought easier to identify from WGS data than resistance due to SNPs, because the coarser resolution of gene presence/absence (as opposed to SNPs) theoretically means sequencing, mapping and assembly errors are more easily tolerated. However, given two beta-lactamases can be only a single SNP different but result in different antibiograms, predicting beta-lactamase resistance requires SNP level accuracy. Further, in the case of imperfectly matching alleles such as synonymous DNA variants, identifying presence of specific beta-lactamase alleles is heavily reliant on sequence alignment to choose the closest matching reference from a large set of possibilities. While new nomenclatures such as naming alleles by genetic identity (with a 2% difference cut-off) have been proposed to alleviate this problem for geneticists, they fail to solve the resistance prediction issue, and have thus far been swiftly rejected by the world-wide beta-lactamase research community ¹⁴⁶. With the exponential increase in beta-lactamases, this number of closely related alleles is ever increasing, which will inevitably make this problem more difficult over time ⁴⁷.

Plasmid-borne nature of resistance

Beta-lactam resistance in *E. coli* is predominantly acquired via horizontal transfer of beta-lactamase genes on plasmids, causing several technical challenges for analysing sequencing data. *De-novo* assembly of plasmids from short-read sequencing data has proved complex^{41,147}, due to the fact that isolates can contain multiple plasmids and the often repetitive nature plasmid sequence. Secondly, isolates are not restricted to containing single genes from any given family¹⁴⁸, which again can exacerbate assembly issues and make consensus mapping approaches complex when attempting to identify what alleles were initially present in the sample. Hence inferences about what genes are in the underlying sequence made by consensus approaches or using data on the relevant abundance of variants are less reliable. In addition to these issues, plasmids/transposons can be biologically complex, rearranging and moving their DNA, resulting in frequent loss or acquisition of both complete and partial genes⁴¹. The functionality of some of these constructs can be difficult to assess, particularly when sequence quality of these regions can be highly variable.

Gram-negative specific issues not affecting E. coli

While predicting beta-lactam resistance in *E. coli* is in itself is a complex task, it is important to note it represents one of the simplest exemplars when assessing the feasibility of WGS-based resistance prediction in Gram-negatives as a whole. Many Gram-negative organisms have complex innate resistance resulting from decreased permeability (Figure 10) (e.g. *P. aeruginosa*)¹⁴⁹ and chromosomal beta-lactamases (*Enterobacter* spp.)⁷⁰ which can be difficult to identify using WGS. Thankfully *E. coli* is generally relatively permeable to most beta-lactam antibiotics and, in the absence of promoter mutations increasing expression, its

only chromosomal beta-lactamase, an *ampC* cephalosporinase, is produced constitutively at low levels.

2.3.5. Published comparisons of WGS-based and observed phenotypes

The question of whether traditional AST could be replaced by WGS-based prediction has been of intense interest over the last 10 years. At the outset of this thesis only a small number of methods/proof-of-principle studies were available ^{28,121,122,141}. However, over the last four years, methods have become increasingly complex and studies into feasibility have increased in size ¹⁵⁰. Published papers to date for *E. coli* are detailed in Table 5 (see Appendix for search strategy). In studies to date WGS-based AMR prediction is generally accurate, particularly on small datasets, but on large datasets, performance is often below recommended FDA or ISO ^{106,113} criteria. Most work to date has been carried out on *M. tuberculosis*, with multiple large population representative studies comparing WGS-based prediction vs observed phenotypes ^{29,121,151,152}. In contrast, studies examining Gram-negatives have been smaller, with resistance mechanisms in different datasets being highly diverse ^{120,152–155}. Further, these studies are hampered by several issues such as the low prevalence of resistance in some isolate sets, unclear choice of AST breakpoints used to define resistance and inconsistency of observed phenotypes on repeat testing leading to concerns over bias and misclassification of susceptibility¹⁴¹. Consequently, the feasibility of WGS-based resistance prediction in *E. coli* is unclear as measurement errors and significant differences in prevalence of resistance mechanisms between datasets prevent drawing a meaningful conclusion from results thus far.

Table 5: Summary of performance of WGS-based beta-lactam resistance prediction for *E. coli*

Study	Year	Approach [†]	Breakpoints ⁺	No isolates (No resistant)	Outcome
Amoxicillin					
Moradigaravand et al ¹⁵⁶	2018	Database guided	Non-standard	218(60)	Sensitivity = 88%, Specificity = 91%
Ampicillin					
Stoesser et al ¹⁴¹	2013	Rule based	EUCAST	74 (68)	Sensitivity = 96%, Specificity = 100%
Zankari et al ²⁸	2013	Rule based	ECOFF	48 (27)	Sensitivity = 100%, Specificity = 100%
Tyson et al ¹²⁰	2015	Rule based	CLSI	76 (72)	Sensitivity = 100%, Specificity = 100%
Peseksky et al ¹⁵⁷	2016	Database guided	CLSI	78 (74) [#]	Number major errors = 0, very major errors = 0
Her et al ⁴⁰	2018	Database guided	Unspecified	59 (.) ^o	Area under curve = 0.97
Moradigaravand et al	2018	Database guided	Non-standard	152 (81)	Sensitivity = 96%, Specificity = 83%
Amoxicillin-clavulanate					
Stoesser et al	2013	Rule based	EUCAST [*]	74 (38)	Sensitivity = 100%, Specificity = 100%
Tyson et al	2015	Rule based	CLSI	76 (34)	Sensitivity = 100%, Specificity = 100%
Moradigaravand et al	2018	Database guided	Non-standard	336 (24)	Sensitivity = 64%, Specificity = 87%
Piperacillin-tazobactam					
Shelburne et al ¹⁵⁸	2017	Rule based	CLSI	31 (26)	Sensitivity = 62%, Specificity = 96%
Cefuroxime					
Moradigaravand et al	2018	Database guided	Non-standard	376 (24)	Sensitivity = 74%, Specificity = 97%
Cefotaxime					
Zankari et al	2013	Rule based	ECOFF	48 (2)	Sensitivity = 100%, Specificity = 100%
Moradigaravand et al	2018	Database guided	Non-standard	358 (20)	Sensitivity = 91%, Specificity = 98%

†: For definition of approach see section 2.3.1.

+: Breakpoint definitions, CLSI – CLSI clinical breakpoints, EUCAST – EUCAST clinical breakpoints, ECOFF – EUCAST ECOFF breakpoints, Non-standard – Defined for the study and not equivalent to any of the three defined breakpoints, Unspecified – Not defined in the paper.

O: Not specified in paper.

#: Only 35/78 isolates in the Peseksky et al paper were *E. coli* with the remaining being other *Enterobacteriaceae*, specific *E. coli* results not presented in their work.

*: Stoesser et al tested amoxicillin-clavulanate susceptibility using both EUCAST breakpoints and E-strips

Table 5 continued

Study	Year	Approach [†]	Breakpoints [‡]	Number of isolates (Number resistant)	Outcome
Ceftriaxone					
Stoesser et al	2013	Rule based	EUCAST	74 (39)	Sensitivity = 97%, Specificity = 97%
Tyson et al	2015	Rule based	CLSI	76 (32)	Sensitivity = 96%, Specificity = 100%
Peseky et al	2016	Database guided	CLSI	78 (47) [#]	No Major Errors = 0, No Very Major Errors = 5
Ceftazidime					
Stoesser et al	2013	Rule based	EUCAST	74 (26)	Sensitivity = 95%, Specificity = 80%
Peseky et al	2016	Database guided	CLSI	78 (47) [#]	Number major errors = 5, very major errors = 5
Shelburne et al	2017	Rule based	CLSI	31 (65)	Sensitivity = 100%, Specificity = 100%
Moradigaravand et al	2018	Database guided	Non-standard	376 (17)	Sensitivity = 79%, Specificity = 99%
Kim et al ¹⁵⁹	2019	Database guided	Unspecified	28 (.) ^o	Area under curve = 0.88
Cefepime					
Peseky et al	2016	Database guided	CLSI	78 (43) [#]	Number major errors = 0, very major errors = 5
Shelburne et al	2017	Rule based	CLSI	31 (58)	Sensitivity = 100%, Specificity = 93%
Kim et al	2019	Database guided	Unspecified	30 (.) ^o	Area under curve = 0.7
Meropenem					
Stoesser et al	2013	Rule based	EUCAST	74 (0)	Sensitivity undefined, Specificity = 100%
Peseky et al	2016	Database guided	CLSI	78 (23) [#]	Number major errors = 0, very major errors = 5
Shelburne et al	2017	Rule based	CLSI	31 (3)	Sensitivity = 100%, Specificity = 100%
Kim et al	2019	Database guided	Unspecified	31 (.) ^o	Area under curve = 1.00

†: For definition of approach see section 2.3.1.

‡: Breakpoint definitions, CLSI – CLSI clinical breakpoints, EUCAST – EUCAST clinical breakpoints, ECOFF – EUCAST ECOFF breakpoints, Non-standard – Defined for the study and not equivalent to any of the three defined breakpoints, Unspecified – Not defined in the paper.

O: Not specified in paper.

#: Only 35/78 isolates in the Peseky et al paper were *E. coli* with the remaining being other *Enterobacteriaceae*, specific *E. coli* results not presented in their work.

*: Stoesser et al tested amoxicillin-clavulanate susceptibility using both EUCAST breakpoints and E-strips

2.4. Thesis outline

The goal of this thesis is to investigate whether, using current methods, short read WGS data could be used to predict beta-lactam resistance in *E. coli*. Beta-lactams and *E. coli* were chosen for several reasons. Primarily, their statuses as the one of the most common infections and the most commonly used class of antibiotics, respectively, makes them an important clinical case in which WGS-based resistance prediction would need to be successful prior to routine use. Additionally, beta-lactam resistance in *E. coli* is relatively common and the genetic basis of it has been extensively studied. Therefore, as *E. coli* is a relatively simple Gram-negative pathogen, it represents an ideal case to test WGS-based resistance prediction before moving on to more complex Gram-negative organisms and less well understood resistance. This work extends the small proof-of-principle studies that had been done prior to the start of this thesis, and aims to give a more realistic assessment of the feasibility of WGS-based resistance prediction in Gram-negatives.

Prior to the analysis chapters, Chapter 3 describes the three datasets used in this thesis, together with the laboratory and computational methods used to generate “baseline” data for each isolate. Chapter 4 then attempts beta-lactam resistance prediction for all isolates in the three datasets (1818 isolates) using the rule-based methods from prior proof-of-principle studies. Chapters 5-7 each assess a possible cause of discrepancy. Chapter 5 investigates bioinformatics causes, Chapter 6 assesses the impact of over-simplifications in the resistance prediction algorithm’s model of resistance, and Chapter 7 investigates the impact of over-simplification of the phenotype. Chapter 7 then ends with testing if phenotype and genotype can be better aligned using a database-guided resistance

prediction approach which better accounts for the nuances of both phenotype and genotype.

Chapter 3: Description of data and methods

3.1. Introduction

Work in this thesis has focussed on three sets of *E. coli* isolates:

- Unselected clinical isolates obtained from blood cultures submitted for testing at the Oxfordshire University Hospitals (OUH) NHS Foundation Trust
- Isolates containing “high-risk” resistance mechanisms (e.g. carbapenemases) selected by Public Health England (PHE) from among isolates sent to them for investigation
- Isolates obtained from pooled pig caeca from UK abattoirs by the Animal and Plant Health Agency (APHA) as part of a point prevalence study of antimicrobial resistance (AMR) in UK pig farms.

This chapter will explain how the isolates were obtained, methods used to generate WGS data (including sequencing method and quality control), and phenotyping methods. Using these data, it will then describe the baseline features of each of these datasets.

3.2. Obtaining datasets

3.2.1. OUH dataset

The OUH dataset is the largest dataset evaluated, and also the most applicable to the primary question of whether WGS-based resistance prediction could replace traditional antimicrobial susceptibility testing (AST) in clinical settings. The aim of the sampling strategy

of this dataset was to obtain a representative isolate from every case of *E. coli* bloodstream infection (BSI) identified by the OUH NHS Foundation Trust microbiology laboratory. Note this drew patients from across Oxfordshire, and did not require them to be admitted to hospital, instead simply requiring that *E. coli* was cultured from their blood in the OUH laboratory. Using this strategy, isolate sequencing data could then be linked to AST data and anonymised patient metadata via the Infectious Diseases in Oxfordshire Research Database (IORD) ¹⁶⁰. To avoid sequencing essentially the same isolates twice, *E. coli* isolated from blood within 90 days of an index BSI isolation were defined as likely recurrence and excluded ¹⁶¹. While isolates have been stored since 2008, work in this thesis focussed on isolates collected between January 2013 and August 2015. There were two reasons for selecting this date range: (i) the time required for sequencing selected isolates and (ii) significant changes in the OUH routine AST method which occurred in late 2012 making routine results from earlier isolates less comparable with the overall dataset. 1003 *E. coli* isolates with complete AST results were successfully recovered from 1000/1039 (96%) cases (997 where one *E. coli* was grown, six with two) of *E. coli* BSI which occurred during this period. This dataset was made publicly available in 2019 as part of publication of work in this thesis ¹⁶².

3.2.2. PHE dataset

As part of their role as an English national reference laboratory for antimicrobial resistance in humans, PHE's Antimicrobial Resistance and Healthcare Associated Infections (AMRHA) reference unit investigates isolates with "unusual" resistance encountered by UK clinical diagnostic laboratories. For *E. coli* specifically, their collection aims to include all carbapenem or colistin resistant isolates identified from human infections across England

¹⁶³. The PHE dataset which was used in this thesis was a subset of 346 *E. coli* isolates from this collection chosen to represent diverse resistance mechanisms and linked phenotype data. These data has not yet been published and are not yet publicly available.

3.2.3. APHA dataset

This dataset, which APHA donated for use in this thesis, consisted of linked sequencing and phenotyping data for 509 *E. coli* (as part of a wider collection of 751 Enterobacteriaceae) . These data were originally obtained by APHA as part of a cross-sectional study of AMR in pig breeder and finisher farms in the UK. This study included 57 pig farms (39 farrow-finisher and 18 finisher only) from across the UK, excluding the Northwest England, Wales and Scotland. It used samples collected by both selective (media containing 1mg/L cefotaxime, 1mg/L ciprofloxacin) and non-selective culture of pooled caecal content (collected at one of the 12 abattoirs these farms use) from 10 randomly selected healthy finisher pigs per farm. Specifically, APHA selected six distinct colonies per pool, preferentially choosing colonies which grew on the selective agar to improve the ability to detect AMR mechanisms circulating in the studied pig farms. This dataset was initially published by AbuOun et al in 2017 ¹⁶⁴.

3.3. Routine processing

3.3.1. Sequencing

Isolates were re-cultured from frozen stocks stored in nutrient broth plus 10% glycerol at -80°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. DNA extracts were then sent to the Wellcome Centre for Human Genetics (WCHG) for sequencing using the illumina HiSeq 2500 platform. Library preparation and multiplexing were performed at the WCHG, sequencing 96 or 192 (depending on the date of sequencing) isolates per flow-cell lane. Post sequencing and de-multiplexing, sequencing data were transferred onto Modernizing Medical Microbiology (MMM) servers and analysed using an in-house pipeline ¹¹⁷. This performed several functions, including; adapter and quality trimming using bbduk ¹⁶⁵(available online at <https://jgi.doe.gov/data-and-tools/bbtools/>), removal of human reads and species identification (using Kraken ¹⁶⁶ and a custom built in house database); mapping to an *E. coli* reference CFT073 (AE014075.1) using Stampy ¹⁶⁷ and *de-novo* assembling using Velvet ¹⁶⁸ and Velvet Optimiser ¹⁶⁹. Summary metrics for each mapping and assembly were obtained using SAMtools ¹⁷⁰ and QUAST ¹⁷¹ respectively. Distributions of summery metrics were then analysed to identify suitable cut-offs which would demarcate evident outliers (Figure 13). Chosen cut-offs were then applied to each isolates sequence data to check sequence quality, with sequences failing any one of these thresholds being excluded (Table 6).

As a baseline, WGS data were used to produce *in silico* multi-locus sequence types (MLST) and resistance genotypes for isolates. This was done using ABRicate (<https://github.com/tseemann/abricate>) to perform BLASTn ¹³⁵ searches of isolate assemblies for (MLST) loci in the Warwick *E. coli* PubMLST database (which follows the Achtman MLST scheme ¹⁷²) and Resfinder ²⁸ transmissible AMR genes respectively. Strict 100% coverage and identity cut-offs were used for defining the presence of specific MLST alleles. Cut-offs for resistance gene identification were set more permissively at 90%

coverage/90% identity, to overcome known assembly quality issues ¹⁷¹ and address the need to prioritise sensitivity and to include novel variants for resistance gene detection. All isolate assemblies were additionally compared using MASH ¹⁷³ to investigate dataset level differences between isolates. Resistance prediction, while performed for all isolates, is the topic of later chapters and hence is not presented in this chapter.

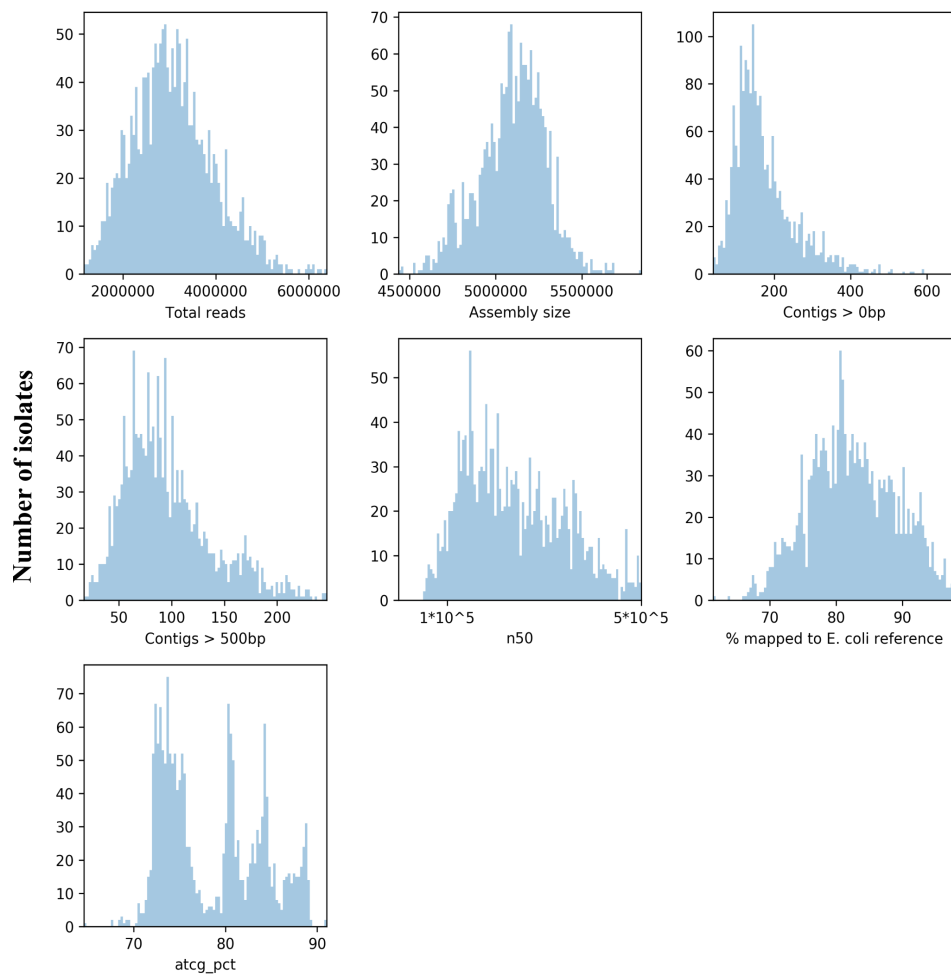


Figure 13: Histograms of sequencing QC metrics

Note: Only samples passing QC are included. Graph for n50 truncated to omit sequences with $n50 > 500,000$ (110 isolates).

Table 6: Isolate sequencing quality control

Feature	Metric	Patterns of quality issues observed (shading indicates failure)												
Total reads	<1000000 bp													
Assembly: total length	>6000000 bp													
Assembly: total length	<4000000 bp													
Assembly: no contigs >0bp long	>800 contigs													
Assembly: no contigs > 500bp long	>250 contigs													
Assembly: n50	<50000 bp													
Mapping: % reads mapped to <i>E. coli</i> reference	<60%													
Mapping: % <i>E. coli</i> reference bases called as A, C, G or T	<60%													
Total number of isolates with pattern of issues	-	2	1	1	1	1	3	2	10	1	12	2	4	1818

Note: Isolate sequencing quality control. Grey shading indicates “failures” of metric cut-off. In total, 15 isolates failed on one metric, 18 on two, three on three and four on four. 1818 isolates had no issues and were included in analyses.

3.3.2. Phenotyping

For each isolate, their linked AST data were obtained using one of two methods. AST for the OUH and PHE datasets (comprising almost three quarters of all isolates studied) was done using the OUH BD Phoenix (Beckton, Dickinson and Company) (an automated AST broth microdilution method). AST for the APHA isolates was done manually using agar dilution by the AMR and Enterics team, APHA. The APHA isolates were not available for BD Phoenix AST, limiting the extent to which these isolates could be compared with the other two datasets.

Of the two, agar dilution is the most “basic”, and was traditionally thought to be the “gold standard” phenotyping method until it was supplanted by broth dilution testing¹⁰⁴. It is still widely used by reference laboratories, given its high accuracy and relatively cheap component cost. In this method, bacteria are inoculated in a controlled way onto agar plates (for agar dilution done in this thesis, this was 5µl of a 1 in 10 dilution of an overnight broth) containing serially doubling concentrations of tested antibiotic. Plates were incubated overnight, and then assessed for growth the following morning. By its nature this method measures the minimum concentration of antibiotic required to inhibit bacterial growth (the isolate’s MIC for that antibiotic). Theoretically the MIC produced by this method is only a function of the microbes’ ability to resist antibiotic action. However, several other factors have the potential to affect test results, such as choice of agar and incubation conditions, although the differences caused by these effects have generally been shown to be small^{174,175}. All agar dilution testing performed for isolates in this thesis followed British Society of Antimicrobial Chemotherapy (BSAC) guidelines¹⁷⁶.

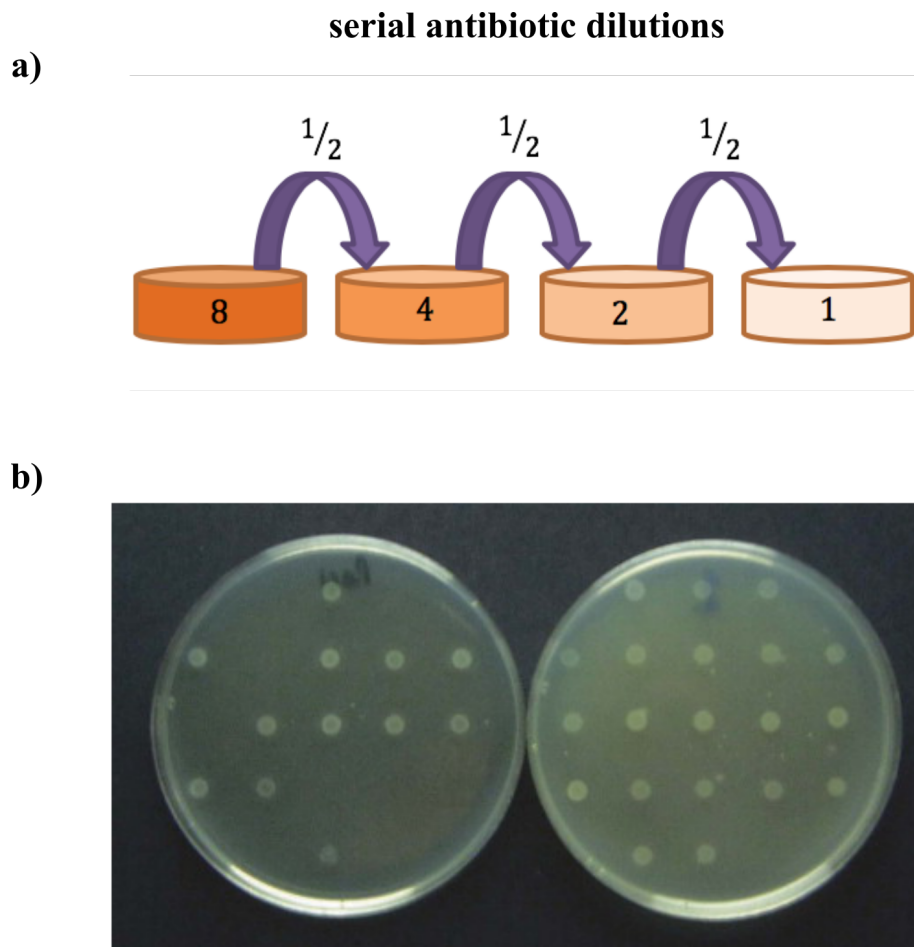


Figure 14: Agar dilution method

Note: a) in agar dilution, plates are prepared by serially diluting antibiotic (normally doubling dilutions) and incorporating into agar. Multiple isolates are then inoculated per plate as in b) and growth assessed after overnight incubation. The plate on the left in b) contains a greater concentration of antibiotic (and hence fewer isolates were able to grow). Image taken from Atlas of Sexually Transmitted Diseases and AIDS (2011), Chapter 2, Gonorrhea¹⁷⁷.

The more common way of performing AST for isolates in this thesis was use of the BD Phoenix (Becton, Dickinson and company, Sparks, MD, USA), an automated AST machine. Automated methods are now standard in most clinical microbiology laboratories, because of their ability to save time and resources and to speed up initial reporting of results while maintaining high degrees of accuracy^{104,178,179}. The BD Phoenix itself is based on broth microdilution, in which a bacterially inoculated suspension containing indicator solution is poured into premade antibiotic containing “panels”. These panels consist of wells

containing multiple serially doubling concentrations of several antibiotics. Steps in this process are standardized and automated as far as possible. The BD machine then continuously monitors for signs of bacterial growth, as measured by both well turbidity and a colour change of the oxidation-reduction indicator solution, normally producing AST results in 4-16 hours ¹⁸⁰. Compared to other automated AST platforms, studies evaluating the BD Phoenix have generally found it to have similarly high accuracy when compared to reference standard tests ^{114,181–183}.

To align with the aim of this thesis, all AST results were interpreted according to EUCAST clinical breakpoints ¹¹⁶. There were two reasons for this choice. Firstly, if WGS-based resistance prediction is to replace traditional AST, it would need to be able to predict clinically significant resistance seen in routine diagnostic laboratories. Other breakpoints, such as the epidemiological cut-off (ECOFF), a breakpoint that separates bacteria into wild and non-wild type organisms, have been postulated as alternatives ³². However, these breakpoints have not been calibrated against clinical outcomes, and even partitioning bacteria into wild and non-wild type populations is not straightforward as these groups can be challenging to define. Comparing against ECOFFs instead is more useful in a surveillance setting, where the aim is to identify important resistance mechanisms and distinguish containing isolates from wild-type ones, as opposed to dictating how individual patients should be treated. The second reason is one of practicality. For some important clinical antibiotics, ECOFFs are thus far undefined, or are set at values outside of the range regularly tested in clinical practice (Table 7). Consequently, comparing WGS-based prediction versus ECOFFs was not possible for many antibiotics studied, (Figure 15).

Table 7: EUCAST clinical breakpoints vs ECOFFs

Antibiotic	MIC Breakpoints (mg/L)			Censored MIC range ¹	
	S ≤	R >	ECOFF	Minimum	Maximum
Ampicillin	8	8	8	≤2	>8
Piperacillin	8	16	8	≤4	>64
Amoxicillin-clavulanate	8	8	- ³	≤2	>8
Piperacillin-tazobactam	8	16	8	≤4	>64
Cefuroxime	8	8	8	≤4	32
Ceftriaxone	1	2	0.125	≤2	>8
Cefotaxime	1	4	0.25	≤0.003	>128
Ceftazidime	1	4	0.5	≤1	>8
Cefepime	1	4	0.125	≤1	>8
Aztreonam	1	4	0.25	≤1	>16
Ertapenem	0.5 ²	0.5 ²	- ³	≤1	2
Meropenem	2	8	0.125 ⁴	≤0.5	>8

¹: Maximum limits of MIC tested, with isolates failing to grow in 2mg/L antibiotic having MICs ≤2mg/L and isolates growing in 8mg/L antibiotic having MIC >8mg/L

²: Ertapenem breakpoints have changed since the start of this thesis, moving from a cut-off of 1mg/L to 0.5mg/L

³: no ECOFF defined

⁴: This ECOFF is used as a screen for carbapenemase, EUCAST state resistance should be reported as found by clinical breakpoints, and that this cut-off should not be used to direct treatment decisions ¹¹⁶.

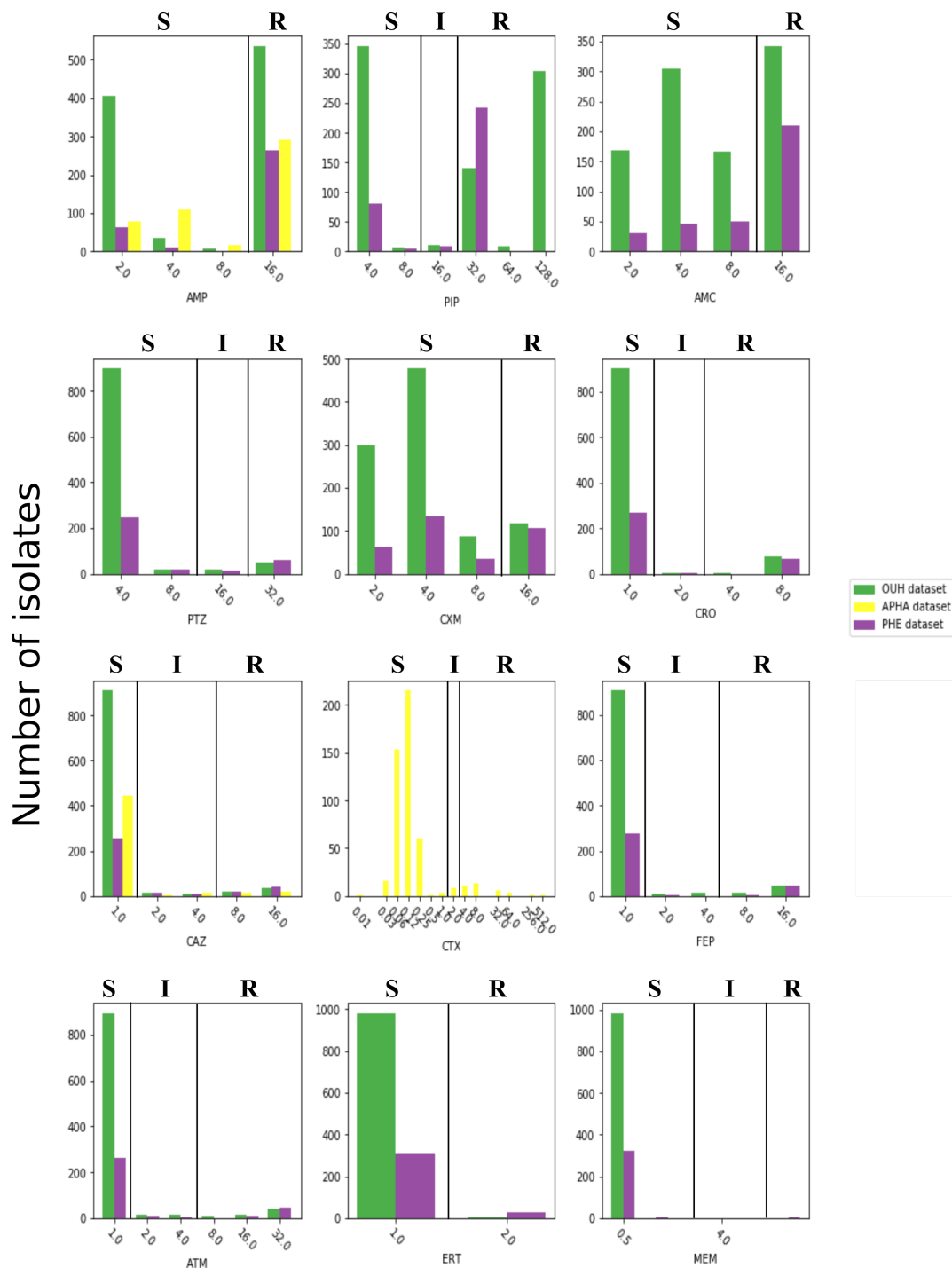


Figure 15: MIC distributions and clinical breakpoints for tested beta-lactams (n=1818 isolates)

Note: NCBI 3 letter coding for antibiotics. Penicillins; AMP (ampicillin) and PIP (piperacillin). BL/BLIs; AMC (amoxicillin-clavulanate) and PTZ (piperacillin-tazobactam). cephalosporins; CXM (cefuroxime, 2nd generation), CRO (ceftriaxone, 3rd generation), CTX (cefotaxime, 3rd generation), CAZ (ceftazidime, 3rd generation) and FEP (cefepime, 4th generation). Monobactams; ATM (aztreonam). Carbapenems; ERT (ertapenem) and MEM (meropenem).

3.3.3. Analysis

Outside the specific methods for reconstructing and analysing sequence data outlined above, data were investigated using BASH, Python 3 and IPython ¹⁸⁴ scripts and analysed statistically using STATA 14.2. Commonly used third-party python modules included:

- Biopython ¹⁸⁵, used for sequence manipulation
- Pandas ¹⁸⁶, Numpy ¹⁸⁷ and Scipy ¹⁸⁸, used for data investigation
- Matplotlib ¹⁸⁹, Seaborn (<https://github.com/mwaskom/seaborn>) and Inkscape (<https://inkscape.org/>), used for data visualisation.

3.4. Dataset description

Of the 1858 *E. coli* isolates with linked AST data available, 1818 passed all sequencing quality measures (Figure 13, Table 6). Of these, 984 belonged to the OUH dataset, 497 to the APHA dataset and 337 to the PHE dataset.

3.4.1. Relatedness of isolates

By MinHASH (MASH) similarity, the APHA isolates were genetically distinct from both the PHE and OUH sets of isolates (Figure 16). This is similar to previous findings which have demonstrated that by both phylogroups and serotypes, strains circulating in animals are generally different to strains circulating in humans ^{190–192}. This was also reflected in the MLSTs where those seen in the animal APHA isolates were largely different to those seen in the clinical isolates (Figure 17) In the APHA set, there were a total of 99 STs seen in 452 isolates, and a further 45 isolates where MLST was unclassified due to either absence from the MLST reference database or inability to find complete versions of all seven MLST genes

in assembly data. STs 58, 10 and 744 were the most common amongst APHA isolates, although there was some overlap for each. In contrast the OUH and PHE datasets were less distinguishable by MinHASH similarity and had the same four most common MLSTs: ST131, 73, 95 and 69 which together accounted for 469/984 (48%) and 160/337 (47%) isolates respectively. A major difference between the two groups was that ST131, a known important and often resistant lineage of ExPEC *E. coli* ¹⁹³, was more common in the PHE set than the OUH set (74/337 (22%) isolates vs 123/984 (12%) isolates respectively, $p < 0.001$). The other common less pathogenic MLSTs were either more frequently seen in the OUH set (e.g. ST73 161/984 (16%) OUH isolates versus 38/337 (11%) PHE isolates, $p = 0.03$ Fisher's exact test) or were similarly distributed across two groups (ST95, 91/984 (9%) OUH versus 20/337 (6%) PHE, $p = 0.29$; ST69, 94/984 (10%) OUH versus 28/337 (8%) PHE, $p = 0.59$).

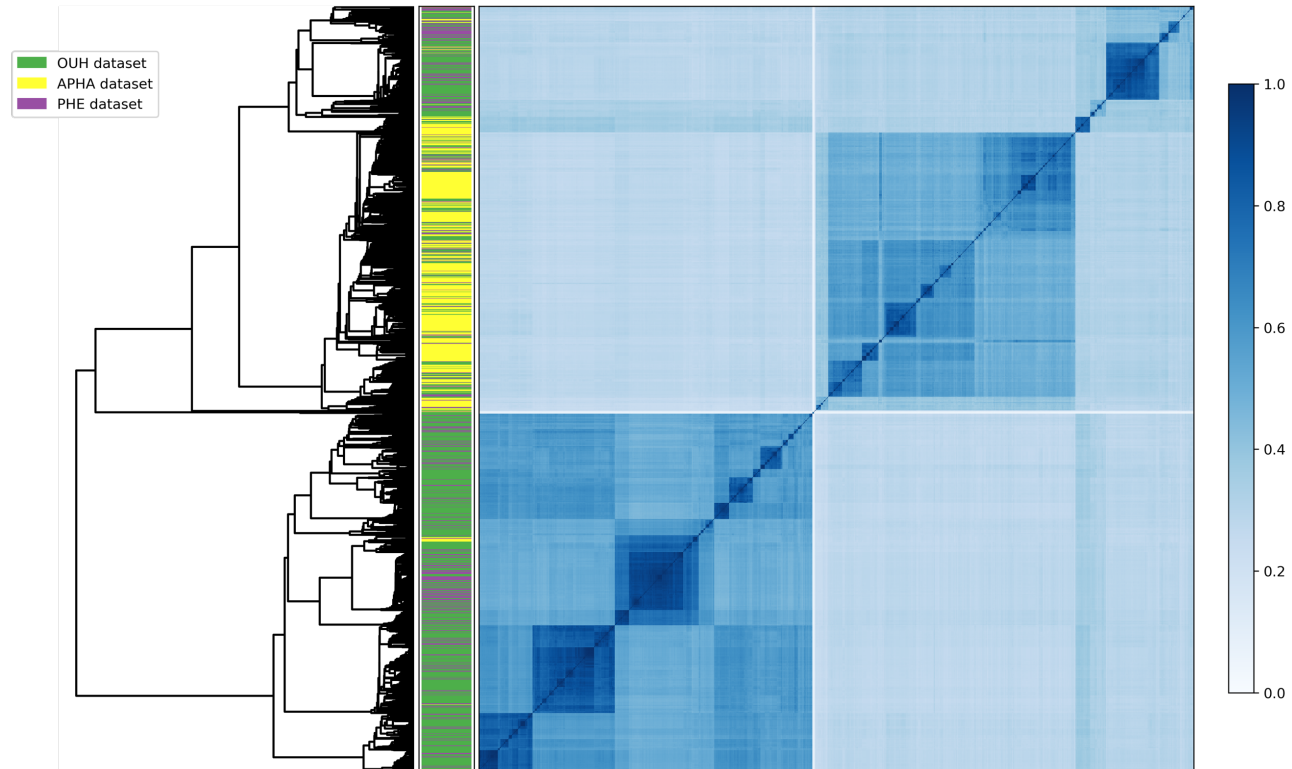


Figure 16: MASH distances of isolates

Note: From left to right, outer left panel is a clustering dendrogram (obtained using the hierarchical clustering single linkage method) based on MASH distances as estimated by SOURMASH ¹⁷³. Inner left panel denotes which dataset the isolate belongs to, and the right panels are a heatmap of sequence similarity (MASH estimated Jaccard's Index) and corresponding colormap.

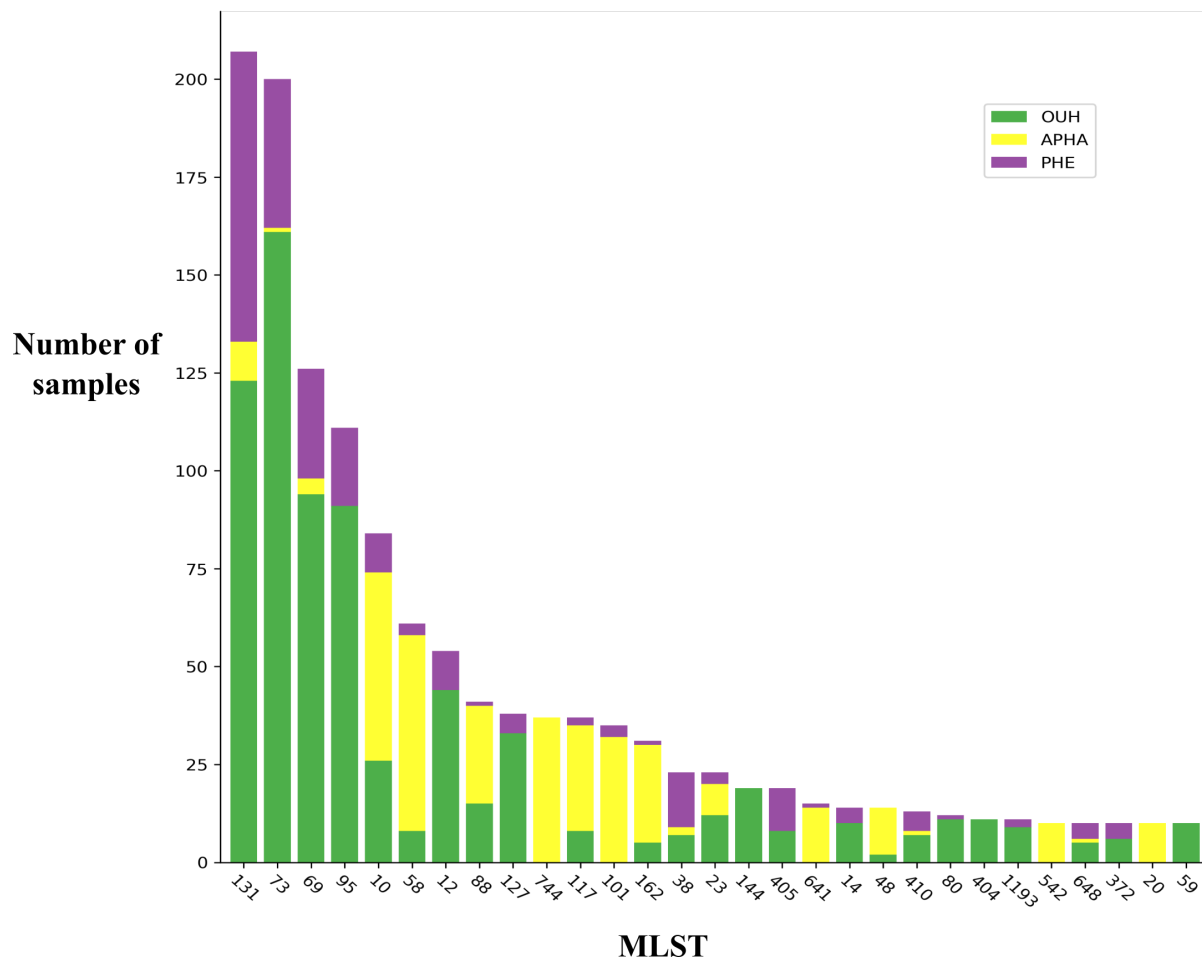


Figure 17: MLSTs in each dataset

Using core-genome SNPs calculated by comparing consensus mappings to *E. coli* CFT073, 404/1818 (22%) isolates were within 10 SNPs of another isolate. This was largely accounted for by isolates in the APHA set, with 282/497 (57%) having at least one neighbour within the APHA set itself whose core-genome was less than 10 SNPs different (as compared to 100/984 (10%) and 22/337 (7%) in the OUH and PHE sets respectively). While choice of reference may play some part (CTF073 is a uropathogenic *E. coli* strain), the close relatedness was still seen even when using mappings to *E. coli* K-12, with 56% of APHA isolates having a neighbour within 10 SNPs (derived using data published in AbuOun et al 2017¹⁶⁴).

This possibly reflects the sampling procedure (based on pooled caeca from animals on the same farm), suggesting that, despite morphological differences, some of the *E. coli* picked for sequencing were the same strain, either with different accessory genomes or with differently expressed genes. In contrast, in the OUH dataset fewer isolates had closest neighbours within 10 SNPs, instead frequently forming more loosely related clusters (Figure 18). This is again perhaps expected as all of these samples derive from a single geographical area, as compared to both the APHA and PHE datasets which are derived from isolates submitted across the UK. Between datasets, isolates were less related. Only six OUH isolates were within 10 SNPs of a PHE isolate, five of which were only related to one PHE isolate and the remaining one related to three PHE isolates. No isolates in the PHE/OUH datasets were within 10 SNPs of any APHA isolate.

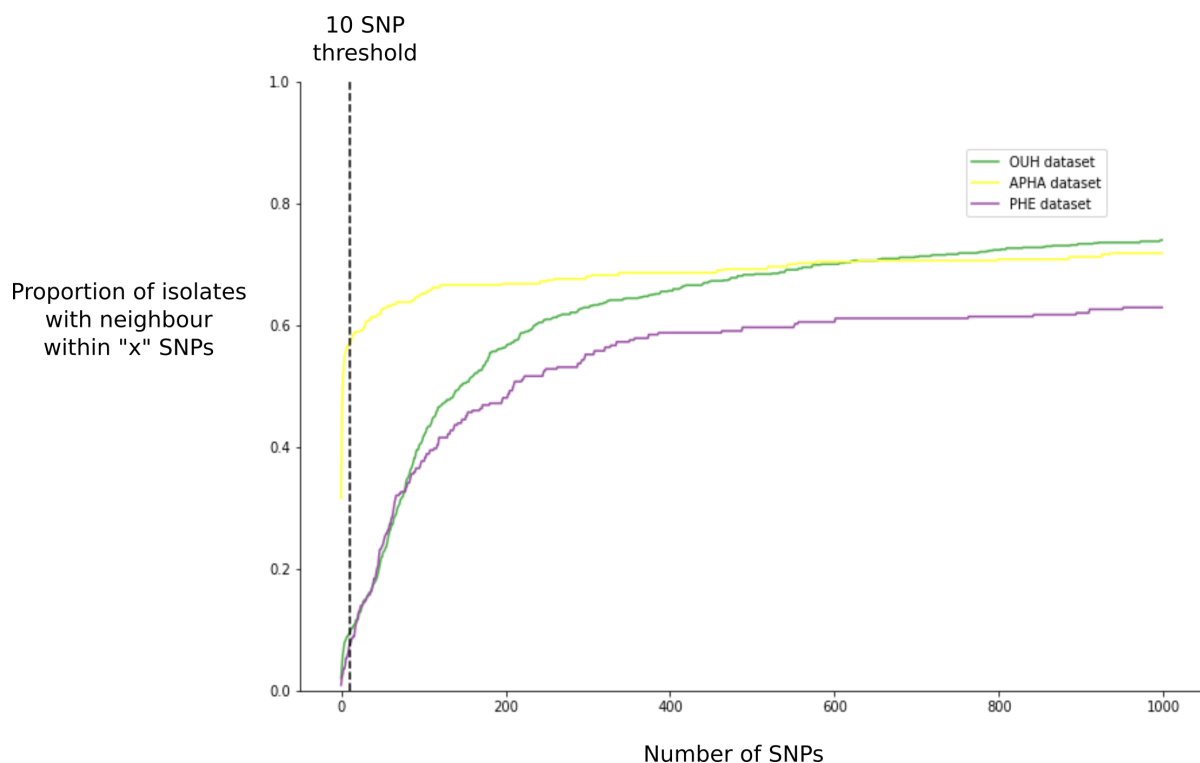


Figure 18: Relatedness of isolates within datasets

Note: SNP distance was identified using an in house script which counts differences in A, C, T, and G base calls between consensus sequences derived from mapping to *E. coli* CFT073.

3.4.2. Beta-lactam resistance genes

The isolates used in thesis collectively contained beta-lactamases from all Bush-Jacoby groups aside from 1e, 2a, 2de and 3b (Table 8). The most common beta-lactamases identified were class 2b *bla*_{TEM} genes (identified in 862/1818 (47%) isolates, 825 of which contained *bla*_{TEM-1}) and the most common ESBL was *bla*_{CTX-M} (identified in 138/1818 (8%) isolates, 100 of which contained *bla*_{CTX-M-15}). There were only 21 (1%) isolates containing carbapenemases, with 20/21 of these belonging to the PHE dataset, which is expected given the PHE sampling frame. These carbapenemase genes consisted of 13 *bla*_{OXA-48}-like, 3 *bla*_{KPC}, 7 *bla*_{NDM}, and 2 *bla*_{IMP} genes (some isolates with multiple carbapenemases). The singular carbapenemase producing *E. coli* in the OUH dataset contained a *bla*_{OXA-48} gene isolated from a haematology patient. There were no carbapenemase producers in the APHA dataset, consistent with the fact that carbapenem use in livestock is restricted. Similarly to the relative segregation of STs between pathogenic human and commensal animal strains, beta-lactamases were also slightly different for the two groups. Most noticeably, the APHA isolates contained no *bla*_{OXA-1}-like genes, compared with the OUH group which contained 61 ($p < 0.0001$, Fisher's exact test). Conversely, *bla*_{CMY} was more common in the APHA isolates than the OUH isolates (6/984 vs 18/497, $p = 0.03$). Both of these effects may be due to the presence of cefotaxime (which would select for cephalosporin as opposed to oxacillin resistance) in agar plates used to select isolates for the APHA study. The specific transmissible resistance genes for non-beta-lactam antibiotics are detailed in Chapter 3 Supplementary Tables 1 and 2.

Table 8: Beta-lactamase genes identified across datasets

Source	Bush-Jacoby class	Enzyme family	No of isolates with mechanism (%)
OUH	1	<i>bla</i> _{CMY}	6 (1)
	2b	<i>bla</i> _{TEM}	407 (41)
	2be	<i>bla</i> _{CTX-M}	73 (7)
		<i>bla</i> _{SHV}	23 (2)
	2br	<i>bla</i> _{TEM}	3 (<1)
	2d	<i>bla</i> _{OXA-1-like}	60 (6)
	2df	<i>bla</i> _{OXA-48-like}	1 (<1)
APHA	1	<i>bla</i> _{CMY}	10 (2)
	2b	<i>bla</i> _{TEM}	237 (48)
	2be	<i>bla</i> _{CTX-M}	13 (3)
		<i>bla</i> _{TEM}	8 (2)
		<i>bla</i> _{SHV}	5 (1)
	2br	<i>bla</i> _{TEM}	4 (1)
PHE	1	<i>bla</i> _{CMY}	18 (5)
		<i>bla</i> _{DHA}	1 (<1)
	2b	<i>bla</i> _{TEM}	182 (54)
	2be	<i>bla</i> _{CTX-M}	52 (15)
		<i>bla</i> _{SHV}	6 (2)
	2br	<i>bla</i> _{TEM}	5 (1)
	2ber	<i>bla</i> _{TEM}	1 (<1)
	2c	<i>bla</i> _{CARB}	1 (<1)
	2d	<i>bla</i> _{OXA-1-like}	55 (16)
	2df	<i>bla</i> _{OXA-48-like}	12 (4)
	2f	<i>bla</i> _{KPC}	3 (1)
	3a	<i>bla</i> _{NDM}	7 (2)
		<i>bla</i> _{IMP}	2 (1)

3.4.3. Isolate phenotypes

Overall, the 1818 isolates studied in this thesis had 34398 linked AST results, of which 19548 were for beta-lactam antibiotics (Table 9, Chapter 3 Supplementary Tables 3-5). Isolates in the OUH and PHE sets were tested using the BD Phoenix against a wide range of beta-lactam antibiotics, including penicillins, 2nd, 3rd and 4th generation cephalosporins, beta-lactam/beta-lactamase (BL/BLI) combination antibiotics and carbapenems. Consequently, the OUH dataset and PHE datasets accounted for 14013/19548 (72%) and 4044/19548

(21%) of linked beta-lactam AST results respectively. In contrast, AST for the APHA isolates was done using agar dilution (according to BSAC guidelines) and only tested three beta-lactam antibiotics: ampicillin (a broad-spectrum penicillin) and two 3rd generation cephalosporins (cefotaxime and ceftazidime). Agar dilution and BD phoenix results from these datasets cannot be compared as the effects of testing methodology cannot be separated from the other dataset specific features.

Table 9: Isolate beta-lactam phenotypes

Antibiotic		No. isolates tested	Phenotypic resistance classification			Percentage resistant
			Susceptible	Intermediate	Resistant	
OUH	Ampicillin	981	445	-	536	55
	Piperacillin	817	352	10	455	56
	Amoxicillin-clavulanate	981	638	-	343	35
	Piperacillin-tazobactam	984	916	18	50	5
	Cefuroxime	980	863	-	117	12
	Ceftriaxone	984	903	2	79	8
	Ceftazidime	984	908	23	53	5
	Cefepime	984	906	21	57	6
	Aztreonam	984	894	29	61	6
	Ertapenem	984	981	-	3	0
	Meropenem	984	983	1	0	0
APHA	Ampicillin	497	205	-	292	59
	Cefotaxime	497	461	11	25	5
	Ceftazidime	497	444	16	37	7
PHE	Ampicillin	337	74	-	263	78
	Piperacillin	337	86	8	243	72
	Amoxicillin-clavulanate	337	127	-	210	62
	Piperacillin-tazobactam	337	265	14	58	17
	Cefuroxime	337	231	-	106	31
	Ceftriaxone	337	268	2	67	20
	Ceftazidime	337	256	20	61	6
	Cefepime	337	278	6	53	16
	Aztreonam	337	263	17	57	17
	Ertapenem	337	307	-	30	9
	Meropenem	337	328	1	8	2
ALL	Ampicillin	1815	724	-	1091	60
	Piperacillin	1154	438	18	698	60
	Amoxicillin-clavulanate	1318	765	-	553	42
	Piperacillin-tazobactam	1321	1181	32	108	8
	Cefuroxime	1317	1094	-	223	17
	Ceftriaxone	1321	1171	4	146	11
	Cefotaxime	497	461	11	25	5
	Ceftazidime	1818	1608	59	151	6
	Cefepime	1321	1184	27	110	8
	Aztreonam	1321	1157	46	118	9
	Ertapenem	1321	1288	-	33	2
	Meropenem	1321	1311	2	9	1

Matching the genetics, the three datasets had different phenotypic profiles. The APHA isolates were predominantly either resistant to ampicillin alone (242/497, 49%) or were fully susceptible (202/497, 41%) (Figure 19). The OUH dataset was slightly more resistant, with 549/984 (56%) isolates resistant to at least one beta-lactam antibiotic, possibly reflecting the difference between the clinical pathogenic isolates in this dataset versus the commensal

nature of isolates in the APHA set. By virtue of the sampling strategy, the PHE dataset had more beta-lactam resistance, and isolates were more frequently resistant to multiple classes of beta-lactams. In all datasets, broad spectrum penicillin (ampicillin/piperacillin) was the most common resistance, consistent with the generally accepted principle that these are the most narrow spectrum agents ¹⁹⁴. In the OUH dataset, the only other beta-lactam resistance occurring in more than 10% of isolates was for amoxicillin-clavulanate (a BL/BLI) and for cefuroxime (a 2nd generation cephalosporin). In contrast, isolates in the PHE set were also commonly resistant to piperacillin-tazobactam (a BL/BLI), 3rd and 4th generation cephalosporins and aztreonam (a monobactam antibiotic).

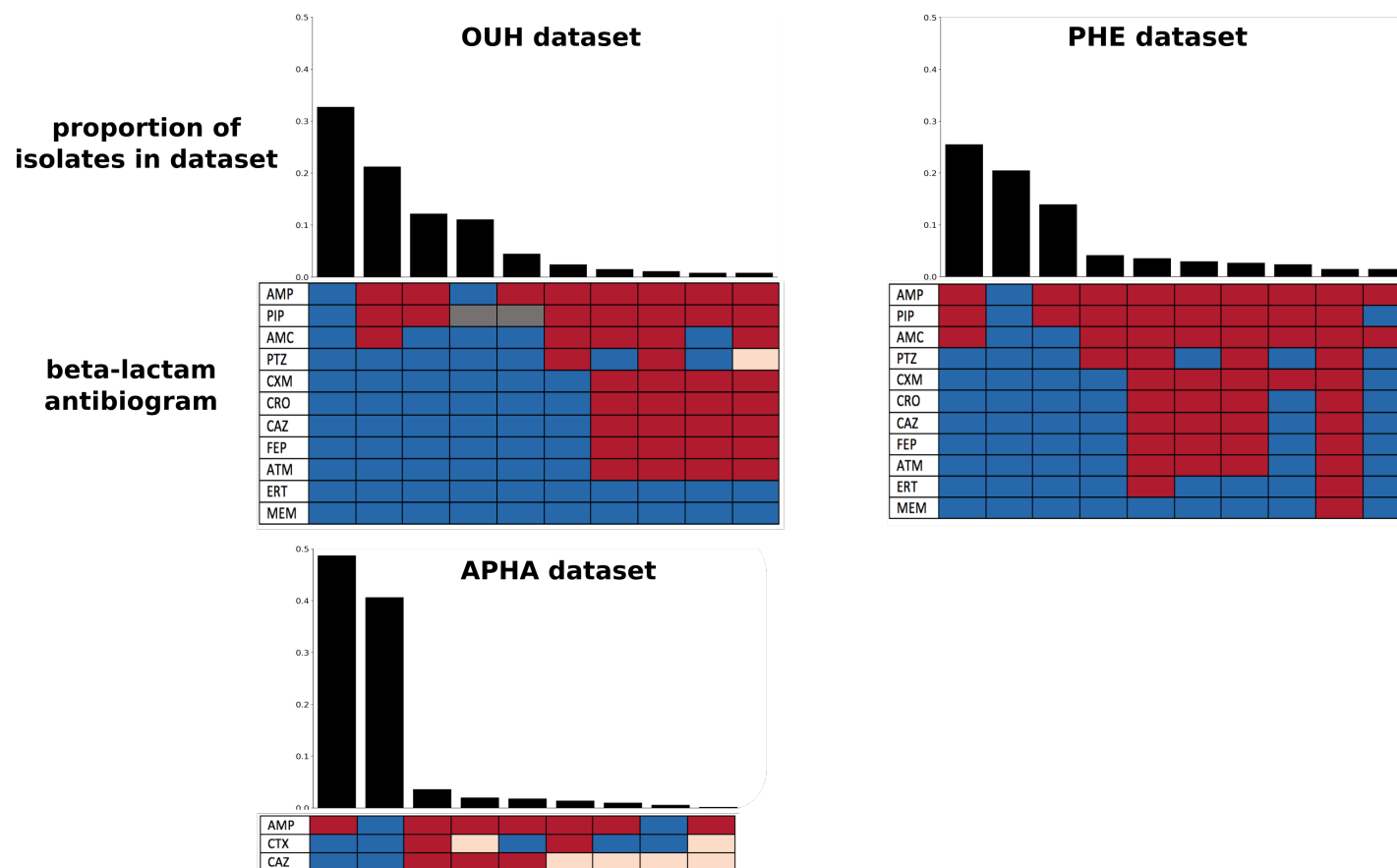


Figure 19: Ten most common beta-lactam antibiograms for each dataset

Note: NCBI 3 letter coding for antibiotics. Penicillins; AMP (ampicillin) and PIP (piperacillin). BL/BLIs; AMC (amoxicillin-clavulanate) and PTZ (piperacillin-tazobactam). cephalosporins; CXM (cefuroxime, 2nd generation), CRO (ceftriaxone, 3rd generation), CTX (cefotaxime, 3rd generation), CAZ (ceftazidime, 3rd generation) and FEP (cefepime, 4th generation). Monobactams; ATM (aztreonam). Carbapenems; ERT (ertapenem) and MEM (meropenem). Blue indicates susceptible, red, resistant, pink intermediate and grey not tested.

Consistent with previous work showing differential impact of resistance mechanisms within antibiotic subclasses (e.g. within 3rd generation cephalosporins), resistance was not consistent across BL/BLIs, cephalosporins or carbapenems in spite of their structural and/or mechanistic similarity (Figure 20) ¹⁹⁵. The issue of non-class-constant resistance does not appear unique to beta-lactams (Chapter 3 Supplementary Tables 1-5), but may be worse in beta-lactams due to the large number of antibiotics in the class, each designed with slightly different spectra.

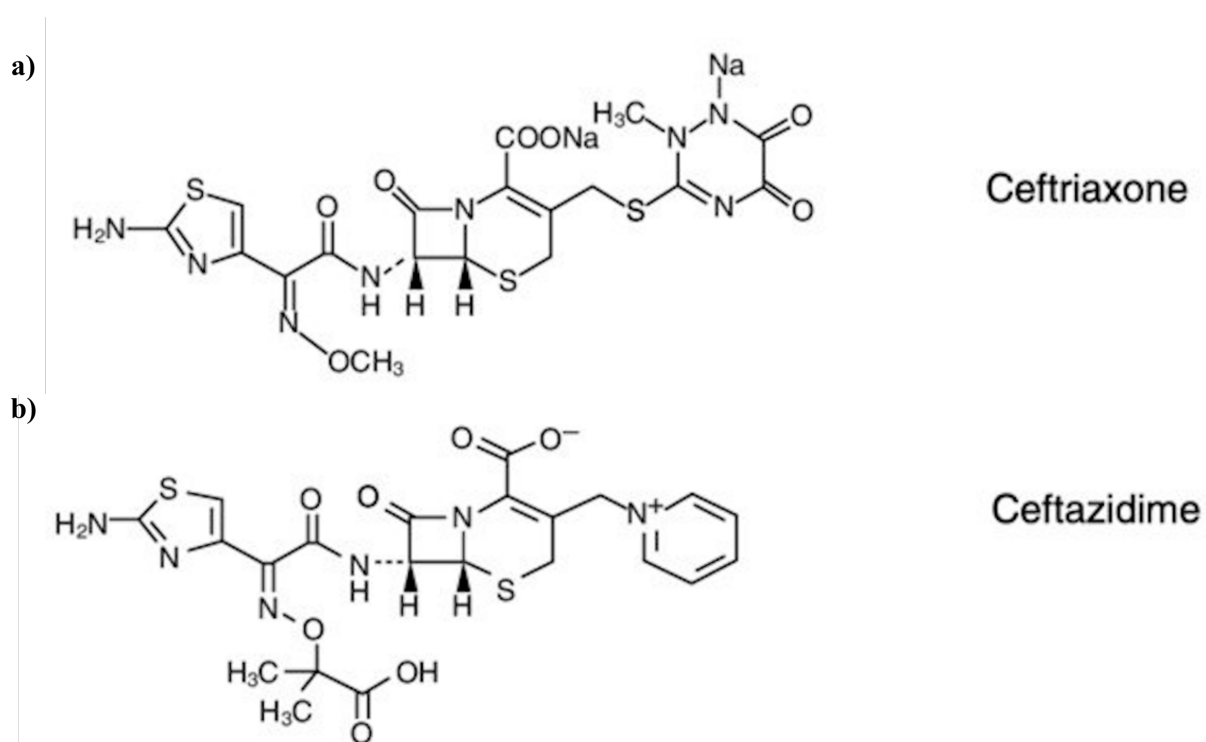


Figure 20: Structural similarity of 3rd generation cephalosporins ceftriaxone and ceftazidime

In addition to AST tests performed to predict patient outcome, some antimicrobial susceptibility tests are used primarily for detecting important beta-lactamases. One such test is MICs for ceftiofur, a 2nd generation cephalosporin. The fact that it both induces AmpC enzyme production (in organisms with inducible AmpC beta-lactamases), and is easily hydrolysed by AmpC enzymes, makes it an ideal screening test for this class of beta-

lactamases¹⁹⁶. Consequently, while EUCAST offer no clinical breakpoints for this antibiotic, they recommend this enzyme as a sensitive screen for AmpC beta-lactamases (using the ECOFF of 8mg/L, with those with MICs greater than this being possible AmpC producers). In this thesis, cefoxitin AST has only been performed for the OUH and PHE datasets, where 30/984 (3%) and 59/337 (18%) isolates screened positive respectively. This test however does not distinguish between the presence of chromosomal mechanisms of cefoxitin resistance (e.g. hyper-expression of *ampC*) and/or cefoxitin resistance mediated through horizontally acquired AmpC beta-lactamases.

3.4.4. Strengths and limitations of datasets

To my knowledge, work in this thesis is the largest study of WGS-based resistance prediction in *E. coli*. Published work to date on WGS-based resistance prediction in *E. coli* has focussed on small sets of highly selected isolates, and may not capture the nuances of resistance gene and phenotypic diversity across settings. In contrast, this thesis likely offers the closest estimate to how WGS-based beta-lactam resistance prediction may work in clinical practice, given its size, representation of both animal and human isolates and diverse resistance mechanisms, and the fact that one of the datasets used directly mirrors the task.

One potential limitation in the datasets is the apparently high relatedness of many isolates. It could be argued isolate sequences should be de-duplicated to make them more “independent” statistically. How to achieve this in practice is less clear as the definition of independent also needs to take account of the accessory genome, where most beta-lactam resistance genes reside. Assessing relatedness of accessory genome content from short read data is complex, with routine methods (e.g. Roary¹⁹⁷) often performing poorly. Given these

complexities, no studies of resistance prediction in Gram-negatives known to me have attempted to account for isolate relatedness. While there was a strong association between pairwise core-genome SNPs and MASH similarity (Figure 21), there was no obvious cut-off of either metric distinguishing related/non-related isolates. In particular, there were 64 pairs of isolates whose core genomes were less than 10 SNPs apart, but had different beta-lactamases. Three of these had core-genomes zero SNPs distant and had whole-genome MASH similarity >99%, yet were distinguished by one of the pair containing an ESBL and the other not. Consequently, in the absence of a good definition of independence, isolates in this thesis were treated as independent throughout. Somewhat reassuringly, while this is likely an over-simplification of a much more complex picture of isolate relatedness, all previous studies of resistance prediction in *E. coli* have either not assessed relatedness at all or have taken the same approach.

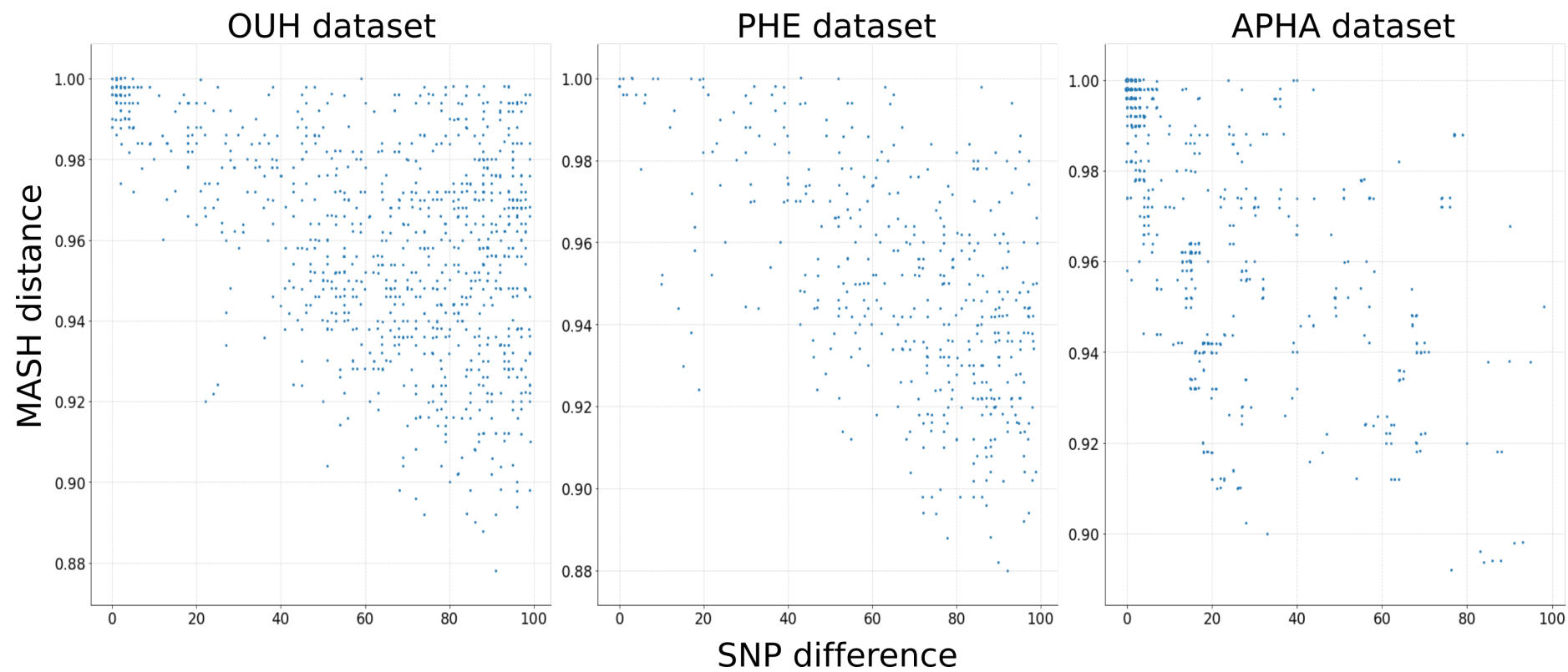


Figure 21: Association between core-genome SNP distance and MASH similarity for highly related isolates

Chapter 4: Large-scale testing of the performance of WGS-based resistance prediction in *Escherichia coli*

4.1. Introduction

At the outset of this thesis, the ability of WGS to predict resistance in clinical isolates had been demonstrated in several proof-of-principle studies. These studies all used *de-novo* assembly and BLAST¹³⁵ based approaches as, for *E. coli*, most of the resistance genes are horizontally acquired and hence not present in standard reference genomes^{28,120,141}. These studies were generally only performed on small sets of highly selected isolates, and hence how the technology would work in routine clinical practice was unclear. This chapter aimed to assess this in a more robust way, attempting to predict resistance in a much larger, relatively unselected clinical dataset using similar bioinformatic and resistance prediction algorithms to those used in the early proof-of-principle studies. To simplify analysis in this thesis, WGS-based resistance prediction with traditional AST were compared on beta-lactam antibiotics alone.

4.2. Methods

4.2.1 Baseline data and resistance prediction

As described in Chapter 3, a total of 1818 isolates (coming from the OUH, PHE and APHA datasets) were available to test WGS-based resistance prediction, with each isolate having the following:

- Quality checked sequencing data, including a *de-novo* assembly produced using Velvet ¹⁶⁸/VelvetOptimiser (<https://github.com/tseemann/VelvetOptimiser>).
- A beta-lactam resistance genotype obtained via assembly and BLAST search ¹³⁵ for genes in the Resfinder database ²⁸ of transmissible resistance genes using ABRicate (<https://github.com/tseemann/abricate>, unpublished), a publicly available program designed specifically for identifying AMR genes. Genes were determined as present if found at >90% identity and >90% width of coverage. In the situation with multiple possible matches for a particular locus, best match was defined as the match with the highest BLAST bit-score (BLAST's normalised sequence alignment score (see Abbreviations) , produced using tBLASTn with default parameters and the BLOSUM-62 (**B**LO**C**kS **S**UBstitution **M**atrix) scoring matrix for matching). Of note, for beta-lactam resistance in *E. coli*, the only known transmissible beta-lactam resistance genes are beta-lactamases.
- MIC data, obtained either via manual agar dilution or in an automated fashion using the BD Phoenix, interpreted according to EUCAST clinical breakpoints ¹¹⁶.

To make WGS-based predictions of resistance, similarly to preceding studies ¹⁴¹, beta-lactamases were classified into their Bush-Jacoby-Medeiros families ¹⁴⁵ (Figure 22) which

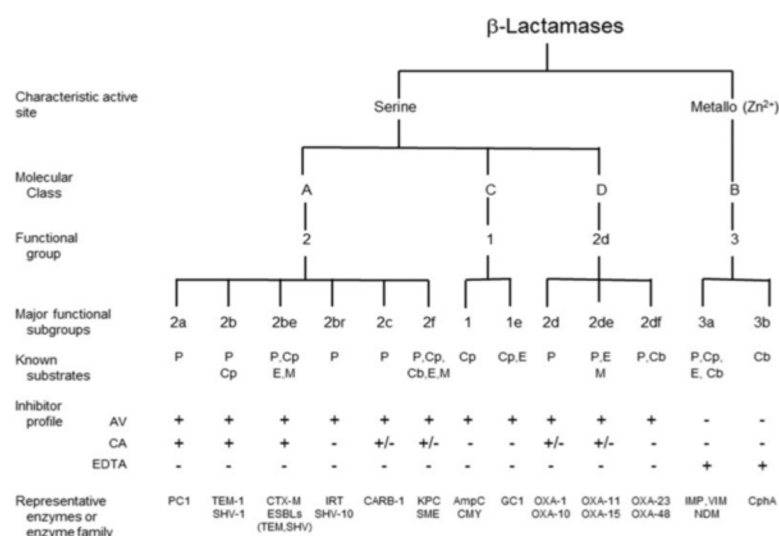
were then used to infer expected beta-lactam phenotypes. These results were then compared with observed AST.

Of note, although this was the standard approach used at the start of this work, there were several hypothesized weaknesses which could lead to suboptimal performance with this approach, including

- Issues with the genotype arising from
 - Omissions from the resistance database
 - Poor assembly quality leading to genes being missed
- Issues with the prediction arising from
 - Inexact matching being too inaccurate
 - Failing to account for *E. coli*'s chromosomal *ampC*
 - Incorrect previous association between phenotype and genotype

Post the initial prediction, this chapter investigates each one of these in turn.

a)



b)

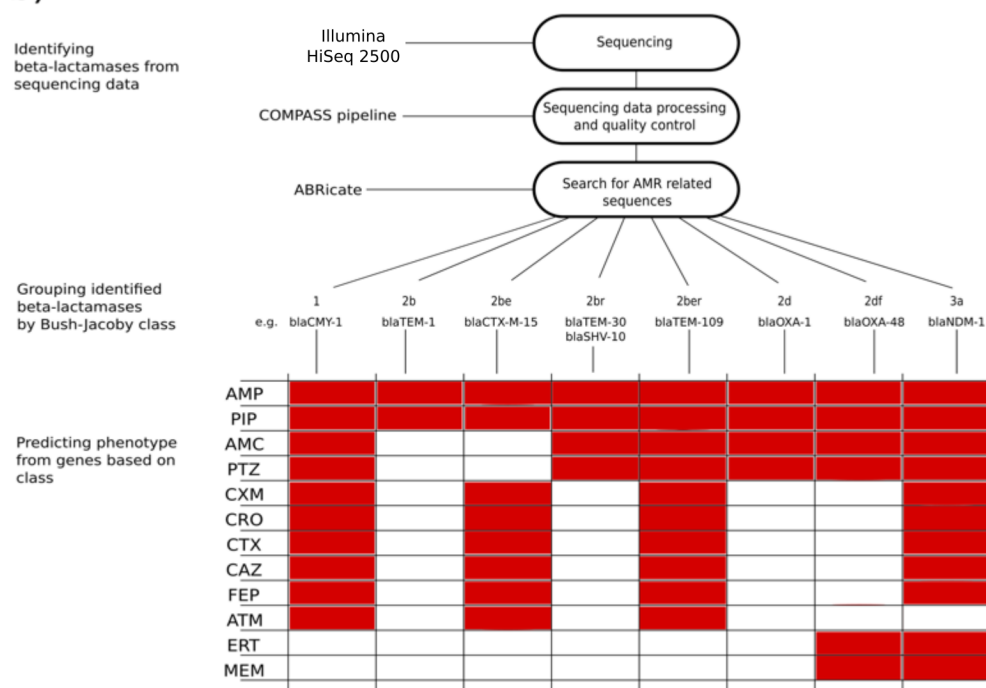


Figure 22: Beta-lactamase grouping and consequent resistance prediction

Note: a) Molecular and functional relationships among beta-lactamases. AV, avibactam; CA, clavulanic acid; Cb, carbapenem; Cp, cephalosporin; E, expanded-spectrum cephalosporin; M, monobactam; P, penicillin. Figure reproduced from Bush 2018⁴⁷. b) resistance prediction for each enzyme (red – resistance, white - susceptible). AMP (ampicillin) and PIP (piperacillin). BL/BLIs; AMC (amoxicillin-clavulanate) and PTZ (piperacillin-tazobactam). cephalosporins; CXM (cefuroxime, 2nd generation), CRO (ceftriaxone, 3rd generation), CTX (cefotaxime, 3rd generation), CAZ (ceftazidime, 3rd generation) and FEP (cefepime, 4th generation). Monobactams; ATM (aztreonam). Carbapenems; ERT (ertapenem) and MEM (meropenem).

4.2.2. Investigating potential genotypic causes of discrepancy

Resistance databases

The method I used, like most resistance prediction methods, is heavily reliant on the resistance database. However, with the rate of discovery of beta-lactamases ⁴⁷ proving a challenge to curate, there is risk that elements may be omitted from databases. To investigate if this was a problem, resistance prediction was performed using the two most widely used resistance databases (as of writing), Resfinder ²⁸ (14/01/2019 update) and CARD ¹³³ (11/10/2018 update). While both databases contain many of the most commonly encountered beta-lactamases (e.g. *bla*_{TEM-1}, *bla*_{OXA-1}, *bla*_{CTX-M-15}), the sets of beta-lactam resistance sequences they contain are not identical (Table 10). The primary differences are that the Resfinder database contained more beta-lactamases in general but the CARD database frequently contained more class C (*ampC*) beta-lactamase genes and more non-beta-lactamase resistance sequences (e.g. porins, efflux pumps). This may reflect a combination of CARD's stricter inclusion criteria (stipulating that the resistance mechanism be documented in a published paper) and the fact that Resfinder only includes transmissible mechanisms of resistance (many of the class C beta-lactamases including both *bla*_{PDC} from *P. aeruginosa* and *bla*_{ADC} from *A. baumannii* are chromosomally encoded).

Table 10: Beta-lactamases by database

Family	Ambler Class	Total proteins	Total proteins in Resfinder	Total proteins in CARD	DNA present in both databases	Protein present in both databases	Protein present in Resfinder only	Protein present in CARD only
<i>bla</i> _{OXA}	D	523	521	290	272	16	233	2
<i>bla</i> _{CTX-M}	A	200	192	139	127	4	61	8
<i>bla</i> _{TEM}	A	186	177	166	146	11	20	9
<i>bla</i> _{SHV}	A	167	150	154	134	3	13	17
<i>bla</i> _{CMY}	C	148	148	108	105	3	40	0
<i>bla</i> _z	A	134	134	0	0	0	134	0
<i>bla</i> _{IMP}	B1	66	66	43	39	4	23	0
<i>bla</i> _{VIM}	B1	53	53	39	39	0	14	0
<i>bla</i> _{ADC}	C	38	1	38	1	0	0	37
<i>bla</i> _{OKP}	A	34	34	33	31	2	1	0
<i>bla</i> _{KPC}	A	33	33	18	16	2	15	0
<i>bla</i> _{ACT}	C	33	13	33	12	1	0	20
<i>bla</i> _{PDC}	C	29	0	29	0	0	0	29
<i>bla</i> _{GES}	A	25	23	25	20	3	0	2
<i>bla</i> _{DHA}	C	24	24	17	16	1	7	0
<i>bla</i> _{NDM}	B1	24	24	14	14	0	10	0
<i>bla</i> _{CARB}	A	24	14	20	10	0	4	10
<i>bla</i> _{OXY}	A	23	22	23	21	1	0	1
Other*	*	472	292	333	141	12	139	180
Total		2236	1921	1522	1144	63	714	315

*: Other consists of beta-lactam resistance sequences belonging to families of size < 20. This group includes smaller transmissible beta-lactamase families (e.g. *bla*_{ACC} (size=10)), chromosomally encoded beta-lactamases (e.g. *Stenotrophomonas Maltophilia*'s *L1* gene) penicillin binding proteins (*mecA* variants) and sequences involved in regulating permeability (e.g. *Klebsiella Pneumoniae*'s AcrAB-TolC efflux pump and OmpK35 porin)

Assembly quality

In the method of resistance prediction I used, generation of the WGS-based resistance genotype and hence resistance prediction relies heavily on assembly quality, which puts it at risk of error ^{137,141}. This can be particularly problematic for Enterobacteriaceae, including *E. coli*, as many of their resistance genes reside on transmissible plasmids. To understand the impact of assembly method on resistance prediction, prediction was attempted using a different assembler (SPAdes ¹⁹⁸ (run with default parameters, aside from the -cov-cutoff parameter which was set to “auto”) and using a range of alternate parameters for percentage identity (PID) and width of coverage (COV) thresholds (initial set both set at 90% or greater) required for gene presence and absence (Table 11). Isolates with either discrepancies between the two assembly methods (in the allele sequence found, i.e. including different alignments as well as allele names) or containing an inexact alleles were identified using the most lax thresholds, and then assessed as to how they affected resistance prediction depending on different PID and COV thresholds.

Table 11: Tested ABRicate parameters

Percentage identity (PID) cut-off (%)	Percentage width of coverage cut-off (COV) (%)	Rationale
75	0	ABRicate defaults
90	90	SRST2 defaults
100	100	Perfect hits only

Information from these analyses was then used to create a “best” genotype from the data, from which to investigate the algorithmic causes of discrepancy.

4.2.3. Investigating potential prediction algorithm causes of discrepancy

Known potential algorithm problems – inexact matching

While previous approaches to predict resistance in *E. coli* have used “closest matching” alleles as a method to infer phenotype ¹⁴¹, the fact that some single SNP differences in beta-lactamase genes can alter the predicted phenotype raises concern that this “closest match” approach may cause errors. Other examples of WGS-based resistance prediction in other organisms have been more conservative, only predicting resistance if all mutations are “known” ¹²¹. To assess if this was a cause of poor performance, a revised resistance prediction was created where samples were only classed susceptible or resistant if all beta-lactamases present in the sample if all beta-lactamases were perfectly identified. Phenotypes were not predicted for isolates where this was not the case were (labelled as “Unknown”). Given some inexact matching alleles were only inexact due to bioinformatic artefact, all inexact matches were first screened and “corrected” if found to be definitively artefactual. This was done by BLAST-ing identified subject sequences against the Resfinder, CARD, and NCBI Redundant database using tBLASTx to find exactly matching proteins (i.e. correcting incorrectly assigned synonymous DNA variants). I assessed performance of the new prediction, discounting samples with “unknown” predicted phenotypes.

*Known potential algorithm problems – omission of *E. coli*’s chromosomal *ampC**

Of note the method of resistance prediction I used, like most previously published studies of resistance prediction in *E. coli*, only searches for transmissible beta-lactam resistance mechanisms as neither of the resistance databases includes any *ampC* promoter sequences. This omits the possibility that beta-lactam resistance could be due to mutations causing constitutive high level expression of *E. coli*’s chromosomal *ampC* gene. This was likely

omitted from previous studies for several reasons including limited understanding of how chromosomal *ampC* genes are regulated ⁷⁰, limited understanding of relevant mutations, and the difficulty assessing relative gene position using short read sequencing ¹⁹⁹. To estimate the impact of this omission on resistance prediction, results for all isolates were compared with those for isolates without AmpC presence suggested by cefoxitin (cefoxitin MIC $\leq$ 8mg/L) ¹¹⁶.

4.2.4. Understanding associations between phenotype and genotype

As the final part of this chapter and in an effort to explain any remaining discordance, the associations between specific beta-lactamases and the phenotypes of the isolates they reside in were examined. This was first done without adjustment for all beta-lactamases present in more than 5 samples and then repeated but only using isolates with singular causes of beta-lactam resistance (i.e. one single beta-lactamase and no other detected mechanisms) and beta-lactamases present *in isolation* in more than 5 isolates.

4.3. Results

4.3.1. Baseline prediction

Bioinformatic prediction parameters

- *Assembler – VelvetOptimiser*
- *Database – Resfinder (14/01/2019 update)*
- *ABRicate cut-offs – PID = 90%, COV= 90%*

WGS-based predictions poorly matched observed phenotype for several antibiotics at initial prediction (Table 12). Very major error (VME) rates for all antibiotics aside from meropenem

and ceftriaxone were worse than required to meet ISO/FDA standards (see Table 3^{105,106}).

Likewise, major error (ME) rates for all antibiotics aside from ceftazidime and cefepime were also below required standards (ME <3.0%). Of note, the WGS-based prediction very major error rates for the beta-lactam/beta-lactamase inhibitor (BL/BLI) combination antibiotics were far higher, with the resistance genotype failing to explain resistance to amoxicillin-clavulanate and piperacillin-tazobactam in 73.2% (405/553) and 70.4% (76/108) resistant isolates respectively.

Table 12:Initial prediction results

Drug	WGS- predicted phenotype	Phenotype			Very major error (VME) rate (%)	Major error (ME) rate (%)	Positive predictive value (PPV)	Negative predictive value (PPV)
		Susceptible	Intermediate	Resistant				
Ampicillin	Susceptible	717	-	94	1.1	8.6	0.99	0.88
	Resistant	8	-	996				
Piperacillin	Susceptible	423	11	34	4.9	3.4	0.98	0.93
	Resistant	15	7	664				
Amoxicillin- clavulanate	Susceptible	762	-	405	73.4	0.5	0.97	0.65
	Resistant	4	-	147				
Piperacillin- tazobactam	Susceptible	1162	31	76	70.4	1.6	0.63	0.94
	Resistant	19	1	32				
Cefuroxime	Susceptible	1065	-	72	32.3	2.7	0.84	0.94
	Resistant	29	-	151				
Ceftriaxone	Susceptible	1139	1	4	2.7	2.8	0.82	1.00
	Resistant	32	3	142				
Cefotaxime	Susceptible	452	4	5	20.0	2.0	0.69	0.99
	Resistant	9	7	20				
Ceftazidime	Susceptible	1549	31	25	16.6	3.7	0.68	0.98
	Resistant	60	27	126				
Cefepime	Susceptible	1097	3	5	7.3	4.5	0.55	1.00
	Resistant	87	24	105				
Ertapenem	Susceptible	1287	-	9	27.3	0.1	0.96	0.99
	Resistant	1	-	24				
Meropenem	Susceptible	1295	1	0	0.0	1.2	0.33	1.00
	Resistant	16	1	8				
Aztreonam	Susceptible	1115	25	8	6.8	3.6	0.72	0.99
	Resistant	42	21	110				

Of note, sensitivity to identify amoxicillin-clavulanate, piperacillin-tazobactam and ceftazidime resistance depended on the source of the isolates (χ^2 $p < 0.0001$) (Figure 23), being generally higher for isolates from the PHE set compared with the OUH set, and particularly low for APHA for ceftazidime (20/37 (54%) resistant isolates correctly identified). Although there was also an apparent difference in sensitivity of ertapenem resistance prediction, this was not statistically significant ($p = 0.18$), potentially reflecting low numbers of resistant isolates ($n = 33$).

4.3.2. Potential genotypic causes of poor predicted versus observed phenotype agreement

Database choice

Bioinformatic prediction parameters

- *Assembler – VelvetOptimiser*
- *Database – CARD (11/10/2018 update).*
- *ABRicate cut-offs –PID = 90%, COV = 90%*

On initial review there appeared to be major discrepancy between the beta-lactam genotypes created by the two databases, with differences in predicted beta-lactamases present occurring in 828/1818 (45.5%) isolates (and specifically 828 (82.5%) of the 1004 isolates containing any beta-lactamase from either database). However most of these discrepant were artefactual, being due to one of two issues in all but 11 cases. The first and most common issue was the mis-identification of alleles caused by DNA-matching as opposed to protein matching, occurring in 728 isolates (674 as an isolated error, 54 in combination with other errors). This most commonly occurred when the search using the

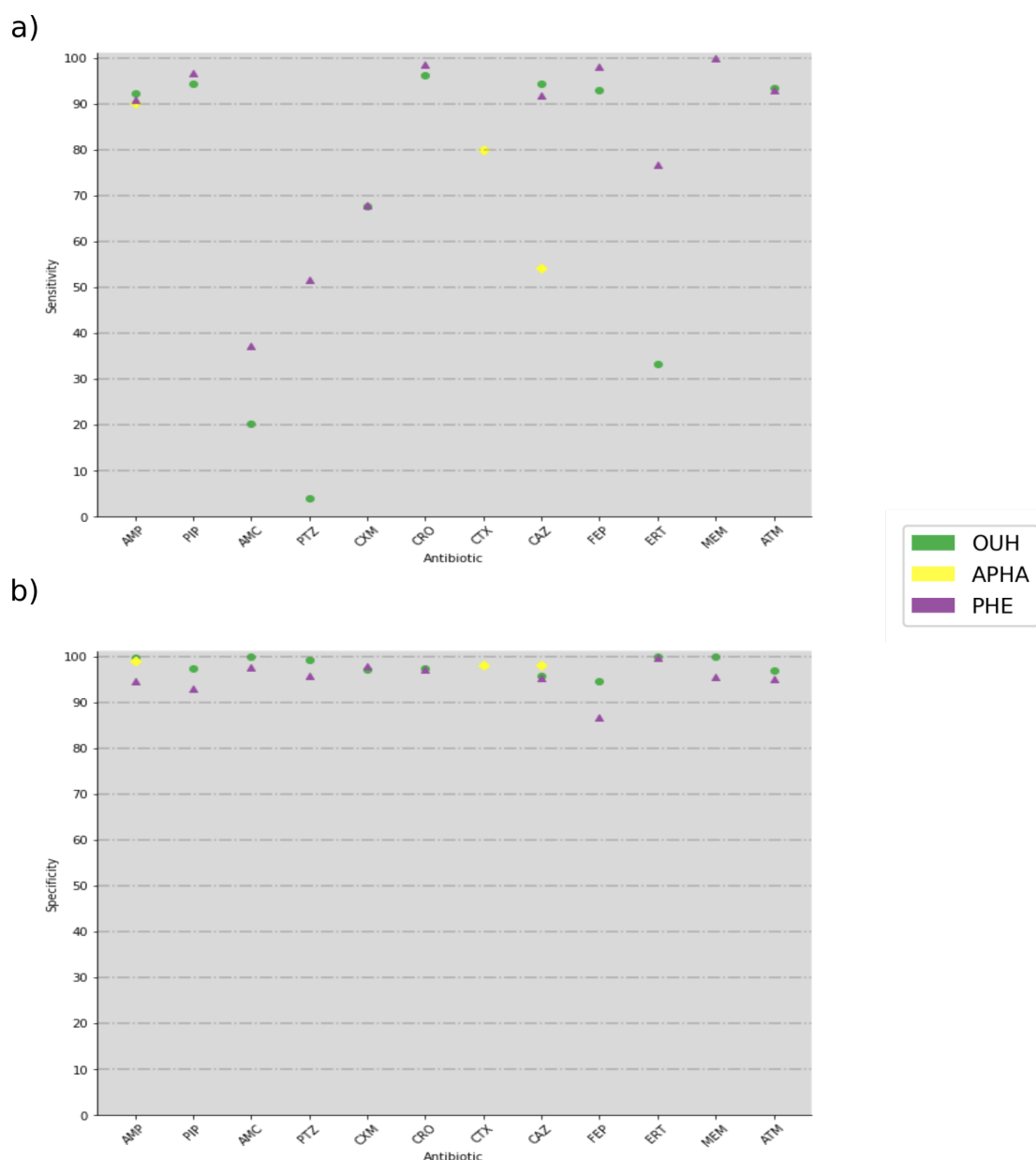


Figure 23: Sensitivity (a) and specificity (b) of WGS-based resistance prediction by isolate set

Note: NCBI three letter codings for antibiotics. AMP -ampicillin/amoxicillin, PIP-piperacillin, AMC - amoxicillin-clavulanate, PTZ – piperacillin-tazobactam, CXM – cefuroxime, CRO - ceftriaxone, CTX – cefotaxime, CAZ – ceftazidime, FEP – cefepime, ERT – ertapenem, MEM- meropenem, ATM – aztreonam.

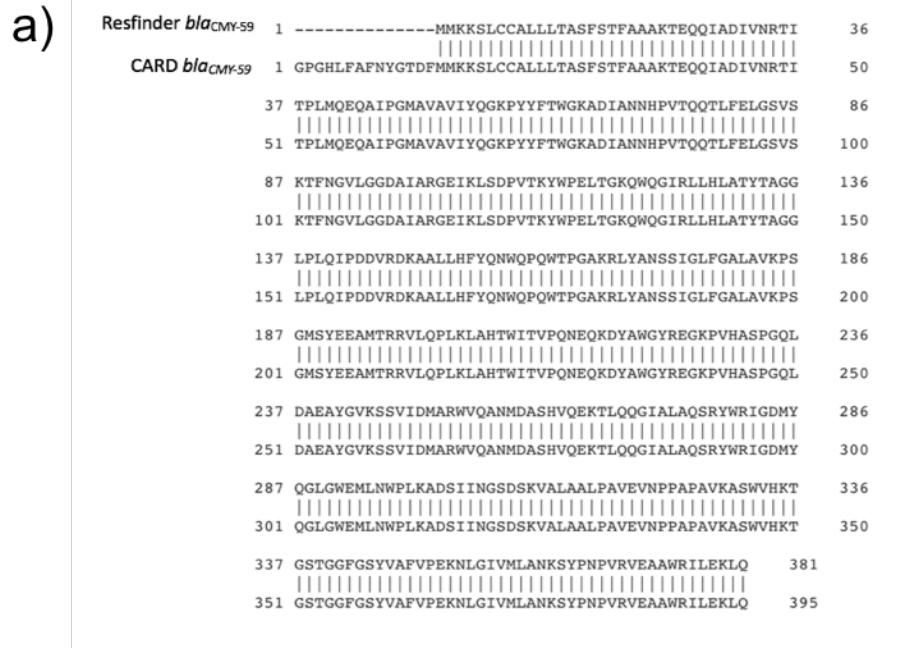
CARD database mis-identified a DNA variant of *bla*_{TEM-1} as *bla*_{TEM-186}, occurring in 659 cases.

This is due to the CARD database only containing one variant of *bla*_{TEM-1}, compared to the

Resfinder database which contains four commonly seen variants. This would not cause

issues if resistance elements in queried sequences were matched by the proteins found (e.g. by using translated BLAST rather than nucleotide BLAST). However, most general resistance typing programs do not take this approach as it requires additional processing time to produce reliable consensus protein sequence from reads and reduces flexibility to look for non-protein elements. For example, ABRICATE uses DNA as opposed to protein searches to enable search for non-protein elements such as mobile elements (e.g. transposons) and virulence factors.

The second issue accounting for 145 discrepant (91 as an isolated error) was errors caused by 'poor quality' database sequences. An example of how this causes error is *bla*_{CMY-59} (<https://card.mcmaster.ca/ontology/38469>) from the CARD database. The CARD sequence for *bla*_{CMY-59} contains extra "coding sequence" prior to the start of the sequence (Figure 24). While there may be a shorter perfectly matching allele in the underlying isolate sequence, longer alleles can produce longer alignments with higher 'alignment scores'. In deciding the optimum allele, many algorithms, including BLAST, rank matching alleles by these scores and hence the perfect match may be missed in favour of longer but less accurate matches.



Gene	Identity (%)	Coverage (%)	Alignment Length	Bit Score
bla _{CMY-2}	100.0	100.0	1146	2117
bla _{CMY-59}	99.9	100.0	1188	2189

Figure 24: Search issues caused by poorly curated sequence

Note: a) alignment of the Resfinder and CARD *bla*_{CMY-59} genes showing the CARD containing additional sequence prior to the Resfinder reference. b) Multiple sequence alignment of a discrepant isolate sequence, the CARD *bla*_{CMY-59} and *bla*_{CMY-2} genes, with alignment scores. Scores based on bit-score will identify *bla*_{CMY-59} as the optimal alignment, whereas scoring systems which use percentage identity will identify *bla*_{CMY-2} as optimal.

As a corollary to this issue, incomplete sequences (e.g. CARD's *bla*_{TEM-117} <https://card.mcmaster.ca/ontology/37360>) causes the inverse issue for methods (e.g. SRST2, considered next) which rank alleles depending on the percentage identity and width of coverage of matches. For example, it is often possible to obtain 100% identity and coverage matches to a subsection of a gene, even when the complete gene is not present in the search database.

Despite the alleles identified being highly discordant, the impact of database choice on resistance prediction was minimal, only affecting 22/828 (3%) discrepant. 19 of these were caused by ABRicate mis-identifying *bla*_{TEM-1} (a Bush-Jacoby class 2b enzyme) as *bla*_{TEM-19} (a Bush-Jacoby class 2be enzyme) when using the CARD database, due to matching DNA as opposed to protein sequences. In all 19 of these cases the phenotype was consistent with the isolate containing a class 2b beta-lactamase. The remaining 3 discrepancies were isolates where ABRicate using Resfinder identified *bla*_{TEM-1} (a class 2b enzyme) and using CARD identified *bla*_{TEM-73} (a class 2br enzyme). When examined, the allele was actually a *bla*_{TEM-33} (a class 2br enzyme) variant which was omitted from both databases. On phenotypic testing, all 3 were resistant to co-amoxiclav; however only one was resistant to piperacillin-tazobactam, meaning the phenotype, while suggestive that the isolate may contain a 2br enzyme, is not a conclusive arbiter in this situation.

Assembly quality

Bioinformatic prediction parameters

- *Assembler – SPAdes (v3.12.0), VelvetOptimiser*
- *Database – Resfinder (14/01/2019 update)*

- *ABRicate cut-offs – PID=75%, COV=0% (ABRicate Defaults)*

Note for this section I have used ABRicate default cut-offs as the difference between assemblers is more identifiable and the causes more evident when using the less stringent cut-offs. The effect of cut-off and assembler selection are then examined together in Figure 26.

In total there were 161/1818 (9%) isolates which contained any inexact beta-lactamase alleles by either assembly method. 106/161 (66%) contained inexact elements for both SPAdes and VelvetOptimiser assemblies, and 43/161 (27%) and 12/161 (8%) for VelvetOptimiser and SPAdes alone respectively. While fewer isolates contained fragmented beta-lactamase genes when using SPAdes assemblies, this difference was not statistically significant ($p=0.056$, Fisher's exact test). 82/1818 (5%) isolates had different resistance genotypes between the two methods when using ABRicate's lax cut-offs, all of which occurred in isolates containing at least one inexactly assembled beta-lactamase (by either method). Correspondingly, 79/1818 (4%) isolates contained inexact elements but the genotype reported using each assembler was the same.

Categorising inexact elements into groups depending on whether or not output from the different assemblers matched gave some insights into the cause of discrepancies and how assembly method/coverage cut-off affected predicted phenotype. 52/79 isolates with inexact elements but matching genotypes contained resistance sequences which were incorrectly matched by ABRicate (25 *bla*_{SHV}, 24 *bla*_{TEM}, two *bla*_{IMP}, and one *bla*_{DHA}) on the basis of DNA alignment, similarly to as described above when investigating the impact of database above. Nine contained novel beta-lactamases (five *bla*_{TEM} variants, three *bla*_{CTX-M} variants and one *bla*_{OXA} variant). All of these elements would be identified using high COV

and PID thresholds (including the original baseline cut-offs of 90%, 90%). Of the remaining 18, only partial genes were found in 14, and genes with missense mutations were found in four. The genes with missense mutations again would be identified using high COV and PID thresholds, unless removed through examining alignments. The partial genes would be removed when applying the cut-offs used in the baseline method.

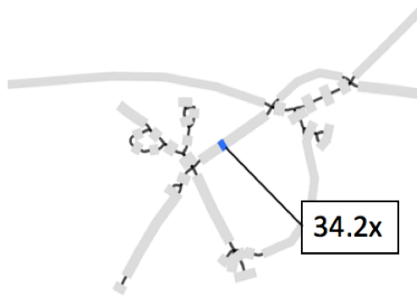
The underlying causes of inexact matching for the 82/1818 where SPAdes assembly and VelvetOptimiser assembly based genotypes disagreed were less clear. The predominant issue, which occurred in 57/82 (70%) discrepant, was fragmentation of *bla*_{TEM} gene assemblies by VelvetOptimiser (i.e. containing contigs which covered the entire gene, but the gene being seen in more than one contig) (this also happened for one *bla*_{CMY} gene). In 43 of these isolates, SPAdes identified a single known *bla*_{TEM} allele, a complete but unknown *bla*_{TEM} allele was identified in six, and for eight SPAdes assemblies were also fragmented. While one could assume these may be sequencing or assembler error, 47/57 contained heterozygous positions (variants at >20% overall coverage depth) in read pileups and had “bubbles” in their VelvetOptimiser assembly graphs, possibly suggesting the presence of more than one allele in the isolate (Figure 25). The majority of the remaining differences between results for each assembler were due to partial genes contained in one or both of the assemblies (12 SPAdes alone, one VelvetOptimiser alone and 12 both).

Almost all of the partial genes or genes with fragmented assemblies would be excluded using stringent PID and COV cut-offs. However, this may be detrimental to resistance prediction as 55/57 isolates with fragmented beta-lactamases and 19/25 isolates with

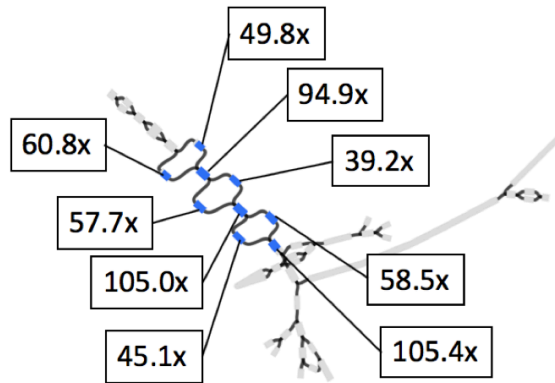
partial beta-lactamases had resistance to at least one beta-lactam antibiotic with no other apparent alternate cause.

Like database choice, the choice of both assembler and cut-off only had small effects on the rule-based resistance prediction for all antibiotics (Figure 26), with effects being dependent on antibiotic. In all cases, method changes altered less than 10% of predictions even when switching between the most discordant results. Consistent with the finding that ABRicate more often found complete elements when searching SPAdes assemblies (see above), ABRicate using SPAdes assemblies was generally less dependent on cut-offs.

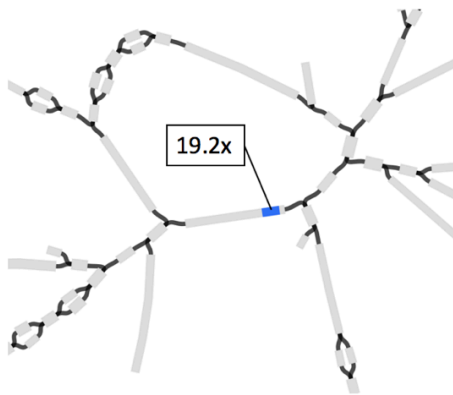
a) *bla*_{TEM} no assembly issue



b) *bla*_{TEM} assembly issue



c) *bla*_{CMY} no assembly issue



d) *bla*_{CMY} assembly issue

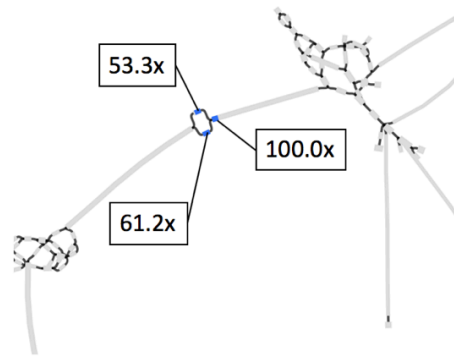


Figure 25: Examples of "bubbles" in assembly graphs for *bla*_{TEM} and *bla*_{CMY} genes

Note: Assembly graphs blasted for *bla*_{TEM} and *bla*_{CMY} genes respectively using Bandage²⁰⁰:
a), c) demonstrating graphs with no-issues in these genes, together with node coverage
depth b), d) examples with errors.

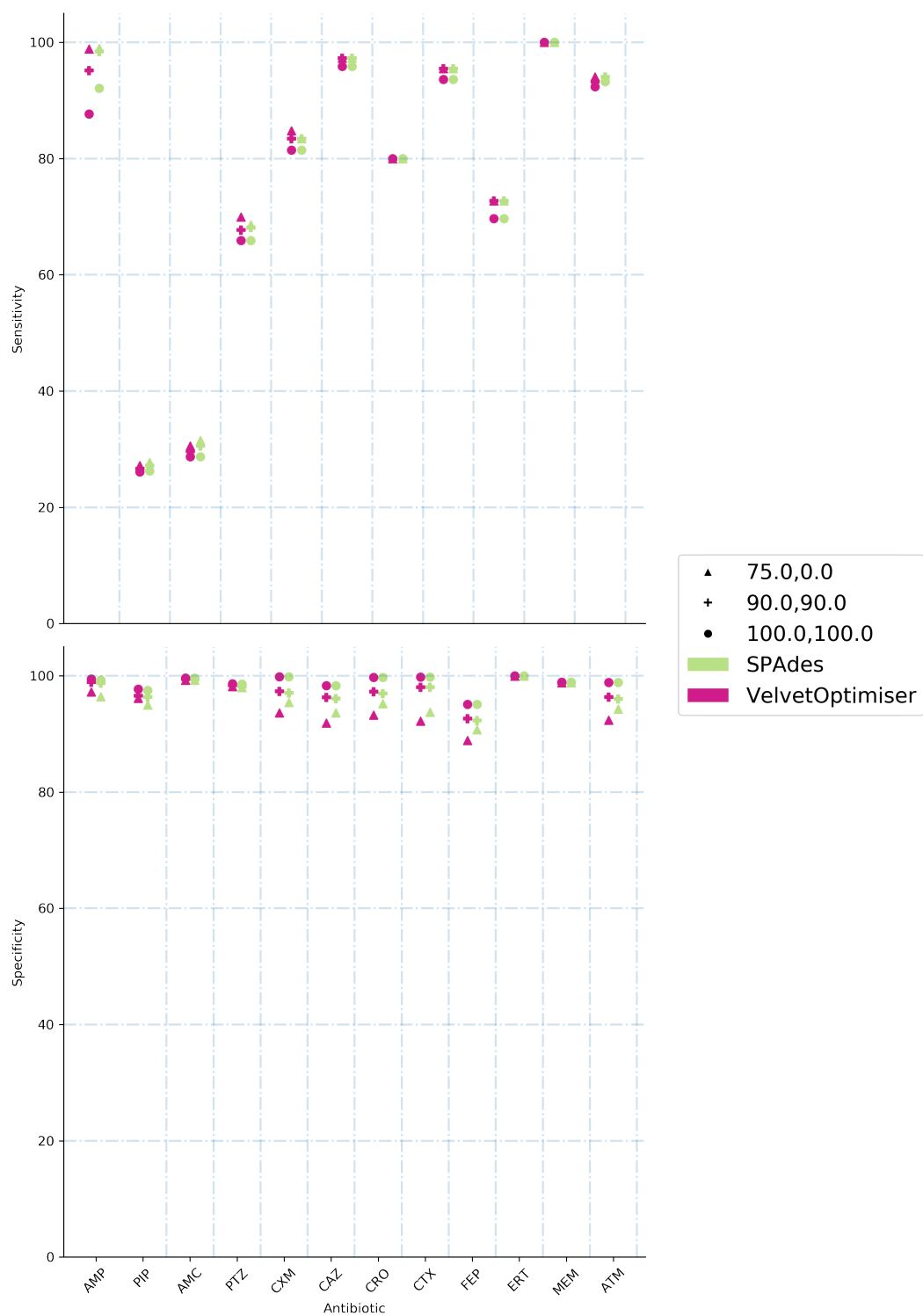


Figure 26: Comparing resistance predictions for different assemblers and cut-offs

Note: Antibiotic abbreviations as in Figure 23. Marker shapes indicate assembly quality thresholds, namely triangle (PID=75%, COV=0%), cross (PID=90%, COV=90%), and circle (PID=100%, COV=100%).

Although genotype, assembly method and cut-offs had little effect on phenotype, there were advantages beyond phenotyping prediction for certain methods. Using the SPAdes assembler resulted in fewer fragmented assemblies and results being less dependent on cut-offs. Likewise, using Resfinder database minimised artefactual errors. Given this, and the 57 isolates with definitively identified artefactual (e.g. protein matching vs DNA based errors, or known (in CARD/NCBI) variants omitted from the Resfinder database), a new baseline prediction was produced. This used the SPAdes assembler, Resfinder database, PID and COV cut-off 90% and 90% respectively and had identified artefactual errors corrected (using tBLASTx of subject sequences, method described as part of the investigation of inexact alleles, see section 4.2.3). The identity and coverage cut-offs were kept at 90% and 90% respectively given the inexact but likely functional elements identified and the need to balance sensitivity and specificity for resistance prediction.

4.3.3. Prediction algorithm causes of discrepancy

Estimating the impact of known potential algorithm problems

To investigate impact of the known problems of prediction using inexact alleles and prediction ignoring the effects of *E. coli*'s chromosomal *ampC*, secondary analyses were performed excluding isolates likely affected by these problems (identified by looking at match statistics and ceftazidime MIC respectively).

Genuine inexact matching (i.e. non-artefactual post correction as part of the revised baseline prediction as described above) only had a small effect on results, affecting only 25/1818 (1.3%) isolates in total. Excluding these from analysis resulted in small increases in specificity at the cost of small decreases in sensitivity (Figure 27, baseline (red triangle) vs

restricted set (purple diamond)). Excluding these elements did not modify any of the VME/ME rates sufficiently to alter the overall assessment of WGS-prediction performance.

To assess the impact of omitting *E. coli*'s *ampC*, 586/1818 (32.2%) isolates were excluded from analysis (497 APHA isolates with no cefoxitin AST results and 89 OUH/PHE isolates with cefoxitin MIC >8mg/L). Excluding these isolates improved WGS-based prediction performance for some antibiotics (cefuroxime, ceftazidime and cefepime), but did not universally improve performance, with far poorer sensitivity for piperacillin-tazobactam resistance in the restricted set. This may be because, by excluding cefoxitin resistant isolates, all isolates producing AmpC-like beta-lactamases are excluded (i.e. excluding isolates which produce AmpC-like beta-lactamases via horizontally acquired genes, e.g. *bla_{CMY}*). Of note however, excluding these isolates reduced ceftazidime VME rates significantly (from 16.6% to 1.4%, $p < 0.0001$) to below the required FDA/ISO thresholds, as well as reducing the VME rates for other cephalosporin and penicillin antibiotics.

When added to methodological improvements, the incremental effects of excluding isolates with either potential issue (604/1818 (33.2%) isolates excluded) were important with final MEs and VME being improved for most antibiotics (Table 13). Importantly, in this final prediction, several antibiotics (namely ampicillin, piperacillin, ceftriaxone, ceftazidime, cefepime and aztreonam) now had VME/ME rates below or close to FDA/ISO thresholds. Several antibiotics, in particular BL/BLIs, however still had VMEs orders of magnitude larger than upper acceptable limits.

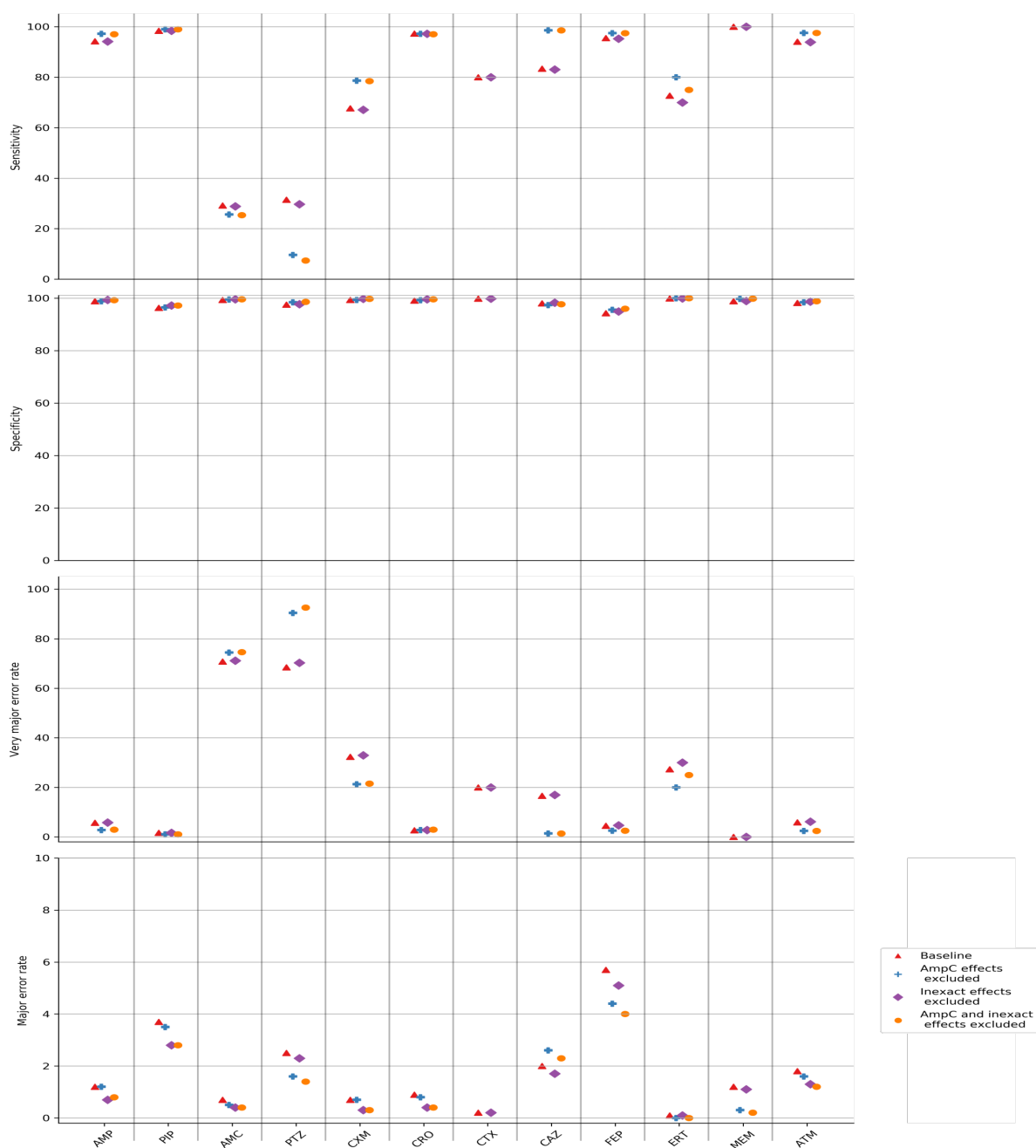


Figure 27: Impact of known potential algorithm problems on resistance prediction

Note: Red triangle, baseline prediction using SPAdes assembly and ABRicate cut-offs (PID=90%, COV=90%). Blue cross, prediction in isolates with cefoxitin MIC ≤ 8 mg/L (i.e. non ampC-like beta-lactamase producers). Purple diamond, excluding isolates with present (PID >90%, COV>90%) but inexact matching beta-lactamases. Orange circle, excluding both AmpC-like beta-lactamase producers and those with inexact elements.

Table 13: Changes in VME/ME rates with method changes

Antibiotic	Initial method*		SPAdes reassembly+		Uncertain isolates excluded and SPAdes reassembly†	
	VME Rate (%)	ME Rate (%)	VME Rate (%)	ME Rate (%)	VME Rate (%)	ME Rate (%)
Ampicillin	1.1	8.6	5.7	1.2	2.9	0.8
Piperacillin	4.9	3.4	1.6	3.7	1.0	2.8
Amoxicillin-clavulanate	73.4	0.5	70.8	0.7	74.6	0.4
Piperacillin-tazobactam	70.4	1.6	68.5	2.5	92.6	1.4
Cefuroxime	32.3	2.7	32.3	0.7	21.5	0.3
Ceftriaxone	2.7	2.8	2.7	0.9	2.9	0.4
Cefotaxime	20.0	2.0	20.0	0.2	-°	-°
Ceftazidime	16.6	3.7	16.6	2.0	1.4	2.3
Cefepime	7.3	4.5	4.5	5.7	2.5	4.0
Ertapenem	27.3	0.1	27.3	0.1	25.0	0.0
Meropenem	0.0	1.2	0.0	1.2	-°	0.2
Aztreonam	6.8	3.6	5.9	1.8	2.4	1.2

*: VelvetOptimiser assembly, ABRicate cut-offs (PID=90%, COV=90%), uncertain (inexact/ ampC producing) isolates included

+: SPAdes assembly, ABRicate cut-offs (PID=90%, COV=90%), uncertain isolates included

†: SPAdes assembly, ABRicate cut-offs (PID=90%, COV=90%), uncertain isolates excluded

O: No isolates with this phenotype to calculate VME/ME rates.

4.3.4. Evaluating associations between genetic elements and phenotype

Looking at associations between beta-lactamase presence and phenotype directly, many of the rules from the prediction algorithm appeared appropriate. However there were some important exceptions (Figure 28) where either the rules appeared wrong, or where the beta-lactamase was found in both susceptible and resistant isolates.

In agreement were that genes of class 2b enzymes (*bla_{SHV-1}*, *bla_{TEM-1}*) were associated with broad spectrum penicillin resistance, genes of class 2be enzymes (*bla_{CTX-M}* genes, *bla_{SHV-12}*) were additionally associated with 2nd and 3rd generation cephalosporin resistance and genes of class 1 enzymes (*bla_{CMY-2}*) were associated with penicillin, inhibitor and cephalosporin resistance. Of note, some of these enzymes (e.g. *bla_{CTX-M-15}*) appeared more resistant on crude analysis (Figure 7a) due to their co-occurrence with other beta-lactamases (e.g. *bla_{OXA-1}*).

The clearly incorrect associations mainly occurred due to the assumption that resistance would be constant across antibiotic subclasses. While all isolates with genes of class 2br inhibitor resistant enzymes (*bla_{TEM-30}*, *bla_{TEM-40}*) were resistant to amoxicillin-clavulanate, none were resistant to piperacillin-tazobactam. Likewise, the same was seen for both some carbapenemases and ESBLs. For example, isolates with *bla_{OXA-48}* were invariably resistant to ertapenem, but only 2/7 were resistant to meropenem (for the remaining five, one having an intermediate MIC (8mg/L) and the four having sensitive MICs (≤ 0.5 mg/L)). Likewise, cephalosporin resistance of isolates with *bla_{CTX-M}* genes depended on the enzyme present.

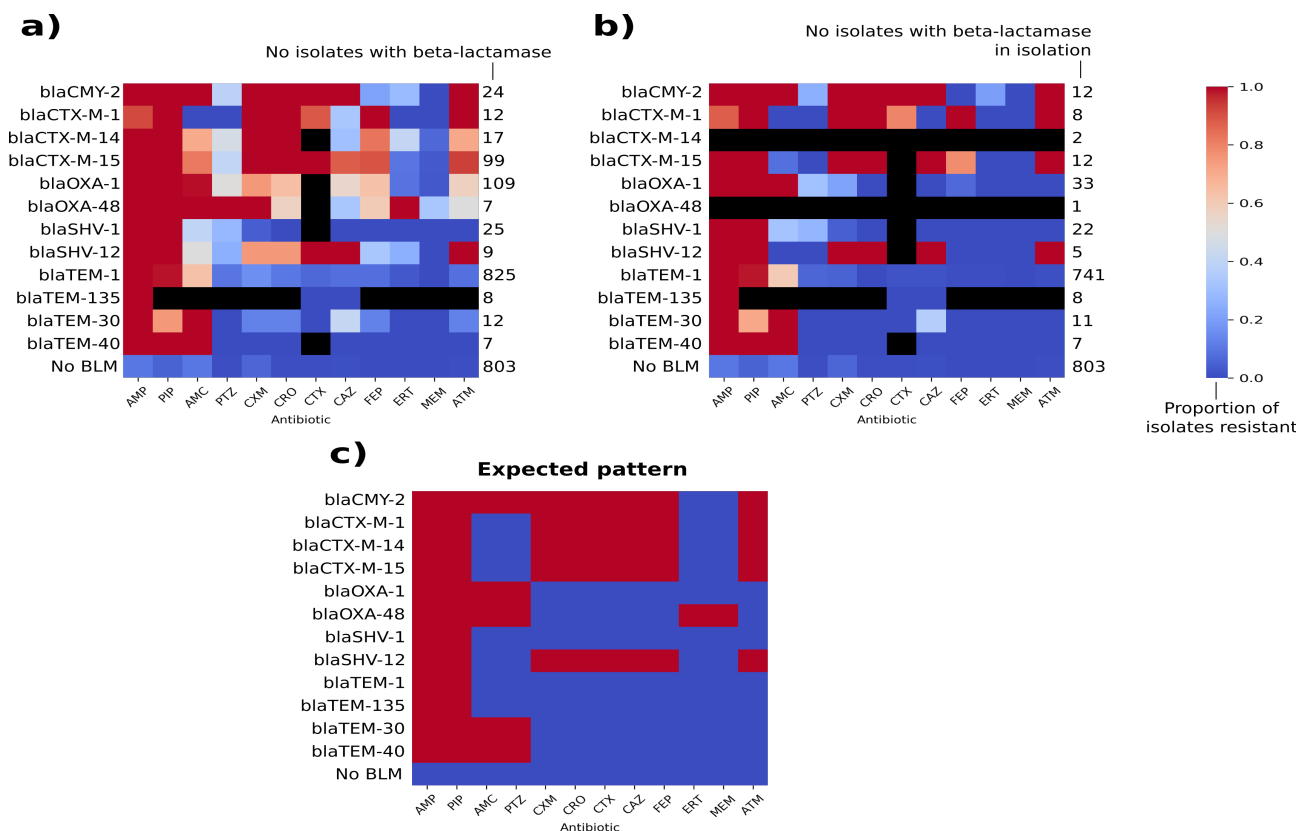


Figure 28: Association between beta-lactamases and phenotypes

Note: a) for each beta-lactamase, proportion of isolates containing this gene that are resistant from blue (0%) to red (100%). Total number of isolates to the right of the graph. Black denotes “missing”, i.e. when there were less than 5 phenotypes to base proportions on. Naming as previously follows NCBI convention, AMP - ampicillin/amoxicillin, PIP- piperacillin, AMC - amoxicillin-clavulanate, PTZ – piperacillin-tazobactam, CXM – cefuroxime, CRO -ceftriaxone, CTX – cefotaxime, CAZ – ceftazidime, FEP – cefepime, ERT – ertapenem, MEM- meropenem, ATM – aztreonam. b) similar to a) but only evaluating isolates containing a single beta-lactamase based on WGS predictions. c) the pattern expected if the category based resistance prediction was correct in all instances. See Figure 22

It is also clear that for certain enzymes and antibiotics, association with resistance or susceptibility was not invariable (e.g. *bla*_{TEM-1} and amoxicillin-clavulanate resistance).

308/510 (60%) isolates only containing a *bla*_{TEM-1} gene were amoxicillin-clavulanate resistant when tested. Only 20/308 (7%) of these isolates were suspected phenotypically as AmpC producers, which, if the class-based resistance prediction algorithm is correct, would suggest amoxicillin-clavulanate resistance was unexplained for 288/308 (93%) isolates.

Lastly, 98 of 803 isolates with no horizontally acquired beta-lactamases had resistance to at least one beta-lactam tested. 80/803 (10.0%) tested with ampicillin were resistant and 58/574 (10.1%) tested with amoxicillin-clavulanate were resistant. Some of this however may be explained by the chromosomal production of ampC, with 25/576 (4.3%) tested isolates having ceftiofur MIC >8 mg/L. 22/25 (88%) and 23/25 (92%) of these were ampicillin and amoxicillin-clavulanate resistant respectively.

4.4. Discussion

In this large dataset, in spite of applying similar approaches to previous highly successful studies^{120,141}, WGS-based beta-lactam resistance prediction was suboptimal, with discrepancies being particularly severe for some drugs. For drugs with only minor discrepancy, accounting for commonly hypothesised issues including inaccurate assembly, poor database curation and omission of AmpC hyperproduction from the prediction algorithm appeared to resolve most discrepancies. However the same was not true for drugs where resistance prediction was poorest. Instead, for these drugs,

the primary issue appeared to be the over-simplistic nature of the class-based resistance prediction approach. In particular, it was apparent that this approach fails account for either within-class phenotype differences (e.g. the differences between ceftazidime and ceftriaxone resistance), or for looser associations between gene presence/absence and susceptibility (e.g. between *bla*_{TEM-1} presence/absence and amoxicillin-clavulanate resistance).

To describe this in more detail, it is first helpful to observe that, for each antibiotic, performance of WGS-based resistance prediction in *E. coli* could be described in one of two ways. For penicillins, 3rd and 4th generation cephalosporins and aztreonam, the initial prediction was in reasonable agreement with phenotype, and, post clean-up of genotyping errors and exclusion of unknown/ampC producing isolates, VME/ME rates almost met the requirements of the FDA/ISO for regulatory approval. For BL/BLIs, cefuroxime and meropenem, performance was far below FDA/ISO standards even once the “optimal” genotype was produced. Each pattern has important implications for WGS-based resistance prediction.

For antibiotics with close to required performance, this work highlighted the importance, and to some extent the difficulty, of obtaining a good quality resistance genotype. Issues seen in this work can be broadly separated into two types, those which are artefactual and those which arise from underlying uncertainty in either the sequence or its interpretation. The artefactual errors primarily resulted from mismatches of the resistance database with its intended purpose. The inclusion of poorly curated elements and only one DNA-variant for each beta-lactamase may be sensible for cataloguing, but when used by a general search

strategy (such as BLAST-ing for all genes in the database) can lead to errors. Solving these problems for the routine use of WGS-based resistance prediction is not however straightforward, due to the fact that correction procedures are difficult to scale, and the fast rate of database change and unpredictable nature of errors making constant re-evaluation of performance essential ⁴⁷. In addition to the artefactual errors, there were several cases where either incomplete or novel genetic mechanisms led to uncertainty in the WGS-based resistance prediction. While only occurring in 25/1818 isolates (1.3%), the level of error these issues can introduce can be important given FDA/ISO standards required for regulatory approval. How to manage these issues is not yet clear, with ongoing research into method of resistance prediction for novel genes. Before assuming this is only a weakness of the WGS-based approach, it is important to stress that similar uncertainty can occur with phenotyping, which is mitigated in clinical diagnostic laboratories by standardised investigation pathways.

For antibiotics with poor performance, a more fundamental examination of the WGS-based resistance prediction method was required to identify likely causes of error (Figure 29). There are multiple ways in which these two processes could come to different conclusions, such as if plasmids are lost in subculture or sequencing, or if resistance sequences are accidentally filtered in bioinformatic processing. The fact however that resistance prediction performs well for some antibiotics argues against many of these hypothesis. Ampicillin is important for this argument, as its range of activity is contained within the range of activity of all other beta-lactams in this study. Consequently, it can be approximately viewed as detecting “any” important beta-lactam resistance. Hence, assuming that the ampicillin phenotype is correct, predicting ampicillin resistance well suggests that the WGS method

performs well at identifying beta-lactam resistance in general, but is unable to correctly identify specific resistance. This could occur for several reasons, for example;

- Specific phenotyping issues (e.g. for amoxicillin-clavulanate) causing error
- Inadequate algorithm to predict resistance from WGS-based data
- WGS failing to correctly identify beta-lactamases to allow for resistance prediction
- Novel mechanisms not searched for from WGS data

While all of these could be possible causes, problems in the resistance prediction algorithm are evident in data.

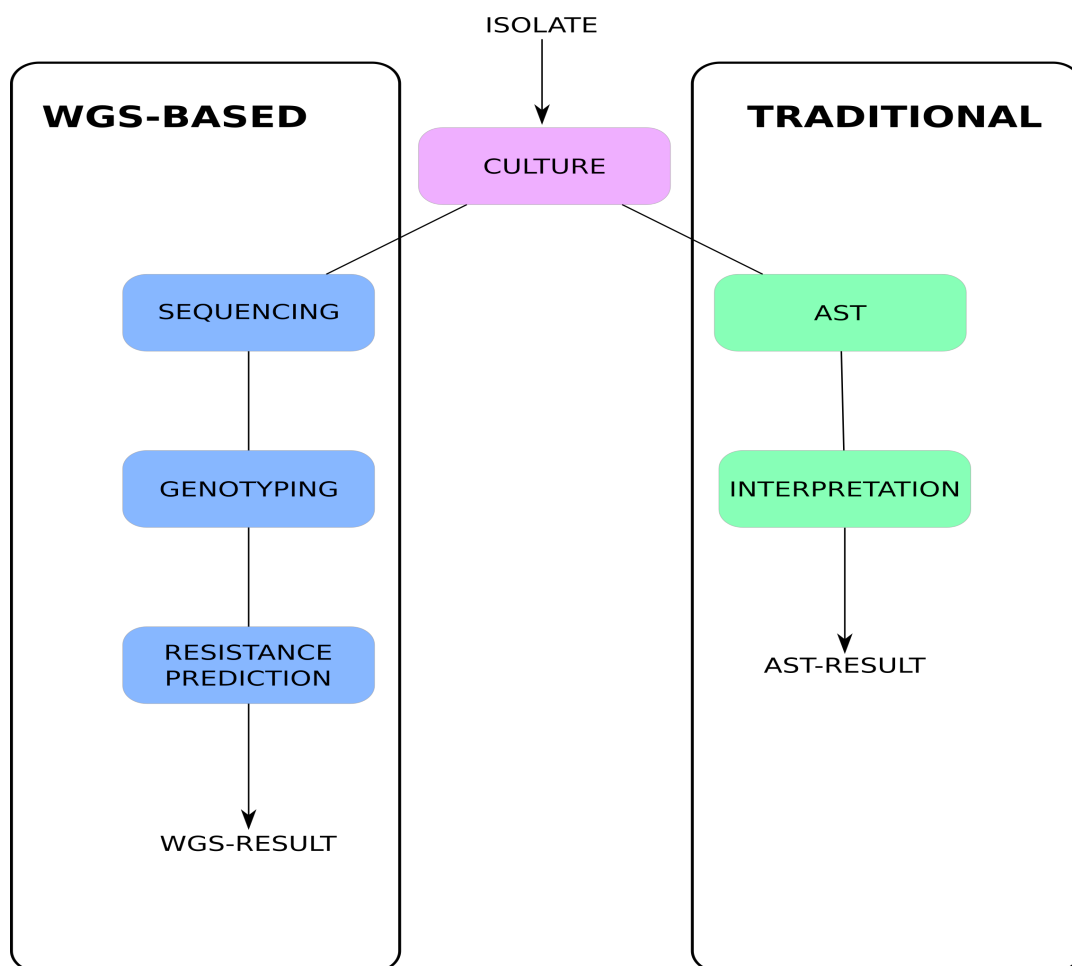


Figure 29: Processing required for WGS-based methods versus traditional AST

Firstly, rules in the resistance prediction algorithm assume that the full beta-lactam phenotype can be predicted using beta-lactamase functional classification alone. This was not the case, with 486/1818 (26.7%) isolates having antibiograms that would be impossible to predict (i.e. there was no possible combination of resistance prediction rules across the different enzyme classes which would match the observed phenotype). A reason for this likely lies in this algorithm mis-interpreting what the functional class of a beta-lactamase signifies. A beta-lactamase's functional class is determined by its substrate profile for a limited set of substrates, measured using purified enzyme. For example, label's such as "2br" or "inhibitor resistant" are specified by the enzymes ability to resist inhibition by clavulanate. However, as was seen from the associations of beta-lactamases with phenotypes (Figure 28), isolates containing these beta-lactamases were inhibited by tazobactam, an alternative beta-lactamase inhibitor. Instead of being indicative of behaviour against each beta-lactam, classes are a simplification of the far more complex interplay between beta-lactamases and beta-lactams, in which each antibiotic and enzyme (or inhibitor) have their own affinities for one another. However, while future work and increased numbers of available isolates may make predictions based on specific genes possible, at the time of writing it is not certain this will be the case for Enterobacteriaceae. Studies such as the CRYPTIC TB project have aimed to do this via linking genetic data with phenotype for thousands of isolates ²⁹; however, with *E. coli* being one of the commonest commensal bacteria and some resistance mechanisms being very rare, exactly how many isolates would be needed to obtain good prediction for *E. coli* is unclear. Further the prediction for individual genes only fails to solve part of the issue, with this work highlighting a second major issue which would even affect gene-based resistance prediction.

The second piece of evidence that the resistance prediction algorithm may be incorrect was the apparent weak association between some beta-lactamases and the phenotypes of the isolates harbouring them. The rule-based WGS-prediction algorithm used in this chapter, similar to those published prior to this thesis, relies on the assumption that each beta-lactamase produces a definitive “R” or “S” phenotype. For some beta-lactamases and antibiotics, such as *bla*_{TEM-1}/*bla*_{SHV-1} and amoxicillin-clavulanate, this was clearly not the case (Figure 28). Again, it would be somewhat less likely (although possible) for this error to be due to un-identified/mis-identified beta-lactamases, given the generally stronger associations between gene presence and amoxicillin-clavulanate susceptibility category for some enzymes. If there were undetected genes, one would expect this to affect associations for other beta-lactamase genes, however most are still strongly associated with either resistance or susceptibility (e.g. *bla*_{CTX-M}). One caveat to this is the potential that isolates contain multiple highly related alleles which cannot be separated on assembly. This is investigated in further detail for this data in Chapter 5 using alternate resistance genotypes which work directly on reads as opposed to assembly.

While these errors taken together are highly suggestive of a resistance prediction algorithm problem, it is important not to discount the possibility of phenotyping errors. These errors could cause antibiotic and even potentially mechanism specific discrepancies (e.g. if particular mechanisms act unpredictably in a given assay). However, with the long established methods of AST quality control and its generally accepted use in clinical practice, this should not be viewed as the most likely conclusion prior to ruling out problems with the WGS approach.

There are several possible reasons why WGS-based beta-lactam resistance prediction was less successful than previous studies ^{141,158}, highlighting the difference between initial proof-of-principle studies and larger studies of feasibility. The first and most obvious difference between this work and previous studies was the sampling frame itself, with prior studies all having used smaller sets of pre-selected isolates. For example, Tyson et al used pre-selected MDR *E. coli* isolates from healthy cattle ¹²⁰, Stoesser et al used isolates containing ESBLs ¹⁴¹, and Shelburne et al used isolates obtained from intensive care patients with haematological malignancies ¹⁵⁸. These isolates likely contain more resistance mechanisms; although sequencing them can be useful for multiple epidemiological studies, they may be a poor proxy to test the applicability of WGS-based resistance prediction in routine clinical practice. As further evidence of this effect, performance of WGS-based prediction in this study was dataset specific, typically struggling more in the (less resistant) OUH as opposed to the (more resistant) PHE dataset (Figure 23). Interestingly however, ceftazidime prediction was worse in the APHA dataset (isolates obtained from pigs), which was selected for cephalosporin resistance. The cause for this is unclear, and could be due to differences in phenotyping approach (agar dilution vs BD Phoenix, see section 3.3.2.) or genetic differences between the human and animal isolates (see section 3.4.1) which were poorly accounted for in the WGS-based resistance prediction approach. The second major difference could be due to differences in the number of samples included. As I found in this study, careful inspection of the initial genotype to identify and correct artefactual errors was important. With smaller sample size these previous studies could individually analyse and manually correct artefactual errors, such as phenotyping issues or assembly breakages ¹⁴¹. Again, while this manual correction is technically possible, it is unrealistic to believe this could be done in a clinically applicable time scale for the large numbers of samples seen in a

routine diagnostic setting. Hopefully with improvements methods to identify resistance genotypes, this correction will become less essential over time; however, in the interim it is important to recognise that manually correcting the genotype obscures the realistic performance of WGS-based resistance prediction.

The next three chapters investigate possible causes of WGS-based resistance prediction errors, including bioinformatic method, resistance prediction algorithm and phenotyping. The next chapter presents some more recent work on the effect of bioinformatic method choice on resistance genotype. While not performed in this chronological order in the doctorate (with this working being performed most recently), it provides a useful baseline to understand some of the complexities of searching for resistance genes from short-read WGS data. Then, the two following chapters focus on issues with resistance prediction algorithms and phenotyping for amoxicillin-clavulanate resistance prediction in detail. The reason for choosing this antibiotic was not only its status as one of the most commonly used broad-spectrum antibiotics worldwide ²⁰¹, but also because the relatively high prevalence of resistance and WGS/phenotype discrepancies in my datasets made it an ideal test case, enabling discrepancies to be analysed on a much larger scale than for other antibiotics.

Chapter 5: Pitfalls of identifying the resistance genotype from WGS data

5.1. Introduction

While unlikely to be the greatest cause of discrepancies for amoxicillin-clavulanate resistance, the bioinformatic method used to determine the resistance genotype from WGS data is a fundamental component of “rule-based” and “database-guided” resistance prediction. Currently, there is no consensus as to the “best” choice. This chapter investigates the impact of bioinformatic method on how resistance genotype is determined from short-read sequencing data.

This chapter compares four different bioinformatic pipelines and is broken into three distinct sections.

- 1) An initial analysis where beta-lactam resistance genotypes are obtained using four different bioinformatic methods for the 1818 isolates tested in Chapter 4. These results were used to identify and analyse the extent and causes of discrepancy between methods.
- 2) While some discrepant had concrete explanations, the cause of others was less clear. Simulations were used to test whether hypothesized causes were plausible.
- 3) Finally, the impact of bioinformatic method choice on resistance prediction (as opposed to just genotyping) is examined. In addition this chapter investigates if

resistance prediction can be improved by combining results from several different pipelines.

5.1.1. Background

Published literature comparing the bioinformatic methods used for resistance genotyping and prediction is limited, with most being done as part of the initial publication of each novel method ^{137–139}. As the comparison between bioinformatics methods is not the primary focus of these papers, the comparisons themselves are typically of poor quality and only focus on phenotypic prediction as opposed to accuracy of genotype. Further they generally only compare accuracy of prediction against a limited set of antibiotics in small numbers of highly selected isolates. These comparisons are at significant risk of bias given the pressure on authors to show superiority of their novel method. To my knowledge, as of July 2019 only one study has attempted to approach this comparison more rigorously ¹⁴⁰; however this only compared bioinformatic methods to identify resistance in *S. aureus*, avoiding many of the complexities in Gram-negative organisms.

5.2. Methods

5.2.1. Comparing resistance genotypes in real data

Selecting bioinformatic methods for comparison

For this study, bioinformatic methods for AMR prediction were chosen for comparison to cover range different of approaches (Table 14). In addition, all included methods had to meet the following inclusion criteria at the outset of the study (March 2018):

- Its program code is publicly available

- It runs on local computing architecture without the need for major modification
- It accepts alternate databases to ensure its continued applicability over time
- It has a command line interface for processing large numbers of samples

The four methods selected were ABRicate (which uses BLAST search of *de-novo* assemblies, as used in Chapter 4), SRST2 (a mapping-based algorithm), ARIBA (used in Chapters 6 and 7, a combined mapping/assembly approach) and KmerResistance (based on k-mer sharing).

Table 14: Commonly used methods for searching WGS data for resistance sequences

Name	Method	Code Public	Published (date of publication)	Flexible database	Interface	Usable March 2018*
ABRicate⁺	Assembly + BLASTn	Y	N	Y	command line	Y
AMRfinder	Assembly + BLASTp/tBLASTx + Hidden Markov Models	Y	N	N	command line	N
ARIBA	Mapping + local assembly	Y	Y (2017) ¹³⁸	Y	command line	Y
ARG-ANNOT	Assembly + Blast	N	Y (2014) ²⁰²	N	web	N
Kmer Resistance	Kmer sharing	Y	Y (2016) ¹³⁹	Y	command line	Y
Mykrobe Genotype	DeBruijn Graph	Y	Y (2015) ¹²¹	Y	command line	Y ^o
Resfinder + Pointfinder	Assembly + BLASTn	Y	Y (2012) ²⁸	N	web +command line	N
RGI (CARD)	Assembly	Y	Y (2017) ¹³³	N	web +command line	N
SRST2	Mapping	Y	Y (2014) ¹³⁷	Y	command line	Y

+: ABRicate was used in for resistance prediction in Chapter 4.

*: Non-usability causes (in March 2018): AMR finder - program not released, ARG-ANNOT – code not publicly available/web interface only, Resfinder/Pointfinder – code not publicly available/web interface only, RGI - code not publicly available/web interface only.

o: Mykrobe genotype was a secondary function of its main program Mykrobe predictor (for AMR prediction for *M. tuberculosis* and *S. aureus*), reducing its functionality resistance prediction. After failing to run this on local architecture despite advice of its creators, this method was removed from analysis.

Bioinformatic methods used in this chapter

ABRicate (<https://github.com/tseemann/abricate>, unpublished)

ABRicate, the baseline method used for resistance genotyping (see Chapter 3, section 3.3.1), uses BLASTn¹³⁵ to identify resistance/virulence genes from pre-assembled sequences. Given the high similarity of many resistance genes, several genes often match to the same contig location. ABRicate uses BLASTn's culling limit option to select a single optimal matching sequence when this occurs. To attempt to mitigate issues caused by assembly fragmentation, ABRicate's default percentage identity and width of coverage cut-offs are 75% and 0% respectively, representing more permissive default settings than either Resfinder or those used by Stoesser et al¹⁴¹.

SRST2 (<https://github.com/katholt/srst2>, Inouye et al¹³⁷)

SRST2 identifies resistance genes directly from raw reads by mapping reads to a resistance database using Bowtie2¹³⁶ (Figure 30). To manage the multiple matching issue, SRST2 uses CD-HIT-est²⁰³ to cluster database elements into groups, and calculates a divergence statistic for each member of the group (reflecting the probability that observed sequence does not match the reference given a 1% per base sequencing error) and selects the allele which minimises this statistic¹³⁷. Best matching alleles are then reported if, and only if, their divergence (difference from the reference allele) is less than 10%, the width of coverage of the identified gene is greater than 90%, and the average mapping read depth across the allele is greater than 5x.

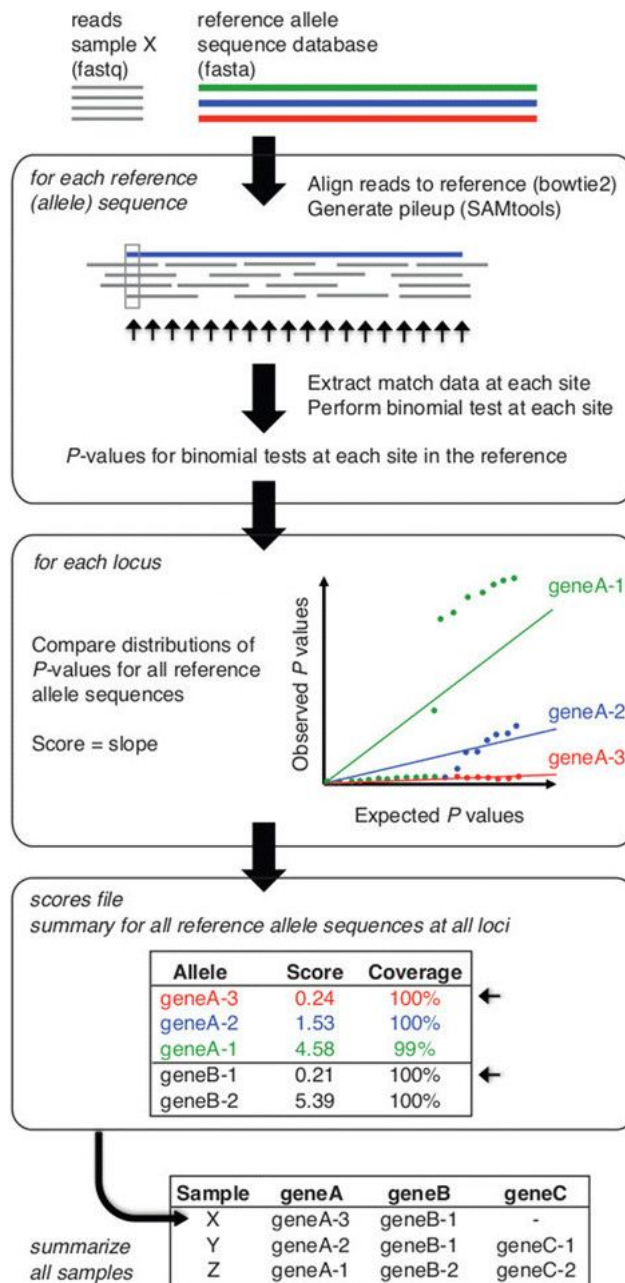


Figure 30: Overview of SRST2 approach

Note: reproduced from Inouye et al¹³⁷

ARIBA (<https://github.com/sanger-pathogens/riba>, Hunt et al¹³⁸)

In an attempt to address the shortcomings of both mapping and assembly approaches, ARIBA uses a combination of both mapping and assembly. This method was used in Chapters 6 and 7 (conducted prior to this work, see section 5.1) as an assumed

improvement over the assembly and blast approach used in the initial prediction in Chapter 4. Reads are initially mapped to resistance sequences using minimap²⁰⁴ and then assembled using fermi-lite (<https://github.com/lh3/fermi-lite>). Like SRST2, ARIBA pre-clusters resistance sequences using CD-HIT-est²⁰³ to deal with multiple matches. Best matching sequences from each cluster are then chosen by comparing assemblies with reference sequence using nucmer²⁰⁵. This again has its own set of default cut-offs (percentage identity = 90%, width of coverage = 20%), and a secondary cut-off requiring 90% of the gene to be assembled into a single contig which ARIBA applies during its “summarize” step.

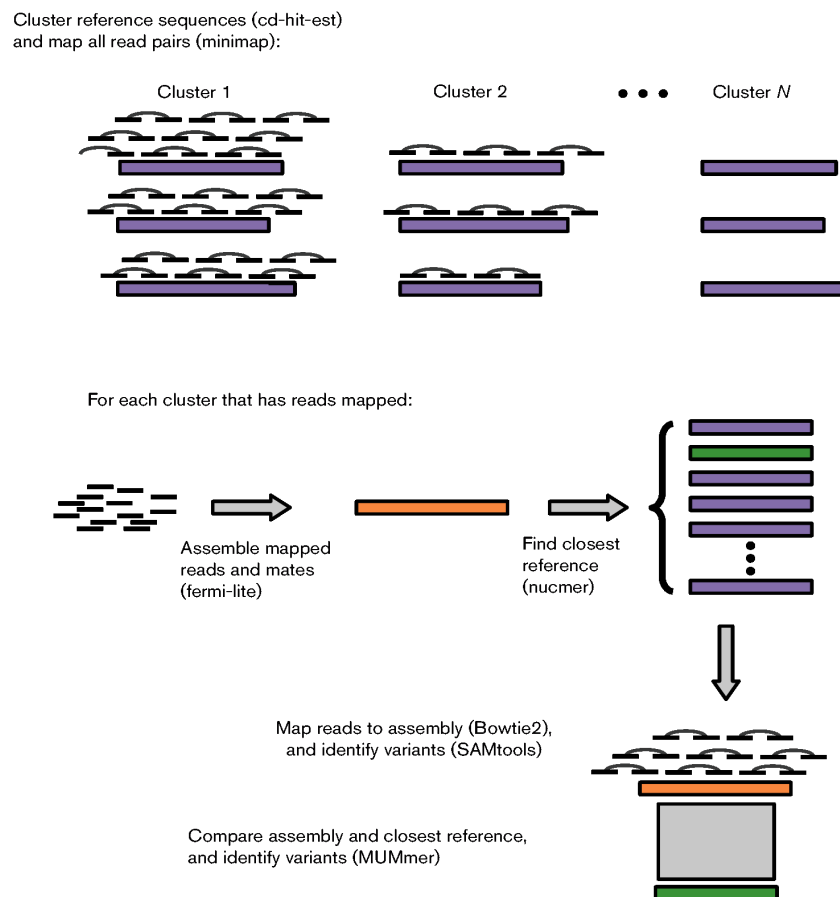


Figure 31: Overview of the ARIBA approach.

Note: reproduced from Hunt et al¹³⁸.

KmerResistance (<https://bitbucket.org/genomicepidemiology/kmerresistance>, Clausen et al

139)

KmerResistance uses an alternative strategy of analysing k-mer sharing between sample reads and the resistance database. For each element in the database, its unique k-mers (in the database) are identified, and used to search k-mers generated from the sample reads. Then all mapping sample read k-mers are iteratively assigned as belonging to elements of the database. This is done by using the “winner takes all strategy”¹³⁹ which, starting at the element with the highest number of matching unique k-mers in the sample read k-mers, assigns all k-mers which map this element and removes them from the sample read k-mer set. The process then moves onto the element with the next highest number of matching unique k-mers and continues until there are no more elements with matching k-mers. As this process relies on exact matching, mismatches and indels both result in reduced k-mer coverage, and hence this method only has one cut-off (identity threshold, default 70%) as opposed to the two (percentage identity and coverage) cut-offs used by all other programs. It also has a process by which it removes what it sees as likely contaminants by checking the relative depth of k-mer coverage of the element compared to a “reference” chromosome of the sample species, chosen by using the k-mer sharing approach between sample reads and complete bacterial genomes from the NCBI ref-seq database (<https://www.ncbi.nlm.nih.gov/refseq/>).

Generating and comparing genotypes

The first part of comparing the tools was to run isolate sequencing data through each method and compare reported resistance genotypes between methods. To encompass a wide range of different possible resistance mechanisms, each isolate’s sequencing data was

searched using the Resfinder database for transmissible resistance sequences related to any of five different classes of antibiotic: beta-lactams, aminoglycosides, quinolones, trimethoprim and sulfonamides. Likewise, genotypes obtained were broken down by “antibiotic class” (defined using Resfinder’s antibiotic specific FASTA files) to simplify analysis, i.e. each isolate’s overall genotype was broken down into its beta-lactam, aminoglycoside, quinolone, trimethoprim and sulfonamide genotypes. The antibiotic classes investigated were chosen as they contained many of the antibiotics most commonly used to treat clinical *E. coli* infections.

Each search aside from ABRicate was performed using default parameters, with ABRicate being run on SPAdes assemblies with percentage identity = 90% and coverage = 90% cut-offs (rather than the defaults of 75% and 0% respectively) to match the final algorithm in Chapter 4.

To simplify analysis, for each antibiotic class-specific genotype, discrepant results were classified according to which of the four methods agreed. For all discrepancies in beta-lactams genotype, the cause was investigated and the impact on predicted phenotype was assessed. To investigate if similar causes could apply to discrepancies in other antibiotic-class genotypes, the causes of discrepancy patterns occurring in more than 1.5% (N=27) isolates for each of the other antibiotic genotypes (aminoglycoside, quinolone, trimethoprim, sulfonamides) were also investigated. The 1.5% cut-off was chosen as, if it also led to a change in resistance prediction, this rate of discrepancy could be sufficient to prevent WGS-based resistance prediction from meeting FDA/ISO standards^{105,106} (section 2.2.4, Table 3).

5.2.2. Investigating discrepant using simulations

The challenge with isolate sequencing data is that the truth is generally unknown, and errors can only be inferred from patterns of concordance between the genotype predictions and phenotype, and understanding of program method. Some errors have relatively concrete explanations, and sequence data can be interrogated to almost definitively identify errors. In other situations however causes are less transparent, as was the case for two types of discrepancy seen in the real data (see section 5.3.2.). The simulations investigated whether hypothesized causes of these two discrepancies could feasibly result in observed patterns of discordance. For each of the two discrepancies patterns, a scenario was constructed to mimic/demonstrate the hypothesized cause (Figure 32).

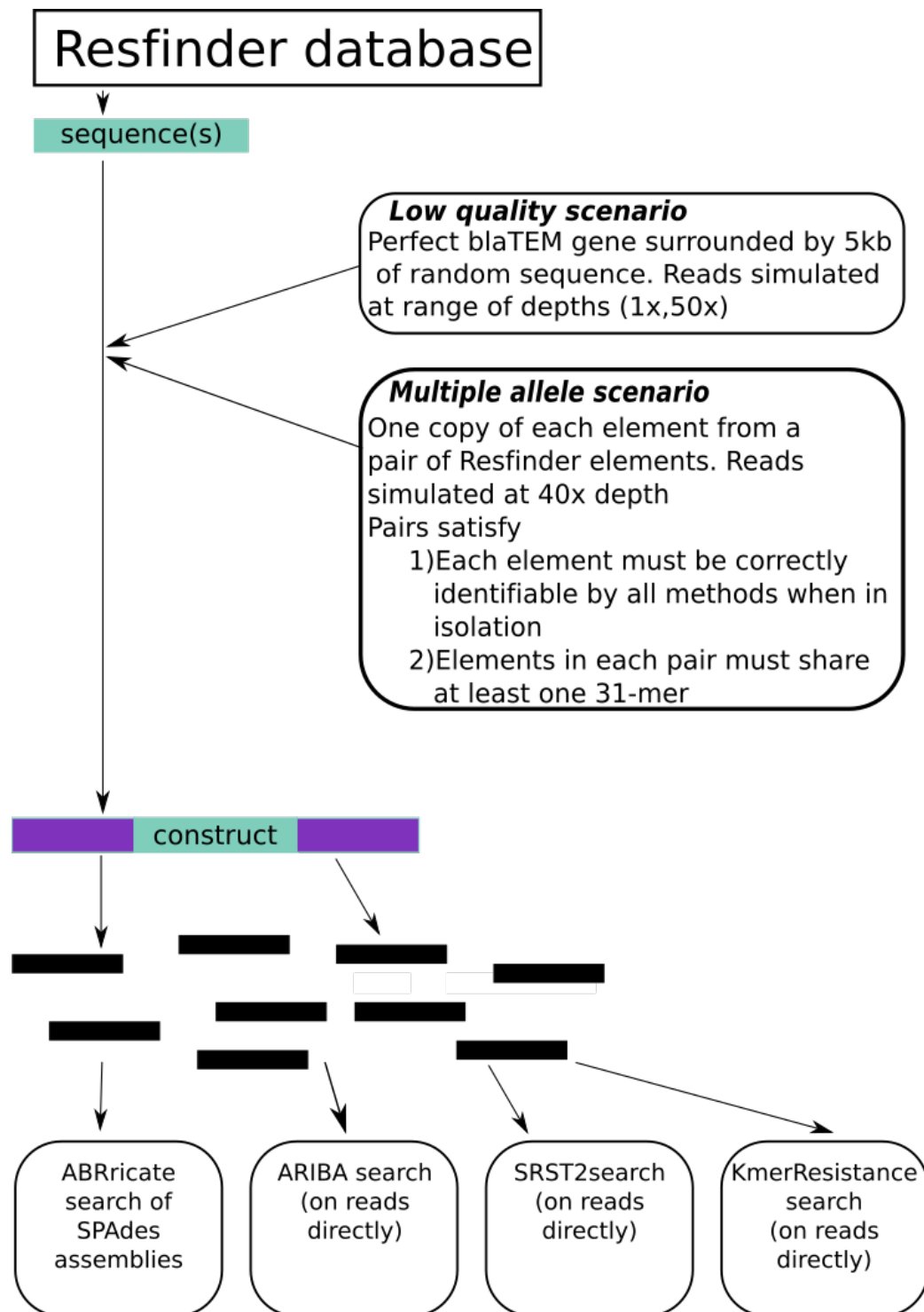


Figure 32: Simulation plan

Note: All programs were generally applied using defaults including database preparation as recommended by program manuals. Exceptions include disabling coverage cut-offs for the low quality scenario, and 5x coverage cut-off for KmerResistance, because by their nature, simulations would disable its inherent coverage cut-off (based on mapping depth to a reference genome).

Low quality scenario

The primary aim of this scenario was to test how each method performed when presented with low quality sequence data, simulated by using low coverage reads. To do this, reads (with error) were simulated from 176 (the number of coding *bla*_{TEM} genes in the Resfinder database) *bla*_{TEM} gene-containing constructs at coverage depths ranging from 1x to 50x using ART, a program designed to simulate short-read sequencing reads *in silico*²⁰⁶ (default illumina HiSeq2500 error profile, total 176*50=8800 simulations). Each construct contained a random perfect *bla*_{TEM} gene flanked by 1kb of random sequence on each side produced using random selection from Python's Numpy module and combined using BioPython^{185,187}. Simulated reads were then processed by each method (Figure 32) using their default settings to create method-specific resistance genotypes and the identified alleles were compared with the actual *bla*_{TEM} genes in each construct. The measure of performance for this scenario was simply the proportion of *bla*_{TEM} genes correctly identified by each method at each coverage level.

As part of the planning for this scenario, the performance of ABRicate at detecting low coverage *bla*_{TEM} alleles in 500 simulated constructs (with coverage ranging from 1x to 50x, i.e. 10 constructs for each level of coverage) was assessed using several different assembly methods. The choice of N=500 constructs was arbitrary, as this analysis was primarily used as a rough guide as to choice of assembler, as opposed to being powered to detect any specific differences. Performance was assessed using different approaches to assembly, namely;

- VelvetOptimiser (<https://github.com/tseemann/VelvetOptimiser>) with k-mer range

31-151

- Velvet ¹⁶⁹ using a single kmer length(k=31, k=51, k=71)
- Shovill (a SPAdes wrapper script, <https://github.com/tseemann/shovill>)
- SPAdes ¹⁹⁸

In this I found that performance was improved by using VelvetOptimiser versus picking a single k-mer alone (Figure 33). Further, using VelvetOptimiser assemblies, correct *bla*_{TEM} alleles could be identified at lower coverage, although there were fewer mistakes when using SPAdes assemblies for simulations with coverage depth 10-14x. Ultimately I chose SPAdes as the assembler for simulations as it provided a balance between accuracy at low coverage and time taken for assembly (using a single core on a 40x depth simulation, VelvetOptimiser took 120-180seconds, Shovill 8-9seconds, and SPAdes 4-5seconds), plus the fact that it matched the approach used in real isolates.

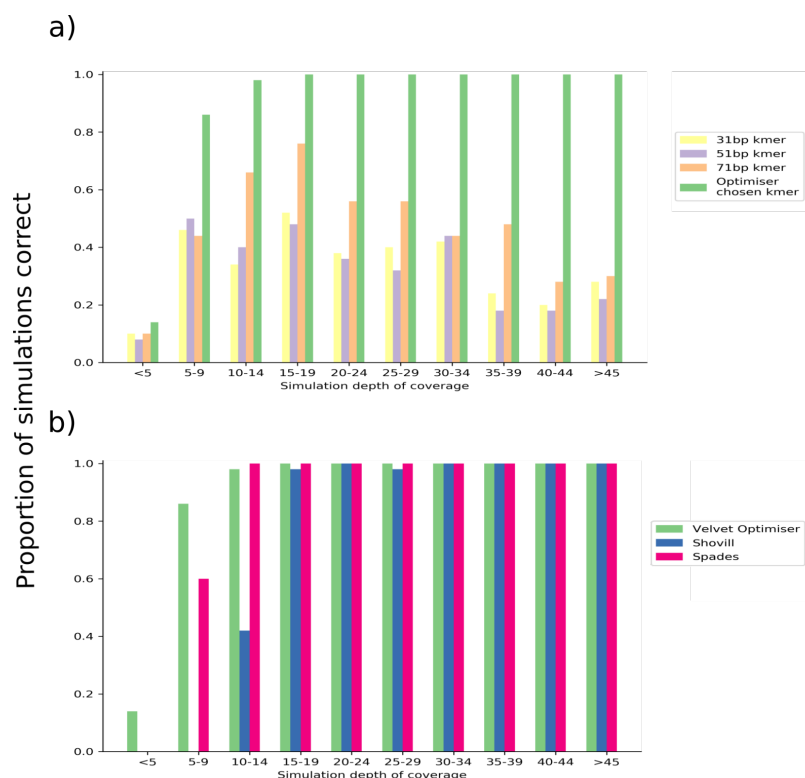


Figure 33: Choosing assembler for simulations using ABRicate

Note: a) Performance with single-kmer velvet assemblies versus VelvetOptimiser b) Performance with VelvetOptimiser versus SPAdes and Shovill

Multiple allele scenario

This scenario aimed to test how successfully algorithms identified isolates containing multiple versions of similar gene sequences. Pairs of similar gene sequences (see below for more detail) were selected and each member of the pair was inserted into its own contig containing the resistance sequence surrounded by 1kb of random flanking (again generated using NumPy and Biopython). These contigs were then placed into the same FASTA file for read simulation (at a ratio 1:1) which was then used to generate reads at 40x coverage using ART (again using it's default illumina HiSeq2500 error profile). This was done for all sufficiently similar pairs (see below), resulting in N=46279 simulations. To assess method performance, I used the following scoring system:

- 5 = Method identifies both sequences correctly and no other additional incorrect sequences
- 4 = Method identified both sequences correctly but with additional incorrect sequences
- 3 = Method identifies only 1 sequence correctly, and no additional incorrect sequences
- 2 = Method identifies only 1 sequence correctly with additional incorrect sequences
- 1 = Method identifies 0 sequences correctly

Scores were then used both to compare methods, and also assess the impact of pair similarity on whether each member of the pair could be identified.

As part of planning this scenario, I had to identify “similar pairs” of sequences and have a measure of similarity between them. Further, for it to be a reasonable test of each program

to identify the pair, all four programs had to be able to identify each element of the pair correctly in isolation, else mistakes could be due to incorrect identification of individual elements as opposed to problems relating to distinguishing highly similar sequences.

The first step to creating a set of similar pairs was to exclude elements which would be incorrectly identified even when they occurred as the only sequence in a construct. To do this, each element from the Resfinder database was inserted into 1kb of random flanking sequence either side of the insertion and reads simulated at 40x coverage using ART. Each method was then tested to see if it could correctly identify the element (3092 simulations, reflecting the number of elements in the Resfinder database). All methods were correct on 2011/3092 (63.5%) genes in the Resfinder database, with KmerResistance being correct on all 3092 simulations, ABRicate on 3082 (99.7%), ARIBA on 2927 (94.7%) and SRST2 on 2120 (68.6%). Most errors were due to a problem relating to how SRST2 clusters sequences by identity to identify the “best matching allele” for a given sequence. This error is explained in more detail in the results below as it is also seen in the real world isolates.

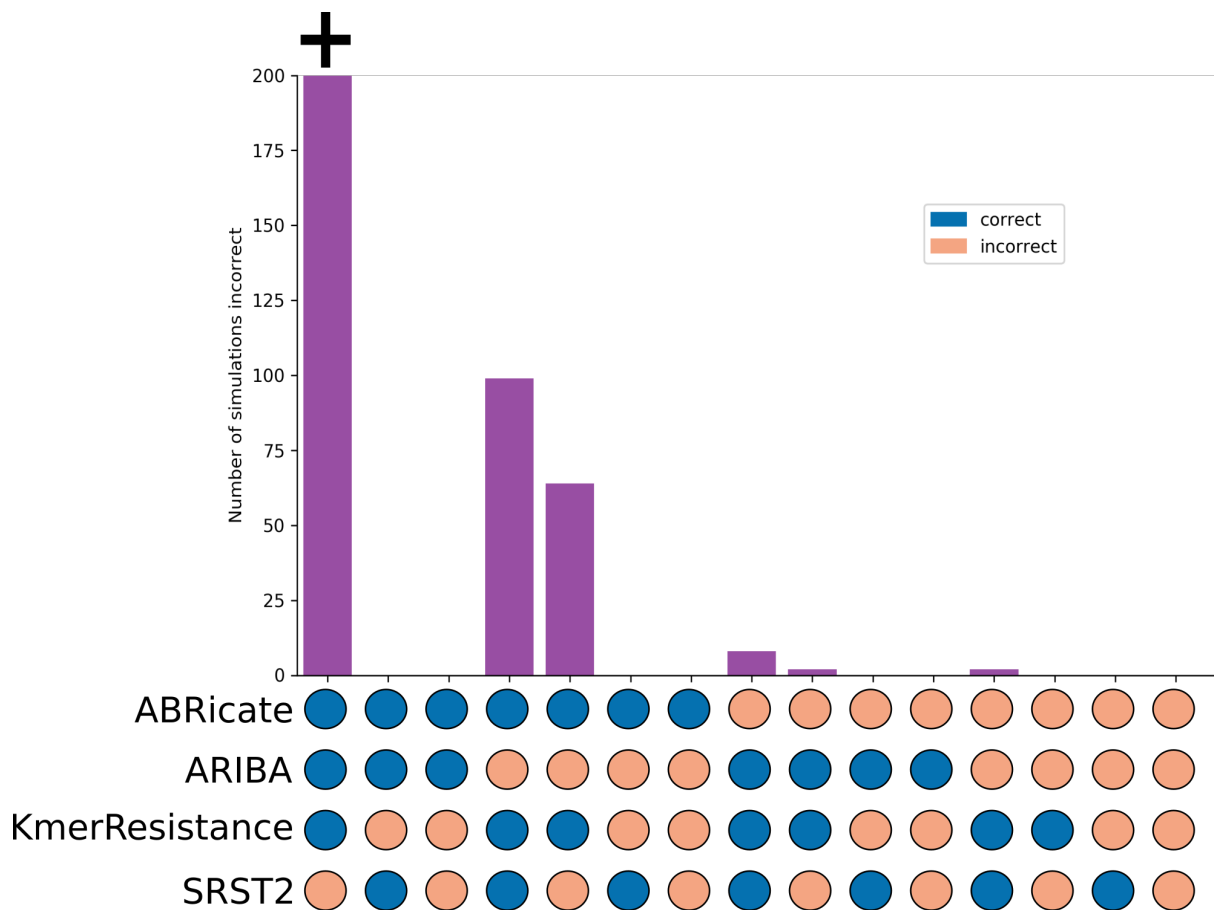


Figure 34: Identifying elements from the Resfinder database in isolation

Note: +, total SRST2 only errors 906, but truncated to 200 to make other errors visible. blue = method correct for these simulations, orange = method incorrect. Only showing simulations where one or more methods was incorrect.

Once elements which would be incorrectly identified in isolation were excluded, pairwise “similarity” between all remaining elements was calculated. To do this, each remaining element (i.e. from the set of all elements that could be correctly identified by all four tools when present in isolation) was split into 31-mers (length 31 k-mers) which were then compared with 31-mer sets from every other non-excluded element using pairwise Jaccard’s similarity indices.

$$\text{Jaccard's similarity index} = \frac{\text{No shared 31 - mers}}{\text{Total 31 - mers}}$$

The choice of 31-mers to identify “similar” resistance genes for each construct was based on preliminary analyses. In this, pairwise Jaccard similarity indices were calculated for four possible lengths of k-mer (11, 17, 21 and 31, the defaults as used by sourmash ²⁰⁷) and compared (Figure 35). Jaccard’s similarity as estimated by 31-mers was highly correlated with that estimated from the other k-mer sizes tested.

Finally, pairs were defined as similar if they shared any 31-mer. This definition of “similar” was chosen for two reasons. Firstly, as this was an exploratory analysis, I was keen to see performance on any pairs which could be defined similar. This approach fitted this mandate as it was able to link even the most loosely associated elements. Secondly, as all of the algorithms rely on k-mers at some point in their approach (including both assembly and mapping based approaches), distinguishing between elements with any similar kmers is likely a problem faced by them all. Again this approach fitted this criterion, as by definition it identified any pairs with similar k-mers. There were a total of 46279 possible pairings which satisfied this criterion, with the distribution of all Jaccard similarity indices as shown in Figure 36a. Figure 36b then demonstrates how this metric compares to two other pairwise similarity metrics, percentage identity of the Needleman Wunsch global alignment ²⁰⁸ and percentage identity of best “high scoring pair” from BLASTn alignment of the two sequences.

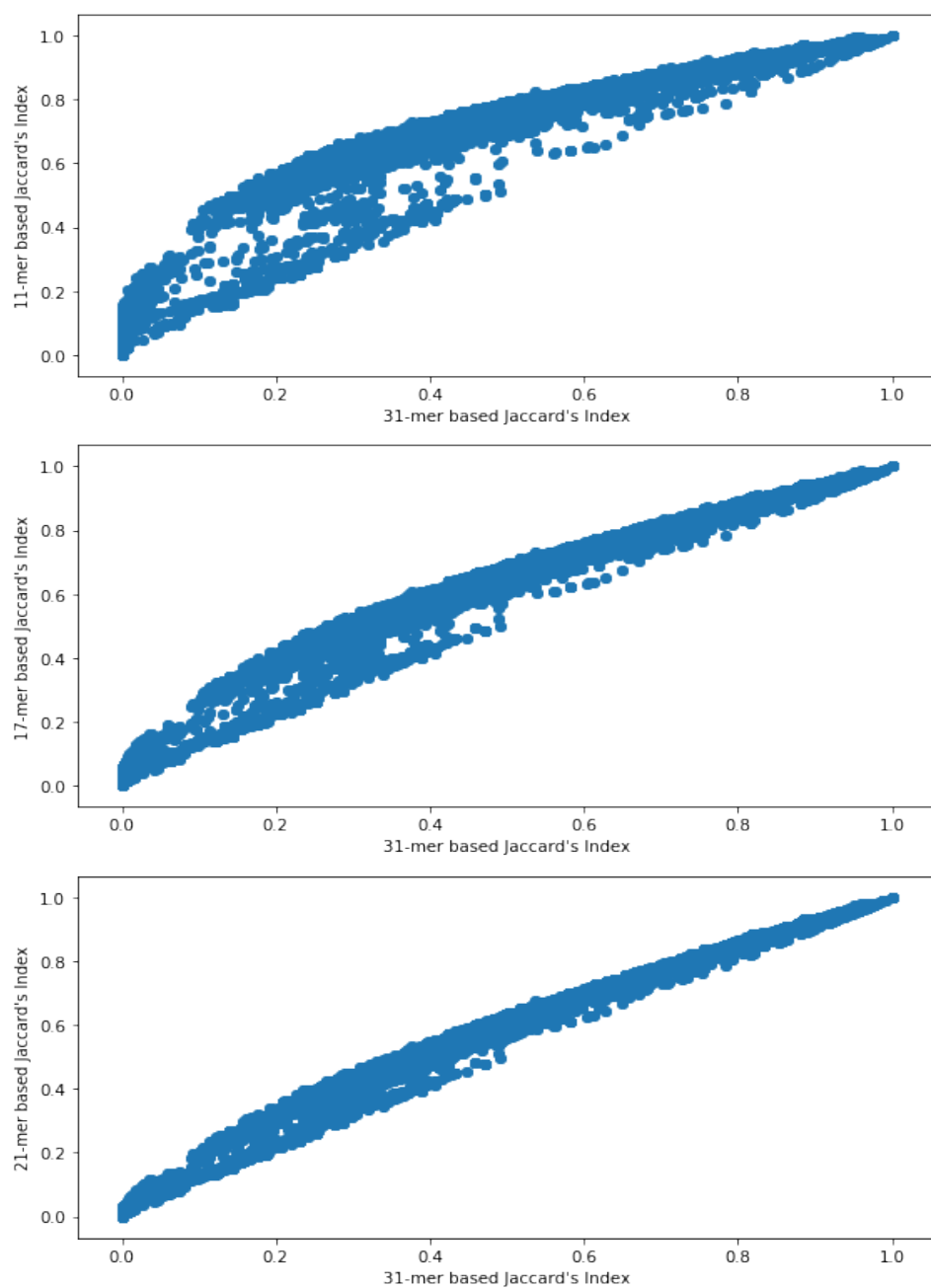
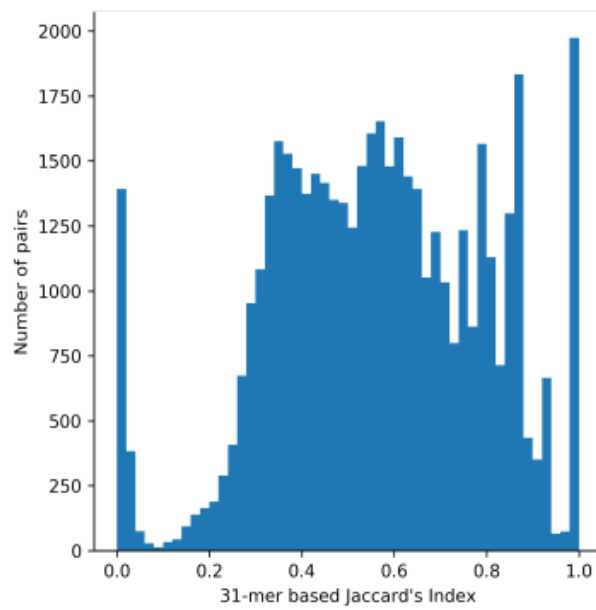


Figure 35: Association between Jaccard index from 31-mers vs A) 11-mers B) 17-mers and C) 21-mers

Note: all vs all comparisons for 2011 included sequences.

a)



b)

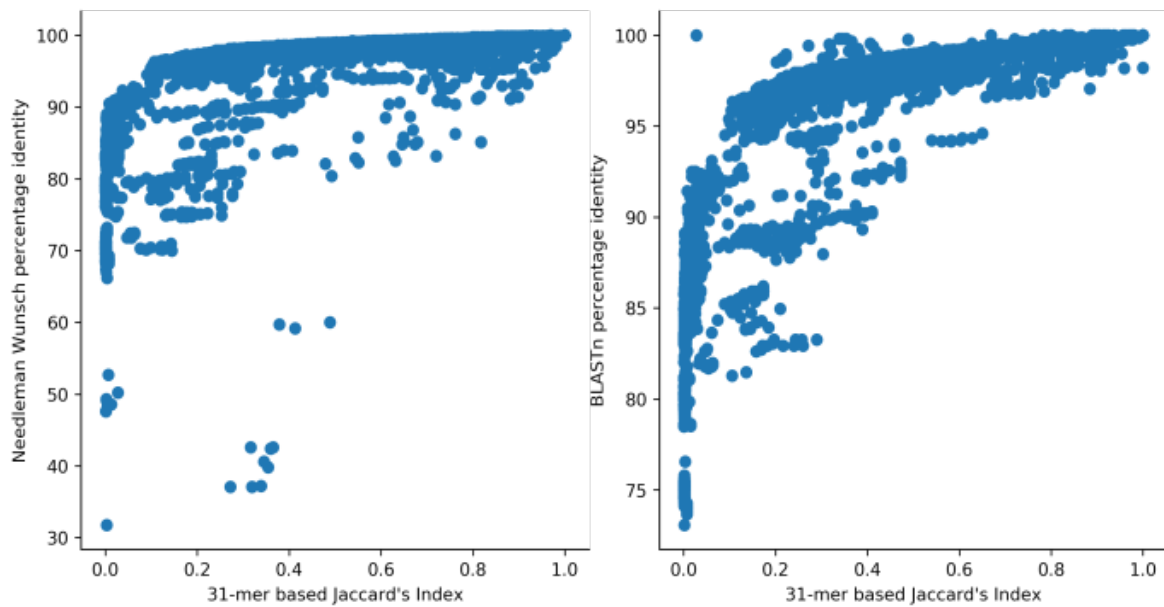


Figure 36: Comparing similarity metrics A) Jaccard index, overall and compared with B) Needleman Wunsch and C) BLASTn percentage identity

For all simulations from both scenarios, any non-depth parameters required for running ART were estimated from metrics derived from sequencing data from 100 randomly selected samples from the real world isolates (Table 15). The 40x depth of coverage used for the multiple allele tests was chosen as (i) it was above what was required for all methods to

identify alleles correctly in the low coverage scenario (Figure 45), (ii) it was well within the distribution of coverage of observed resistance sequences (Figure 37), and (iii) choosing a value at the lower end of expected coverages minimised storage requirements and processing time.

Table 15: ART parameters

Metric	Estimated value	ART parameter
ART error profile	-	HISEQ2500
Mean DNA fragment length	477bp	480bp
Fragment length standard deviation	164bp	150bp
Read length	151bp	151bp

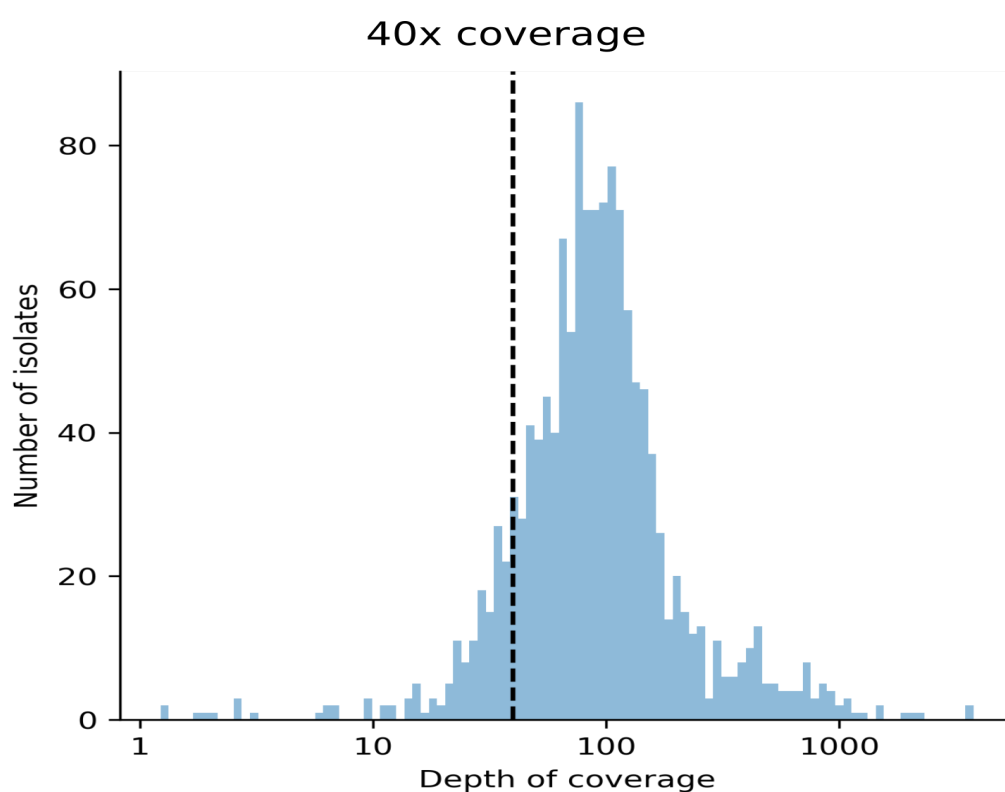


Figure 37: Beta-lactamase depth of coverage from real (non-simulated) samples

Note: Depth of coverage of all SRST2 identified beta-lactamases sequences as estimated by SRST2, used for parameter derivation

5.2.3. Assessing combined approaches on real world data

The rule-based approach used in section 4.2.1 was used to create predicted beta-lactam phenotypes for each method. In addition to comparing the resistance predictions of each method, methods were combined create combination predictions based on the following (Figure 38).

- Susceptible-preference (S-preference) phenotype: an isolate is predicted susceptible unless all tools in the combination identify it as resistant
- Unanimous: An isolate is predicted susceptible/resistant if and only if predictions from all methods agree, otherwise isolates are classed as unknown (and excluded from sensitivity/specificity analysis)
- Consensus: An isolate is predicted susceptible/resistant if and only if a majority of methods identify the isolate as susceptible/resistant respectively. If no majority exists, the isolate is classes as unknown (and excluded from sensitivity/specificity analysis)
- Resistance-preference (R-preference) phenotype: an isolate is predicted resistant if any method in the combination identifies the isolate as resistant.

Both combination and individual method predicted phenotypes were compared with observed AST results in isolates screened negative by cefoxitin AST (i.e. not chromosomal AmpC producing, N=1232). Of note this set contained APHA isolates and no meropenem resistant isolates.

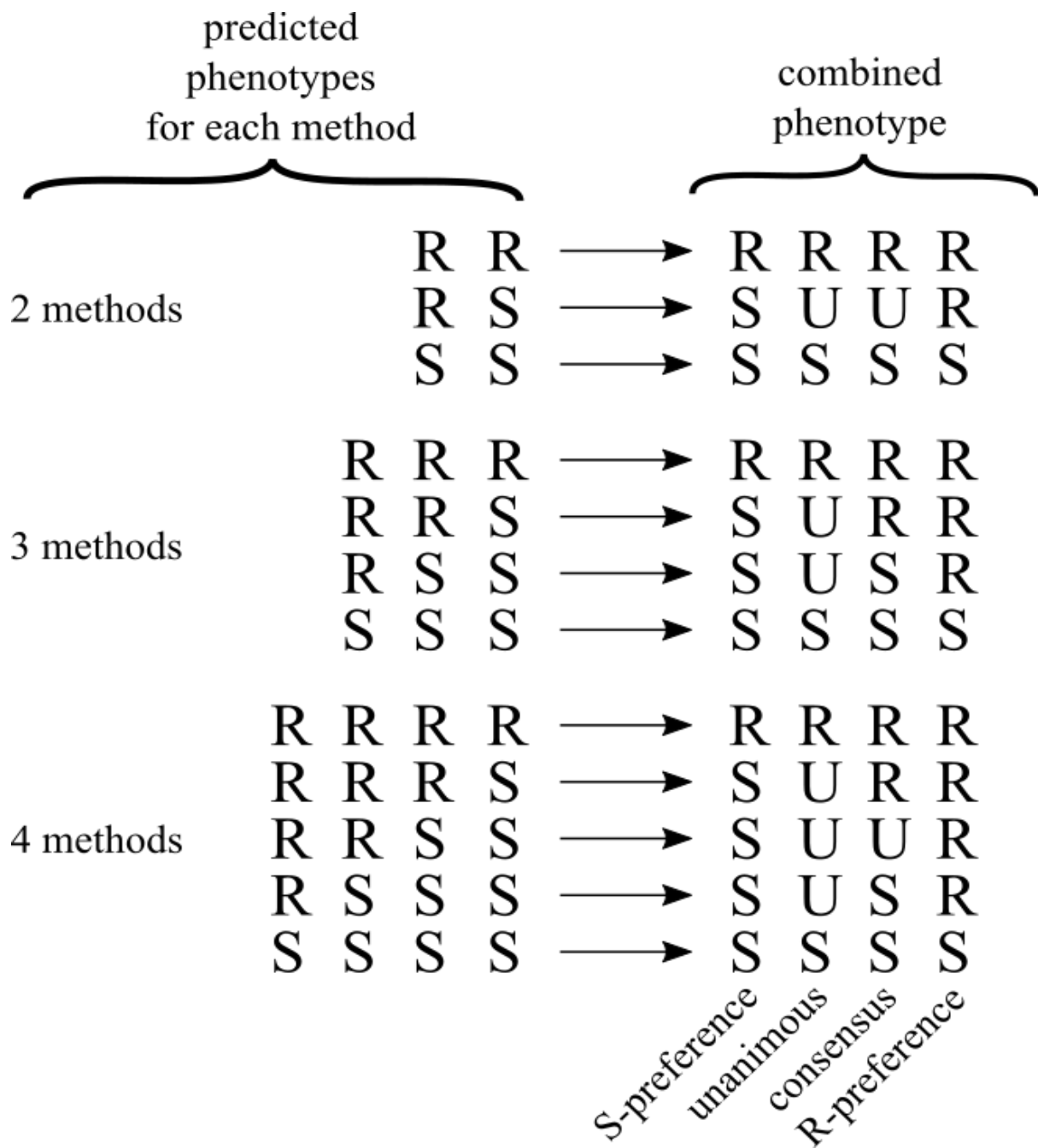


Figure 38: Methods of combining predicted phenotypes across the 4 bioinformatic tools

5.3. Results

5.3.1. Comparing resistance genotypes

Concordance of methods

Over the five classes of antibiotic, there was a large number of differences in the genotypes detected by the 4 methods, with 934/1818 (51%) isolates having discrepancies for at least one antibiotic-specific genotype (Figure 39). The frequency of discrepancies was method and antibiotic specific. For beta-lactams and trimethoprim, methods were generally concordant, with discrepancies in less than 20% of isolates with detected resistance mechanisms (147/1071 (14%), 133/759 (18%) respectively). Further, for these antibiotics, most commonly discrepant were due to a single method, generally SRST2, differing from the other methods. By contrast discrepancies for aminoglycoside, quinolones and sulphonamide resistance genes were more common, occurring in over half of isolates with detected transmissible resistance mechanisms (582/1028 (57%) aminoglycoside, 103/194 (53%) quinolone and 464/915 (51%) sulfonamide). Further, for these antibiotics, there was often less agreement between programs, with many isolates having three or more reported possible genotypes. For all antibiotics, unsupported genotypes (i.e. only suggested by a single method) were most commonly produced by SRST2 (Figure 40).

		beta-lactam	aminoglycoside	quinolone	trimethoprim	sulfonamides	any
ABRicate ARIBA KmerResistance SRST2	Ab Ar Km Sr	1671	1236	1715	1685	1354	884
	Ab Ar Km Sr	45	21	72	81	135	146
	Ab Ar Km Sr	29	58	17	20	14	57
	Ab Ar Km Sr	10	13	1	0	3	9
	Ab Ar Km Sr	8	9	1	24	6	11
	Ab Ar Km Sr	6	52	2	4	1	38
	Ab Ar Km Sr	1	0	1	0	0	0
	Ab Ar Km Sr	3	0	0	0	1	2
	Ab Ar Km Sr	11	13	0	1	5	42
	Ab Ar Km Sr	1	51	5	0	289	174
	Ab Ar Km Sr	3	3	1	0	1	3
	Ab Ar Km Sr	12	216	3	0	1	185
	Ab Ar Km Sr	0	3	0	1	0	3
	Ab Ar Km Sr	9	3	0	2	0	6
	Ab Ar Km Sr	9	140	0	0	8	258

Figure 39: Concordance of resistance genotyping methods by antibiotic class

Note: Panel on the left indicates which methods (if any) were in agreement (joined by lines). Ab = ABRicate, Ar = ARIBA, Km = KmerResistance, Sr = SRST2. Number on the right indicates the number of isolates which had each agreement “pattern”. Groupings by antibiotic were done by specifically looking for resistance genes in each of Resfinder’s beta-lactam, aminoglycoside, quinolone, trimethoprim and sulphonamide fasta files. The any column represents any discordance found using all elements in Resfinder (isolates can have multiple errors e.g. if only ABRicate and ARIBA agree for one gene, and only SRST2 and KmerResistance agree for another, then combined all methods would give different results for that isolate).

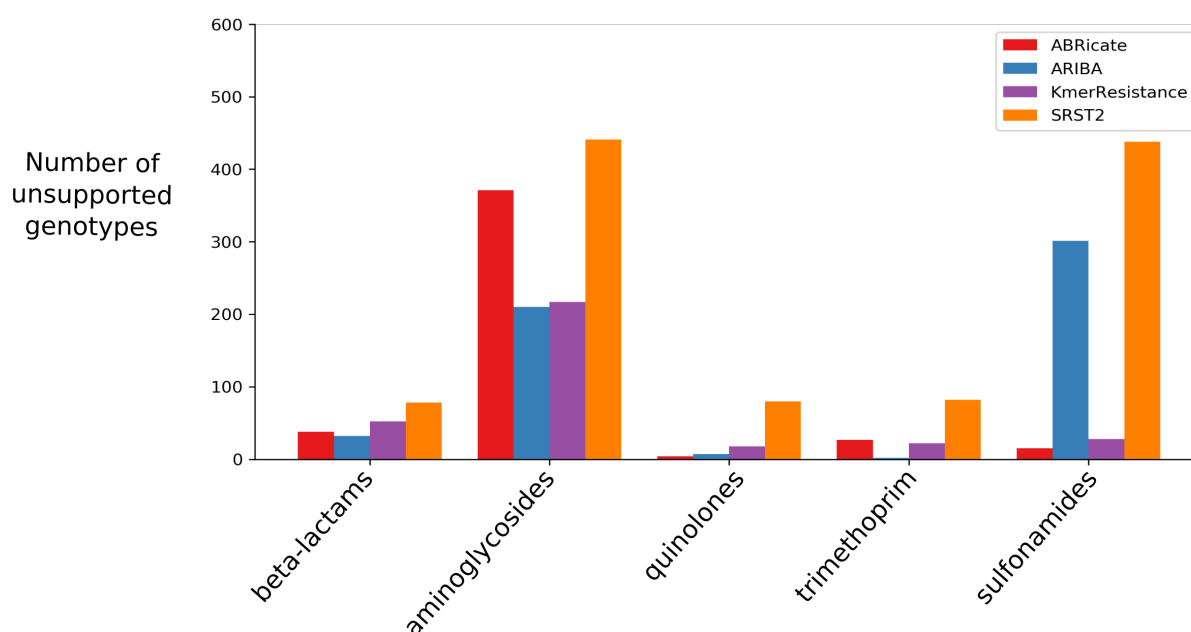


Figure 40: Frequency of unsupported genotypes (suggested by a single method alone) by antibiotic and bioinformatic method

5.3.2. Characterising genotypic discrepancies

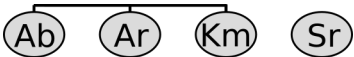
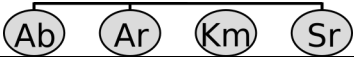
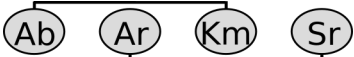
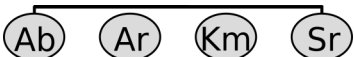
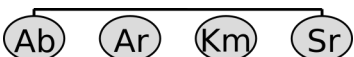
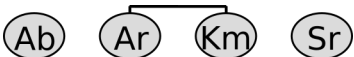
Beta-lactam antibiotics

More than half of all beta-lactam genotype discrepancies (80/147) (54%) between the four programs were identified artefacts, being attributable to one of two errors:

- SRST2 identifying spurious additional beta-lactamases due to how it clusters elements of resistance databases by similarity to identify “best matching allele”
- Incorrect choice of allele based on DNA-matching.

When isolated, these errors typically caused only a single method to be discordant, but when combined could make all four methods disagree (Table 16).

Table 16: Causes of discrepancy of beta-lactam genotypes

Discrepancy Pattern ⁺	No isolates	Causes ^o	Identified Artefact ^o
	43	Clustering	Yes
	1	DNA matching	Yes
	1	Clustering, DNA matching	Yes
	23	Partial/low coverage	No
	6	Possible collapsed variant	No
	5	Possible collapsed variant	No
	3	Fragmented assemblies	No
	1	DNA matching	Yes
	1	Cut-offs	No
	3	Fragmented assemblies	No
	2	Partial/low coverage	No
	2	Possible collapsed variant	No
	1	Novel	No
	4	Cut-offs	No
	2	Fragmented assemblies	No
	1	Novel	No
	2	DNA matching	Yes
	1	Partial/low coverage	No
	6	Clustering, DNA matching	Yes
	4	Cut-offs	No
	1	Clustering, possible collapsed variant	No
	1	Clustering, Partial/low coverage	No
	2	Partial/low coverage	No
	1	Possible collapsed variant	No
	12	Clustering, DNA matching	Yes
	8	DNA matching	Yes
	1	Possible collapsed variant	No
	8	Clustering, DNA matching	Yes
	1	Possible collapsed variant	No

+ : lines join methods with the same predicted genotype.

O : for causes and classification into identified artefacts/not see section 5.3.2.

Note: Ab – ABRicate, Ar – ARIBA, Km – KmerResistance, Sr – SRST2.

The SRST2 clustering issue was the most common type of error encountered, occurring in 71/147 (48%) isolates with discrepancies (43 in isolation with no other type of discrepancy). In each of these cases, SRST2 identified more than one *bla*_{CMY}/*bla*_{SHV}/*bla*_{IMP} gene; however, there were <10 reads mapping individually to each allele, no clear heterozygosity on read pileups, and no fragmentation in assembly graphs. The plausible cause is that, in order to select a “best allele” match, SRST2 clusters the resistance database into “families” based on a single CD-hit identity threshold. Reads are then (multiply) matched to these families and any matches with sufficient identity are reported. However, families of beta-lactamases are not easily defined by sequence identity, with the number and size of gene families created being highly dependent on the chosen cut-off (Figure 41). Default database preparation in SRST2 partitions each of the *bla*_{IMP}, *bla*_{CMY} or *bla*_{SHV} families into multiple subfamilies. However, given how highly similar these genes are (Figure 42), SRST2 can often map reads to genes in either of multiple subfamilies, hence finding multiple genes when only one existed in the underlying isolate. In this instance, none of these clustering-related artefacts changed the predicted phenotype because these errors only caused within family reporting duplications and none of the extra reported genes were associated with a different phenotype to the primary gene.

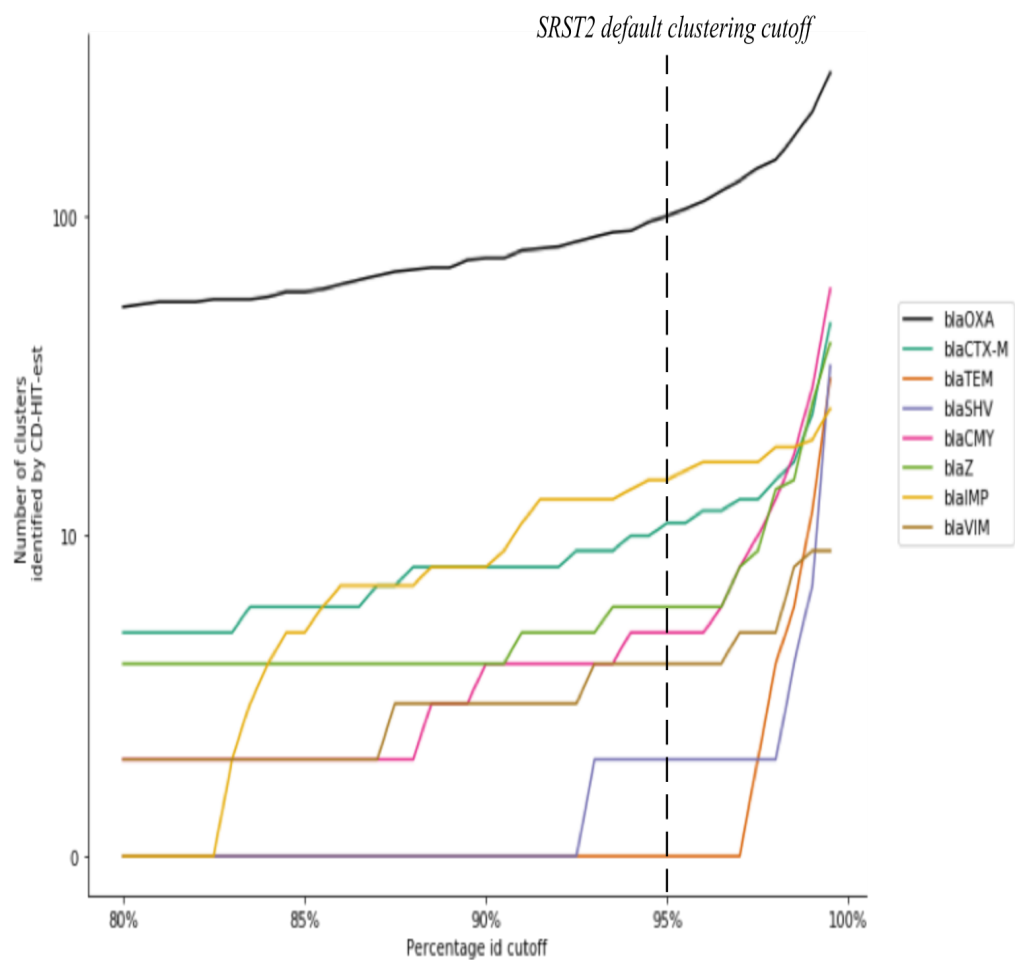


Figure 41: The impact of threshold on database clustering

Note: The (log) number of clusters created for each beta-lactamase family, dependent on clustering threshold.

```

1 MMKSLCCALLLTASFSTFAAAKTEQQIADIVNRTITPLMQEQAIPGMV 50
  |||:|||||
1 MMKSLCCALLLTASFSTFAAAKTEQQIADIVNRTITPLMQEQAIPGMV 50

51 AVIYQGKPYFTWKGADIANNHPVTQOTLFELGSVSKTFNGVLGGDAIAR 100
  |||:|||||
51 AVIYEGKPYFTWKGADIVNHPVTQOTLFELGSVSKTFNGVLGGDAIAR 100

101 GEIKLSDPVTKYWPELTGKQWQGIRLLHLATYTAGGLPLQIPDDVRDKAA 150
  |||.|||||
101 GEIKLSDPVTKYWPELTGKQWQGISLLHLATYTAGGLPLQIPDYVTDKAE 150

151 LLHFYQNWQPQWTPGAKRLYANSSIGLFGALAVKPSGMSYEEAMTRRVLQ 200
  |||.|||||
151 LLRFYQNWQPQWTPGAKRLYANSSIGLFGALVVKPSGMSYEEAMTRRVLQ 200

201 PLKLAHTWITVPQNEQKDYAWGYREGKPVHVSPGQLDAEAYGVKSSVIDM 250
  |||:|||||
201 PLKLAHTWITVPQSEQKNYAWGYREGKPVHVSPGQLDAEAYGVKSSVIDM 250

251 ARWVQANMDASHVQEKTLQQGIALAQSRWIRGDMYQGLGWEMLNWPLKA 300
  |||.|||||
251 ARWVQANMDASHVQEKTLQQGIELAQSRWIRGDMYQGLGWEMLNWPLKA 300

301 DSIINGSDSKVALAALPAVEVNPPAPAVKASWVHKTGSTGGFGSYVAFVP 350
  |||:|||||
301 DSIINGSDSKVALAALPAVEVNPPAPAVKASWVHKTGSTGGFGSYVAFVP 350

351 EKNLGIVMLANKSYPNPVRVEAAWRILEKLQ 381
  |||.|||||
351 EKNLGIVMLANKSYPNPARVEAAWRILEKLQ 381

```

Figure 42: Pairwise alignment of two genes identified by SRST2 in the same isolate

Note: Needleman-Wunsch²⁰⁸ alignment of the *bla*CMY-2 and *bla*CMY-81 proteins, both identified by SRST2 in the same isolate. Alignment has zero gaps and 1080/1146 (94.2%) identities. When using a percentage identity threshold of 95%, these genes are identified by CD-hit-est as belonging to two different clusters.

As seen when comparing databases in section 4.3.2, the other common artefactual error encountered was incorrect choice of allele based on DNA matching. To recap, this occurs when the identified DNA sequence fails to match any element in the database, however there is a perfect protein match. Based on optimal DNA matching (as opposed to protein matching), algorithms incorrectly select a non-exact matching protein. This error occurred in 38 of the 147 isolates with discrepancies, and in all cases when choosing between best-matching *bla*_{TEM} or *bla*_{SHV} alleles. Given that *bla*_{SHV} alleles also cause clustering errors for SRST2, this error only occurred in isolation in 12 isolates. All methods produced an allele-choice error at least once, but for ABRicate and ARIBA misnaming alleles was more common, occurring in 35 and 28 isolates respectively. Further, incorrect allele naming was

not consistent between methods, with different alleles being incorrectly called two different ways for 17 isolates.

In contrast to the clustering errors, DNA matching errors did cause different antibiograms for 8/12 cases where they occurred in isolation. Seven of these were mis-identified *bla*_{TEM-135} alleles (a class 2b enzyme) which was correctly identified by SRST2 and KmerResistance. Isolate phenotype matched this, only being resistant to ampicillin. While ARIBA identified these as *bla*_{TEM-1} genes which would agree with this phenotype, ABRicate identified *bla*_{TEM-106}, an ESBL *bla*_{TEM} variant. In the remaining case, all methods mis-identified a *bla*_{TEM-190} gene (confirmed when translating consensus sequences produced by each program).

Interestingly, this gene is thought to produce an ESBL, but phenotypically the isolate was only penicillin and BL/BLI resistant. This may however also represent uncertainty in classification, as there is little evidence confirming this enzyme's phenotypic profile ²⁰⁹.

The remaining discrepancies in beta-lactam genotypes were less easily confirmed as artefacts. They broadly followed three patterns:

- Partial/low coverage elements discrepantly identified by the four methods
- The assembly of the element was broken, and output from the four methods suggested presence of more than one variant in the isolate
- The assembly of the element was broken, but there was no suggestion of more than one variant in the isolate.

Partial/low coverage beta-lactamase genes were discrepantly found in 39 isolates (Table 17). *bla*_{TEM} like genes were the most common type of poor quality gene found, being

identified in 29 isolates. KmerResistance reported all of these genes as present, with the other programs being more variable on which they reported. In all but 4 cases, KmerResistance reported that the depth of the gene was less than 5x (Figure 43). Of the 4 discrepant present at greater than 5x depth, 3 (present at depth > 100x) were omitted from ARIBA reports as ARIBA assemblies contained mis-sense mutations and the final one (present at depth 17x) failed to assemble for ABRicate.

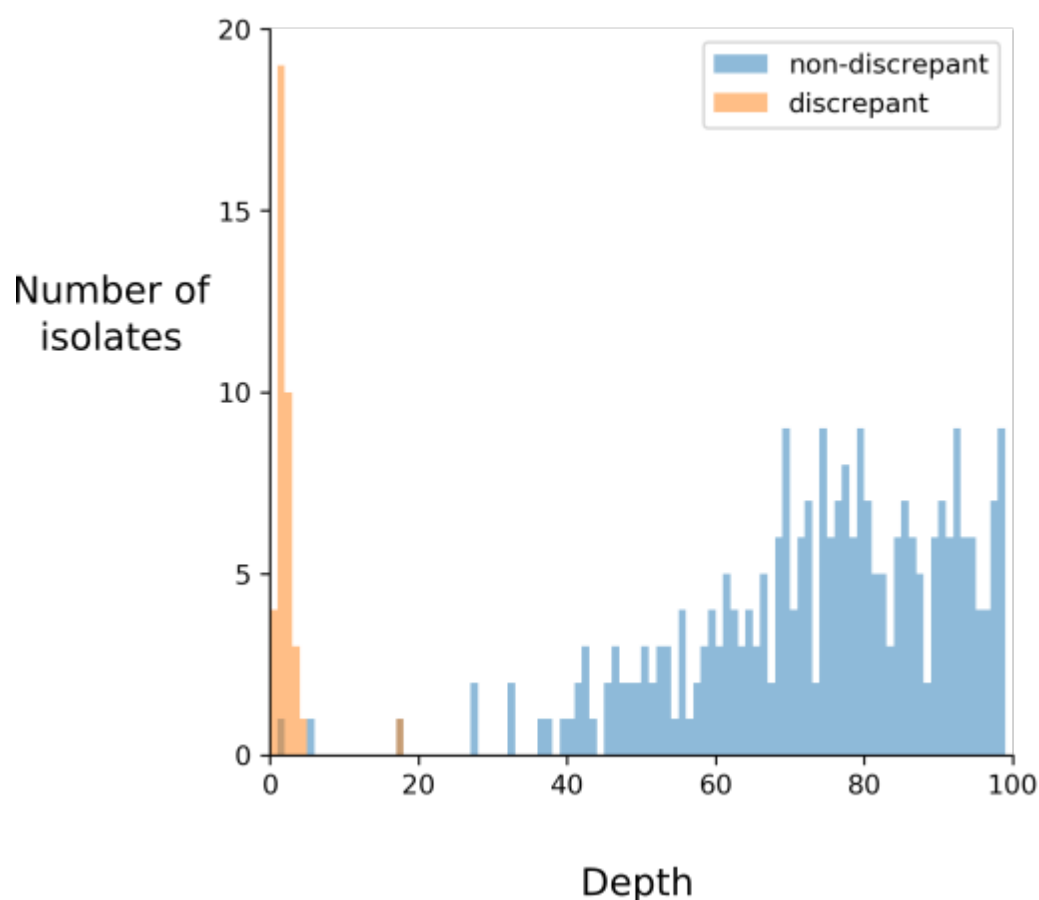


Figure 43: KmerResistance reported beta-lactamase gene depth

Note: Orange - inconsistently reported genes. Blue – consistently reported genes. Depths above 100 not shown. Only 3 inconsistently reported genes had depth greater than 100, all of which were truncated on ARIBA assembly (see above).

For these errors, both whether the gene was present in the isolate, and if present, reasons why only certain bioinformatic methods identified it were unclear. Any observed beta-

lactam resistance only occurred 6/29 (21%) cases where these partial elements were present in isolation (Table 17), suggesting many of these predicted genes are artefactual. However, given each method has its own approach to filtering spurious matches which are not directly comparable, it is unclear which, if any of these discrepancies, are due to a genuine difference of between methods in their ability to identify resistance genes from poor quality data as opposed to simple cut-off differences.

Table 17: Isolates containing low coverage/partial genes variably identified by different methods

Discrepancy pattern (grey indicates not found)				All isolates		Only isolates with no other beta-lactamase present	
Ab	Ar	Km	Sr	Number isolates	Number ampicillin resistant	Number isolates	Number ampicillin resistant
				23	6	17	3
				9	4	7	2
				3	1	2	0
				2	2	1	1
				2	0	2	0
Total				39	13	29	6

Note: Ab = ABRicate, Ar = ARIBA, Km = KmerResistance and Sr = SRST2. Discrepancy pattern indicated by left panel with white indicating gene found by this method, grey indicating not found. E.g. the first line represents 23 isolates which KmerResistance only identified as containing partial/low coverage genes.

The other unclear error resulted from fragmentation of genes in ABRicate and/or ARIBA assemblies, occurring in 24 isolates. As found in section 4.3.2., there was evidence that 16 of these isolates contained multiple closely related beta-lactamase alleles. For these 16, ARIBA and SRST2 always found single alleles, but to highlight the possibility of heterozygosity ARIBA flagged “variants_suggest_collapsed_repeat” and SRST2’s maximum “minor allele frequency” was high (>20%). KmerResistance often reported two alleles, one with high depth, PID and coverage, and one much less accurately. This likely reflects KmerResistance’s

“winner takes all strategy”, where, once matching unique k-mers are counted, the allele with the most matches is also assigned all reads with non-unique kmer-matches (Figure 44). This then leaves only reads with unique k-mers matching the secondary allele, which together then struggle to assemble a complete gene. Finally for ABRicate, assemblies were variably fragmented by both Velvet (15/16 cases) and SPAdes (4/16) with Bandage graphs²⁰⁰ similar to those in section 4.3.2. Only 4/16 of these differences were associated with changes in predicted phenotype, which in all cases agreed with the single allele predictions of ARIBA and SRST2.

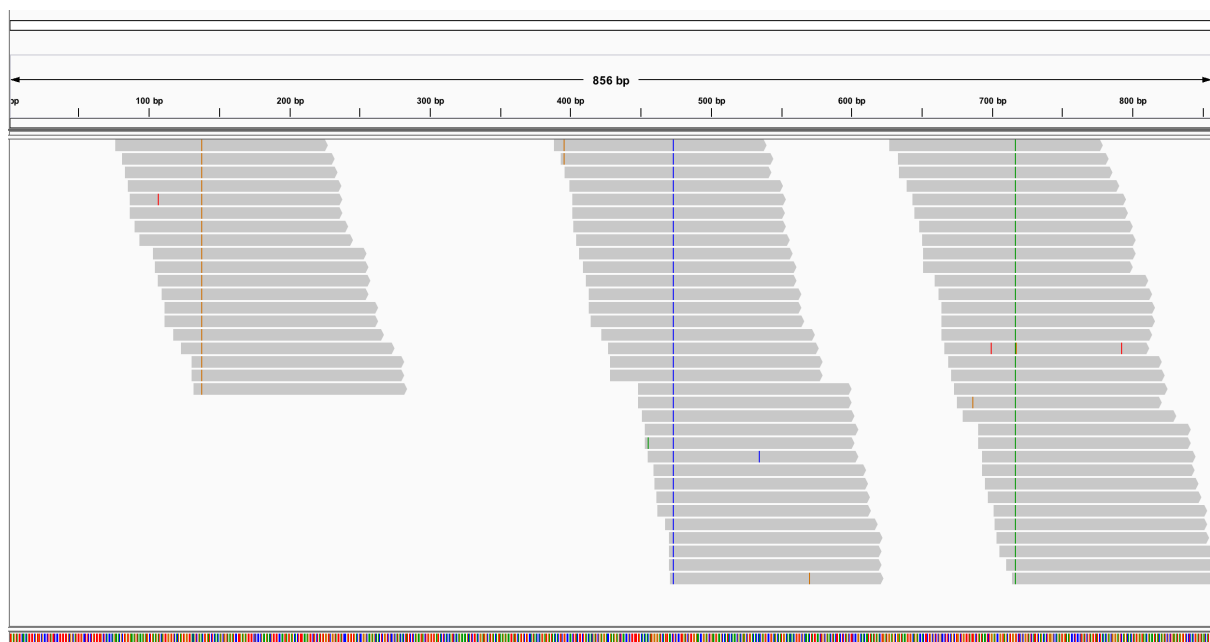


Figure 44: Reads assigned to secondary incomplete bla_{TEM} allele only by KmerResistance

However, the presence of multiple closely related alleles being the sole cause of assembly problems was unlikely. This was evidenced by the fact that the apparent occurrence of secondary alleles were neither necessary nor sufficient to cause assembly errors.

Non-beta-lactam antibiotics

For non-beta-lactam antibiotics, most discrepancies appeared to be similarly due to artefacts (Table 18). All errors for the quinolone and sulfonamide genotypes were due to bioinformatic artefacts alone. In addition to those already described in this chapter, the issue of sequence length, encountered in section 4.3.2, caused frequent discrepancies among the non-beta-lactam genotypes. This however is not simply database curation, but also partially reflects the biological complexity of some resistance mechanisms.

Aminoglycosides are a case in point where the context of some genes (e.g. the *aadA* gene cassettes) are prone to frequent rearrangement. Consequently, complex elements such as fusion proteins, and coding variant genes “contained” within one another (i.e. one sequence is a subsequence of another), are common. The situation is further complicated for aminoglycosides because some genes, such as the *strA-strB* linked gene pair frequently have large insertions, as previously demonstrated ²¹⁰ and seen in this dataset, which are poorly identified by KmerResistance and SRST2.

Table 18: Causes of discrepancies for non-beta-lactam genotypes

Discrepancy Pattern ⁺	No isolates	Causes ^o	Identified Artefact ^o
	31	Partial/low coverage	No
	14	Novel	No
	9	Possible insertion	No
	4	Partial/low coverage, Length	No
	52	Possible insertion	No
	46	Clustering, Length	Yes
	4	Fragmented assemblies, Clustering, Length	No
	4	Clustering, Length Novel	No
	215	Clustering, Length	Yes
	1	Clustering, Novel	No
	90	Clustering, Length	Yes
	35	Fragmented assembly, Clustering, Length	No
	12	Clustering, Length, Insertion	No
	2	Clustering, Length, Novel	No
	1	Clustering, Length, Partial/low coverage	No
	72	Length	Yes
	81	Length	Yes
	20	Partial/low coverage	No
	19	Length	Yes
	5	Partial/low coverage	No
	96	Clustering	Yes
	39	Clustering, Length	Yes
	289	Clustering, Length	Yes

+ : lines join methods with the same predicted genotype.

O : for causes and classification into identified artefacts/not see section 5.3.2.

Note: Ab – ABRicate, Ar – ARIBA, Km – KmerResistance, Sr – SRST2.

5.3.3. Understanding discrepancies using simulations

Low quality simulations

Mirroring what was seen in real isolates, in the simulated datasets where the “true” *bla*_{TEM} genotype was known and only one single gene was present, KmerResistance was able to identify alleles at lower coverage than any of the other methods. As found when compared against ARIBA by Hunt et al ¹³⁸, KmerResistance’s accuracy initially dropped as coverage increased, possibly reflecting issues with its “winner takes all” strategy. Above 15-20x coverage depth, all methods performed well at identifying the 176 *bla*_{TEM} alleles in constructs (Figure 45). Further, all methods were able to identify all of the *bla*_{TEM} alleles correctly at least once. However, increasing depth of coverage did not invariably correct mistakes made by methods, with examples existing for all methods where the correct allele was identified at low coverage, but then mis-identified at higher coverages. In general, ABRicate and SRST2, while requiring greater coverage to identify correct *bla*_{TEM} alleles initially were more accurate at higher coverage, making mistakes in only 1/176 (0.6%) and 0/176 (0%) of *bla*_{TEM} alleles respectively in simulations with greater than 20x coverage depth. In contrast, beyond 20x coverage ARIBA and KmerResistance made mistakes for 23/176 (13%) and 6/176 (3%) alleles respectively. Beyond 40x coverage (the threshold used for the single and multiple allele scenarios) ABRicate was incorrect for one (0.6%), ARIBA for four (2%), KmerResistance for one (0.6%), and SRST2 for zero (0%) alleles.

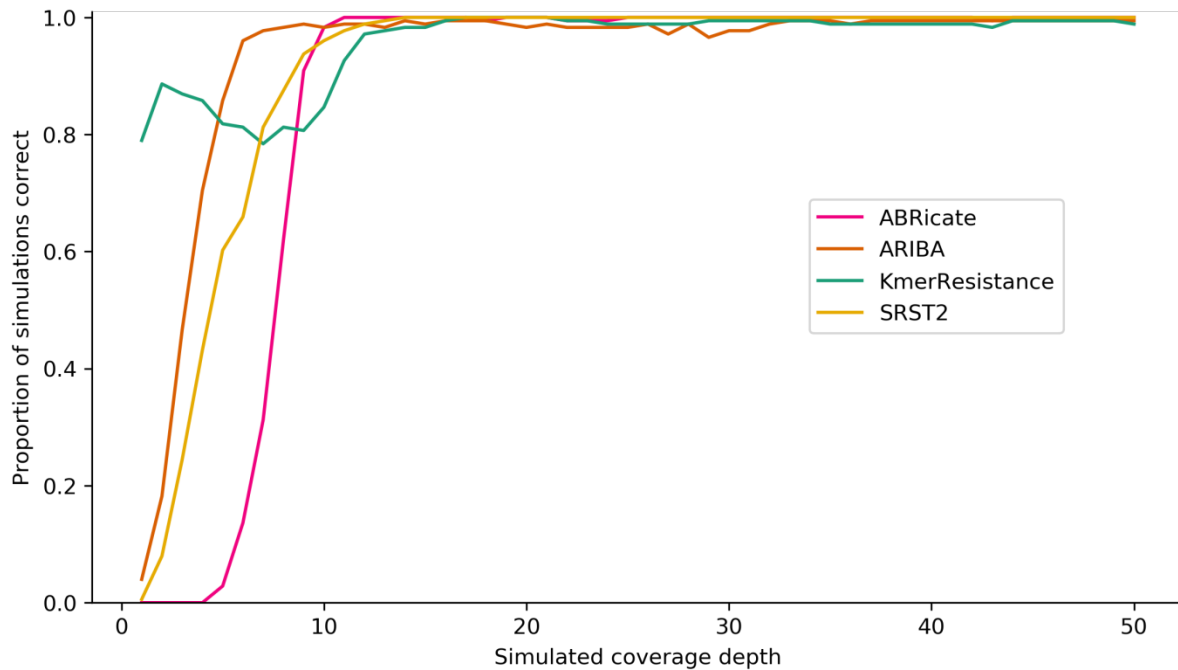


Figure 45: Simulating the impact of coverage on resistance allele identification

Multiple allele simulations

The multiple allele simulation caused significant problems for assembly based algorithms, with ABRicate reporting fragmented/incomplete alleles for 32194/46279 (70%) simulations and ARIBA reporting no alleles meeting its assembly quality requirement for 32987/46279 (71%) simulations. Nevertheless, ARBicate did manage to construct the correct alleles in 11382/46279 (25%) of simulations, whereas ARIBA only managed to correctly identify both members of the pair in 2494/46279 (5%) (Table 19). SRST2, as expected generally only found a single allele as dictated by its clustering parameters. Of the four programs, KmerResistance performed the best, often identifying both alleles correctly.

For all four programs, they were most likely to make mistakes, and mistakes were more severe, for closely related alleles (Figure 46, Chapter 5 Supplementary Figures 1-3).

Table 19: Multiple allele simulation performance

Score	Interpretation	Number with score (% of total 46279 simulations)			
		ABRicate	ARIBA	KmerResistance	SRST2
1	Nil correct	17145 (37%)	36150 (78%)	489 (1%)	9898 (21%)
2	One correct but additional incorrect matches	2419 (5%)	2 (0%)	1452 (3%)	152 (0%)
3	One correct, no incorrect matches	15333 (33%)	7634 (17%)	2203 (5%)	33077 (71%)
4	Two correct but additional incorrect matches	0 (0%)	1 (0%)	3309 (7%)	613 (1%)
5	Two correct, no incorrect matches	11382 (25%)	2494 (5%)	38826 (84%)	2539 (5%)

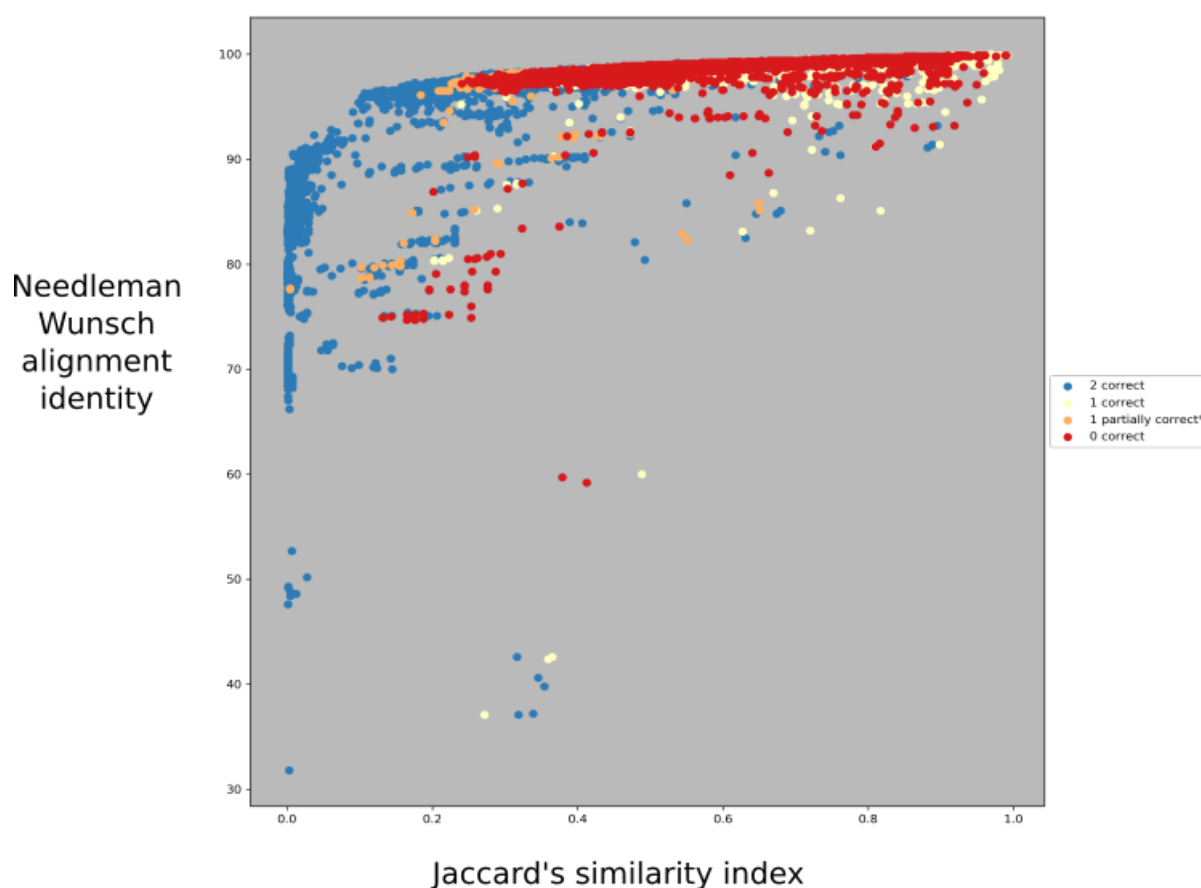


Figure 46: Simulation score by pair similarity for ABRicate

Note: Scores obtained by ABRicate for each pair of genes in a simulation denoted by colour. Dark blue = 2 correct alleles identified, no incorrect alleles identified. Yellow = 1 correct allele identified, no incorrect alleles identified. Orange = 1 correct allele identified, at least one incorrect allele identified. Red = 0 correct alleles identified.

5.3.4. Impact of bioinformatic method on beta-lactam resistance prediction

Consistent with the relatively small number of genotypic changes that caused alteration of the predicted phenotype (see section 5.3.3), differences in sensitivity and specificity between each approach was small (Figure 47-48, Chapter 5 Supplementary Figures 4-11). This was true even for highly discrepant antibiotics amoxicillin-clavulanate and piperacillin-tazobactam, where no alternate approach significantly improved upon the performance seen in Chapter 4 with ABRIcate alone. Likewise, differences between combination approaches were small and there was no obvious specific combination of methods that outperformed the others.

For each individual approach, there was generally a trade-off between sensitivity and specificity. For example, the lower coverage alleles KmerResistance detected led to a small improvement in sensitivity, but at a heavy cost to specificity (Figure 47). Looking at combined approaches, as expected the R-preference and S-preference were the most sensitive and specific respectively, but by their nature however, these were also more vulnerable to the effects of 'outlier' individual methods. Correspondingly, results for the other two prediction combination methods (unanimous/consensus) were less volatile reflecting the fact that most discrepancies in the beta-lactam genotype occurred for single methods (section 5.3.2). The unanimous approach generally had high sensitivity and specificity, at the expense of not predicting a phenotype for between 1-5% of isolates for all samples. The "consensus" prediction offered a compromise between the conservative "unanimous" approach and the more extreme R-preference/S-preference approaches. While not providing superior sensitivity and specificity to individual methods, it classified more than 98% of isolates for all testable antibiotics, and results were generally still resilient

to the effect of outlier methods, providing good performance for both sensitivity/specificity. While outlier methods appear to influence the consensus approach as much when using three methods as compared to two, this is likely artefactual, resulting from the fact that the consensus approach is equivalent to the “unanimous” approach when using two methods (Figure 38), but instead classifies all isolates when using three.

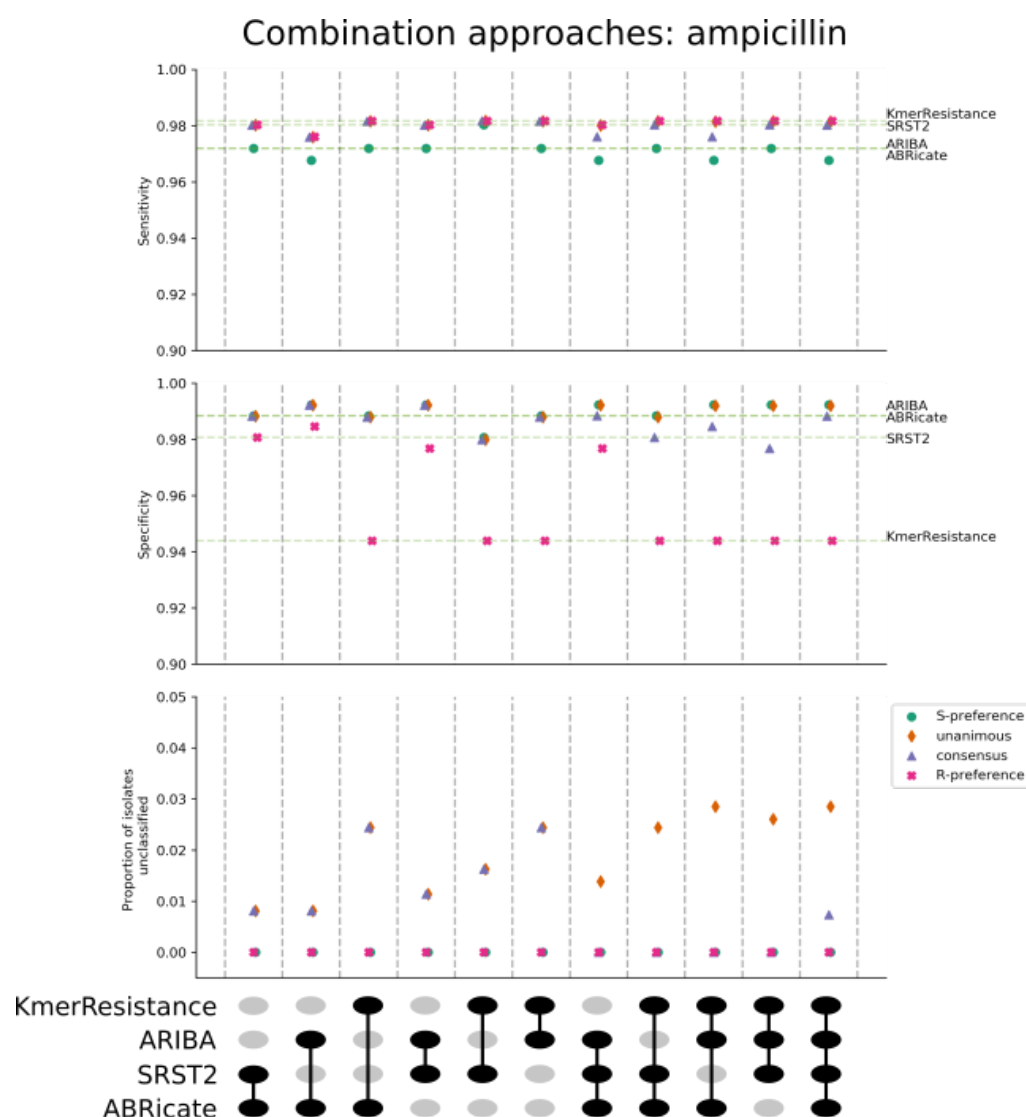


Figure 47: Sensitivity and specificity of combination approaches for predicting ampicillin resistance

Note: Top panel) For each combination (as indicated at the bottom of the figure), sensitivity of combined phenotypes (S-preference, unanimous, consensus and R-preference, definitions in section 5.2.3). Horizontal green dashed lines indicate the sensitivity of individual methods. Middle panel) Specificity for each combination. Bottom panel) Number

of isolates with unclassified 'U' predicted phenotypes due to disagreement of predicted phenotype between individual methods (section 5.2.3).

Combination approaches: amoxicillin-clavulanate

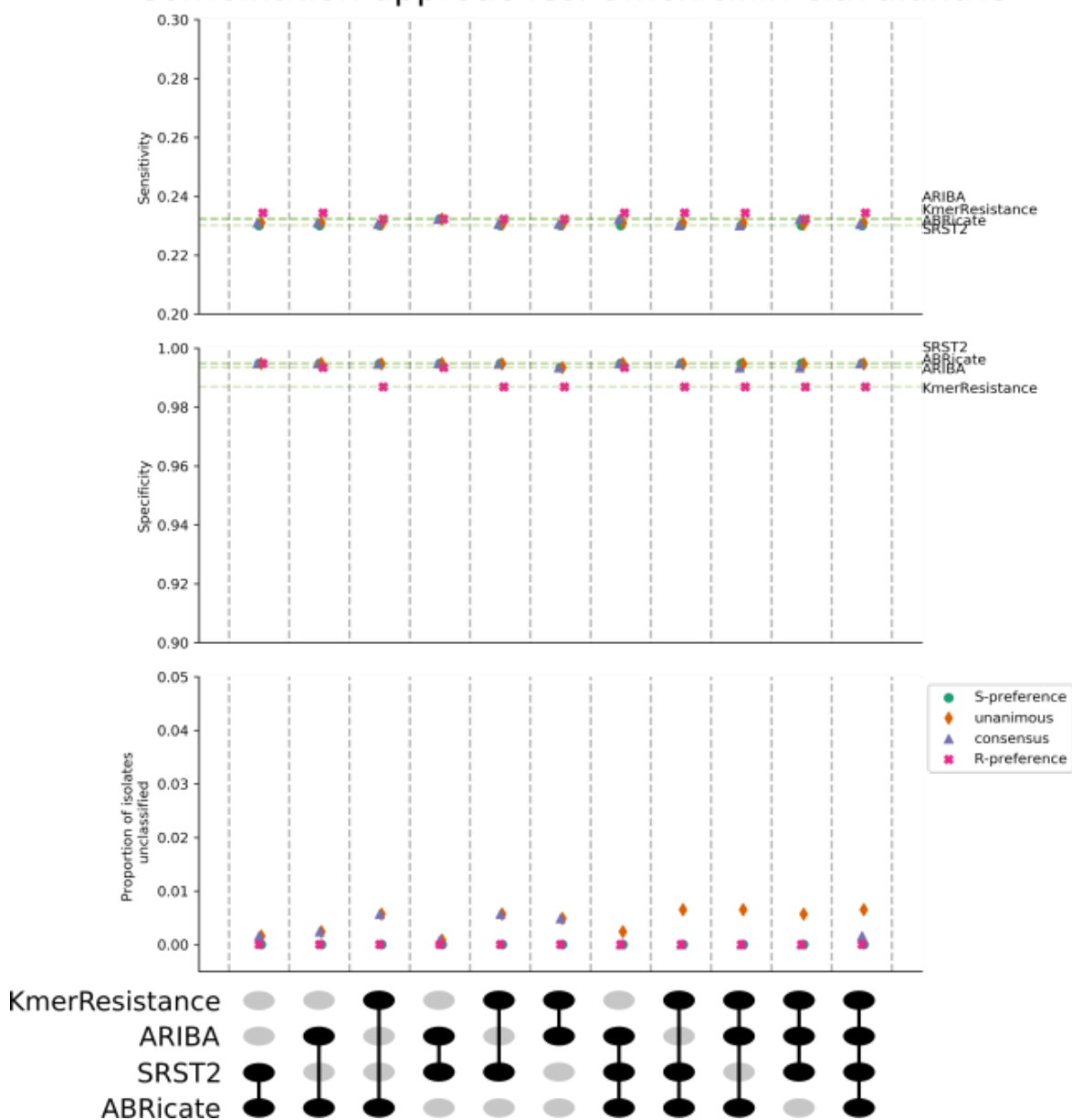


Figure 48: Sensitivity and specificity of combination approaches for predicting amoxicillin-clavulanate resistance

Note: Top panel) For each combination (as indicated at the bottom of the figure), sensitivity of combined phenotypes (S-preference, unanimous, consensus and R-preference, definitions in section 5.2.3). Horizontal green dashed lines indicate the sensitivity of individual methods. Middle panel) Specificity for each combination. Bottom panel) Number of isolates with unclassified 'U' predicted phenotypes due to disagreement of predicted phenotype between individual methods (see section 5.2.3).

5.4. Discussion

Contrary to the belief that WGS produces reliable resistance genotypes, the choice of bioinformatic method had a significant impact on the genes identified from a single database in large sets of *E. coli* isolates and in simulations. Despite the large number of discrepancies, most were definitively “artefactual”, caused by methods not properly accounting for the nuances of resistance databases. These discrepancies were most commonly due to poor performance of a single method, making a combination of methods an attractive approach as it both makes the error visible and mitigates its impact. In addition to the “definitively artefactual” discrepancies, there were a second group of discrepancies where underlying sequence was less certain. In these discrepant, isolates contained low coverage/partial elements or elements which were fragmented on assembly. Low coverage discrepant were hypothesized to be potentially due to a genuine difference in performance at low coverage. Fragmented assembly discrepant were hypothesized to be due to isolates containing multiple closely related resistance sequences. Simulations demonstrated both of these hypothesis were plausible. KmerResistance, which by its nature of matching requires the least sequence to choose a best match outperformed other methods in both of these scenarios. Consistent with findings when investigating discrepant, consensus approaches helped reduce the impact of artefactual errors on resistance prediction and performed with comparable sensitivity and specificity to the best performing single method. However, as was suspected in Chapter 4, modifying bioinformatic approaches had little effect on the overall level of discrepancies between genotypic prediction and observed phenotype seen for amoxicillin-clavulanate.

The artefactual errors, while seemingly trivial in nature, pose serious questions concerning how data on resistance genetics is stored and searched. For example, beta-lactamase allelic naming is based solely on protein sequence, and resistance databases typically only contain one variant per protein. To avoid missing DNA variants, either they need to be included in the database, or isolate sequence must be translated to match on a protein basis. Both of these approaches however have downsides. Including extra DNA sequences in the database makes them larger, slower to search, and could worsen the task of identifying DNA variants for other alleles given the increased number of alleles to choose from. On the other hand, translating isolate sequence and protein matching requires consensus sequence, again slowing search. The other two common artefactual errors encountered in this chapter, due to either differences in matching based on alignment scores versus identity or database clustering, reflect the fact that resistance sequences from gene families do not necessarily conform to the simplifying assumptions required by these bioinformatics methods. Consequently, these also have no straightforward solutions and are likely to become more problematic as resistance databases become more complex. Further, with the increasing number of resistant variants within genetic families (e.g. the complex mutant *bla*_{TEM} genes), artefactual mistakes may affect prediction more drastically in future. The other challenging aspect of these errors is their unpredictable nature, as they can occur with any change in database content. While generally easily identifiable/rectifiable on manual interrogation, whether or not this approach will be scalable to routine clinical practice is unclear.

Simulations demonstrated the “plausibility” of hypotheses as to the cause of the non-artefactual errors. Simulations of low coverage alleles demonstrated difference in performance of the bioinformatic methods with KmerResistance (as seen in real world data)

more successfully identifying low coverage alleles. However, while simulations suggest matching sequence may genuinely be in isolate sequence data (as opposed to purely due to bioinformatic error), the importance for resistance prediction is unclear given 80% of these isolates were fully susceptible on AST. Predicting resistance from low coverage elements is similar to the problem predicting resistance in metagenomic data, which as of July 2019 is largely still open ²¹¹. The multi-allele simulations demonstrated these situations still cause significant problems for most short-read sequencing methods. Further, these simulations represented an ideal challenge, where flanking sequence around each allele is unique and can be used to distinguish alleles. Real-world data is likely to be tougher, with more genes inserted in potentially highly similar backgrounds. However, while this simulation demonstrated that multiple-alleles would lead to fragmented assembly, the significantly worse performance of ARIBA in the simulations as compared with real data suggest that the simulations were not accurately capturing all issues seen in the real world data. Interestingly, both of these simulations also test the tasks required for metagenomic resistance genotyping. This therefore highlights that methods able to working with less complete sequence such as KmerResistance may perform better in this setting.

This work demonstrated that combining methods may be useful in both resistance genotyping and, to a lesser extent, phenotype prediction. Using a combined approach enables easy identification of both artefactual errors (as these are only caused by a single program) and of situations where the underlying sequence is uncertain. This has two potential benefits. Firstly, it helps counter the unpredictable nature of artefactual errors introduced by database changes. Secondly, it makes identification of artefactual errors potentially automatable. This would be useful if whole genome sequencing is used clinically,

as, while manually checking the typically large numbers of isolates analysed in diagnostic laboratories is unfeasible, it may help restrict the problem of isolate sequences that need checking to a more manageable size. These benefits may also apply for resistance prediction, although prediction algorithms like the one used in this chapter often have greater tolerance for artefactual errors because multiple genes give the same phenotype. As a final benefit for combining methods, this approach could utilise the specific strengths and weaknesses of programs to get the best performance of all of them. For example, KmerResistance outperformed other methods at identifying multiple and low coverage alleles (Figure 45, Table 19), but performed poorly when identifying insertions (Table 18). Combining this approach with assembly based program (e.g. ARIBA) may help ameliorate this weakness. Further, as errors are commonly algorithm and gene specific, matching the best-performing algorithm for each gene could significantly improve genotyping performance.

This chapter has several limitations. The first major limitation is that, while simulations have demonstrated my hypotheses as plausible, patterns seen in the multiple allele simulations were different to those in the real data, suggesting the biological causes of these discrepancies are not clear. Not knowing the underlying cause of difference makes it difficult to take a view as to which method performs “best” in these situations. Likely these errors represent complexities of the accessory genome, including but not limited to the presence of multiple similar *bla*_{TEM} alleles and insertion sequences affecting assembly. Possible reasons for why my simulations failed to replicate this behaviour are the over simplistic nature of the randomised flanking sequence (when in reality *bla*_{TEM} is known to be associated with specific insertion sequences) and the lack of sequence representing the rest

of the *E. coli* genome, which can also affect assembly. While I could modify these simulations in future to try to better recreate scenarios affecting assembly, the loss of simplicity of the simulations might make them overall less interpretable. An alternative approach in future would be to re-sequence isolates with discrepancies using long-read sequencing to attempt to fully resolve accessory genome structures. Causes for mis-assembly could then be investigated by comparing *bla*_{TEM} flanking sequences both with each other and with the rest of the isolate genome.

The other major limitation of this work is that while I have demonstrated that many of the errors for beta-lactam genotyping apply to other antibiotic classes, I have not exhaustively investigated non beta-lactam discrepant. Although this potentially misses some types of discrepant, it is unlikely to affect the main outcome of this chapter, that current resistance genotyping in *E. coli* is prone to artefactual error. The rate of artefactual error has already been shown to be substantial, and hence even if the un-investigated errors are not, this does not detract from the size of the problem already identified. Another limitation was that an alternative databases were not used; but given that the Resfinder database is arguably the simplest and most well curated general purpose resistance database, and the fact less artefactual errors were identified when using it versus the CARD database in section 4.3.2, it is more likely that using alternate databases would exacerbate artefactual errors.

Regarding amoxicillin-clavulanate resistance prediction, this chapter demonstrates two useful points. Firstly, it argues against “hidden alleles” (i.e. samples containing two versions of highly related alleles, one inhibitor resistant and one susceptible, but only reporting the

susceptible allele). Even supposing all isolates where “collapsed repeats” were reported by ARIBA and SRST2 contained inhibitor resistant alleles, given their low frequency this would not account for the large difference between predicted and observed amoxicillin-clavulanate resistance. Secondly it highlights that while genotype can be significantly altered by method, this mostly translates to little difference in prediction of phenotype. Two observations, namely that bioinformatic prediction error is less plausible (given above) and the good resistance prediction for ampicillin (meaning any beta-lactam resistance is generally well identified) further reinforce the view that discrepancy is likely due to the resistance prediction algorithm itself or phenotype.

Chapter 6: Moving to a more realistic model of beta-lactam resistance

6.1. Introduction

This chapter uses prior knowledge of the specific mechanisms of amoxicillin-clavulanate resistance to construct an enhanced resistance prediction. This is then compared once again with phenotype to evaluate the impact of the predication algorithm changes.

6.1.1. Background and rationale

Amoxicillin-clavulanate

Amoxicillin-clavulanate is a combination antibiotic containing a beta-lactam (amoxicillin) and beta-lactamase inhibitor (clavulanate) (BL/BLI) which uses clavulanate to extend the activity of amoxicillin to include isolates which produce extended spectrum penicillinases^{212,213}. Of note, clavulanate itself is a beta-lactam antibiotic, however given its weak affinity for penicillin binding proteins (PBPs) and its kinetics strongly favouring inhibition for many enzymes, it is traditionally thought of as an inhibitor⁶⁵.

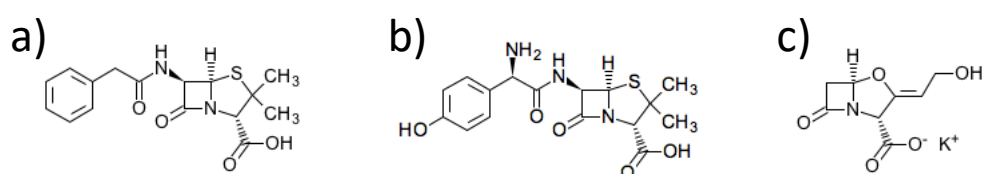


Figure 49: Chemical structure of a) benzylpenicillin b) amoxicillin and c) potassium clavulanate

Note: reproduced from the Kyoto Encyclopedia of Genes and Genomes pathway database (www.genome.jp/kegg/pathway.html).

Given its broad spectrum, proven use against common infections (e.g. urinary tract, skin and chest infections), and availability of an oral formulation, amoxicillin-clavulanate is currently the most commonly used antibiotic in Europe ²⁰¹. As with all antibiotics, resistance to amoxicillin-clavulanate is due to one or more of three mechanisms: hydrolysis of the active enzyme (beta-lactamases), modification of the target substrate (modification of PBPs), and reduced access of the antibiotic to its active site (reduced outer membrane permeability/increased antibiotic efflux) ⁵⁷. Further, as it is a combination antibiotic, each of these could occur for amoxicillin, clavulanate or both of them.

Identifying the limitations of the initial WGS-based prediction

The WGS-based approach for beta-lactam resistance prediction used in Chapter 4 only considered the presence/absence of transmissible beta-lactamase genes. Compared to other Gram-negative bacteria, *E. coli* is a relatively straightforward organism, with only a single non-inducible minimally-expressed chromosomal beta-lactamase (AmpC) ^{64,214} and exceptionally rare naturally occurring penicillin binding protein (PBP) mutations conferring beta-lactamase resistance ⁹⁹. Nevertheless, this basic WGS approach to resistance prediction still fails to account for several important mechanisms. For example, decreased permeability through porin loss and increased drug efflux have long been known to confer resistance to a wide range of beta-lactams in *E. coli* ^{215–218}. This, however, was unlikely to be the major cause of discrepancy, as it would be expected to affect agreement for a wide range antibiotics, contrary to the observed results in which extremely poor prediction was generally restricted to BL/BLI combinations ^{92,94}.

A more likely cause of poor prediction performance is the interpretation of how beta-lactamase presence affects an isolate's predicted antibiogram. To explain this, it is helpful to think about the biological interpretation of the resistance prediction algorithm. In the initial prediction, beta-lactamases, if present, were presumed to have fixed binary effects (i.e. the presence/absence of beta-lactamases alone would be sufficient to predict all isolates as susceptible or resistant). Biologically, this model assumes that for a given beta-lactam antibiotic, either the bacteria produces beta-lactamases that efficiently degrade all the antibiotic to which the bacteria is exposed (making it resistant), or it does not (making it susceptible). This model of resistance worked well for prediction of resistance to many beta-lactam antibiotics, notably ampicillin, the narrowest spectrum beta-lactam which acts as a proxy for any versus no beta-lactam resistance in isolates. However, BL/BLI combination are well-known to be more complex, with relative quantity of antibiotic and enzyme at the antibiotic's site of action playing an important role in resistance ²¹⁹ (Figure 50).

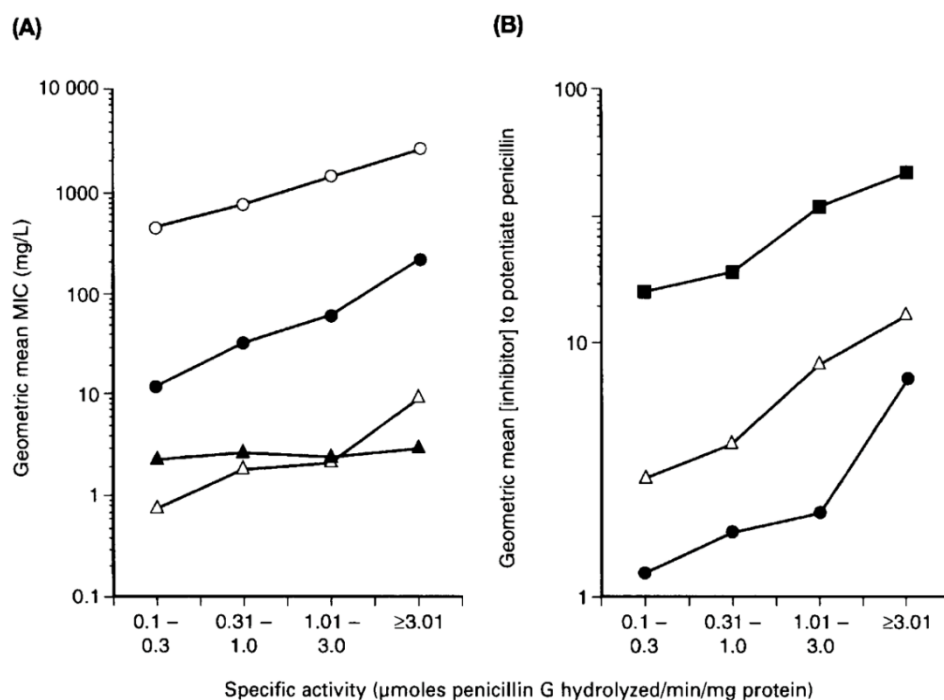


Figure 50: Impact of quantity of TEM-1 beta-lactamase on levels of resistance in 36 *E. coli*

Note: Reproduced from Livermore et al ²¹⁹. The isolates are grouped according to beta-lactamase specific activity. (A) The geometric mean MICs of: ○, ampicillin; ●, piperacillin; △, cefoperazone; and ▲, cefuroxime. (B) The mean concentrations required to potentiate penicillin MICs to breakpoint as follows: ■, sulbactam to reduce ampicillin MICs to 8 mg/L; △, clavulanate to reduce amoxicillin MICs to 8 mg/L; and ●, tazobactam to reduce piperacillin MICs to 16 mg/L.

Hence, increased expression of beta-lactamase (both chromosomal and plasmid-borne) through promoter mutations ⁷¹, gene copy number effects ²²⁰ and the inclusion of multiple different beta-lactamases per organism ²²¹ can all increase resistance to beta-lactam antibiotics and increase the amount of beta-lactamase inhibitor required to inhibit them. When combined with reduced permeability, these effects can become so pronounced as to make isolates containing only *bla*_{TEM-1} resistant to 4th generation cephalosporins ⁹⁶.

Consistent with the expectation that these effects may account for much of the missed resistance, previous studies on similar sets of isolates ^{222,223} have suggested that much of the amoxicillin-clavulanate resistance in *E. coli* is due to hyper-production of penicillinases and hyper-expression of *E.coli*'s chromosomal ampC beta-lactamase. However, previous studies of these effects, such those by as Ortega et al ²²² and Beceiro et al ⁹⁶, have relied on combinations of laboratory techniques such as phenotypic assays, quantitative polymerase chain reaction and sodium dodecyl sulphate–polyacrylamide gel electrophoresis to identify enzyme hyper-expression and reduced isolate permeability. Techniques such as these, while more definitively characterising isolates, are complex and time consuming and are often performed only on small numbers of isolates with abnormal/unexpected phenotypes. They are therefore not practical for assessing the causes of resistance in larger datasets, and hence a WGS-based method to identify these effects rapidly on large datasets would be highly desirable. However, only small numbers of studies have attempted even basic prediction of porin disruption using WGS ^{158,217}. Further, to my knowledge, to date none have attempted to quantify increased expression of transmissible beta-lactamases.

6.2. Methods

6.2.1. Sample baseline selection, antimicrobial susceptibility, and sequencing data

The OUH dataset (see section 3.2.1) was used to investigate the impact of modifying the resistance prediction algorithm for amoxicillin-clavulanate, with baseline sequencing and phenotyping as outlined in Chapter 3. Using this set had advantages both logistically and analytically. Logistically, it had complete amoxicillin-clavulanate phenotype data, the

originally cultured isolates were available locally and isolate sequencing was complete at the outset of this work. Analytically, the sampling frame for this dataset was not only the most representative of the aim of this thesis, it was also, because of the exhaustive nature of its sampling, likely less biased towards either resistance (PHE) or carriage (APHA) than the other two datasets in this thesis.

6.2.2. Resistance genotyping

In order to produce a more realistic model of beta-lactam resistance, searches for several new features were added to the initial resistance genotype method to create an “extended” genotype. In addition to the beta-lactamases, this “extended” approach additionally identified

- DNA copy number for each transmissible beta-lactamase identified
- *bla*_{TEM} promoter mutations likely to cause increased *bla*_{TEM} expression
- *ampC* promoter mutations likely to cause increased chromosomal AmpC expression
- mutations in *ompF/ompC* (genes encoding the two predominant non-selective porin channels in *E. coli* ⁹² suggestive of gene loss of function.

In this chapter, both the presence/absence of beta-lactamases and the additional features above (the “extended” genotype) were identified using ARIBA ¹³⁸, a combination assembly/mapping approach (see section 5.2.1 for more detail). The reason for switching from ABRicate search of *de novo* assemblies was increasing concern that assembly errors were a significant shortcoming for resistance genotyping in Gram-negatives ¹³⁷. ARIBA was preferred over a mapping only method as it was widely believed its assembly component would more accurately depict indels and the fact that its assembly output could be easily

interrogated. Each feature identified by ARIBA was defined as present if its “length adjusted assembly depth of coverage” was greater than 10x.

$$\text{Length adjusted assembly depth of coverage} = \frac{\sum \text{Contig Coverage} \times \text{Contig Length}}{\text{Feature Reference Length}} \geq 10$$

ARIBA itself does not apply a coverage cut-off, instead giving the user a choice as to whether or not to only report “complete” matches, with complete being defined on the identity (>90%) and length (>90%) of the match. However, to avoid the need to assemble complete genes (which would reduce the benefit of switching to ARIBA over the initial ABRicate produced genotype), I opted to use non-complete matches, additionally applying a depth cut-off. The depth cut-off chosen was slightly stricter than used in other resistance genotype predictions (SRST2, default depth cut-off 5x) due to the fact there appeared to be a natural separation between in the distribution of length adjusted cut-offs at 10x (Figure 51), suggesting potential biological relevance to this threshold.

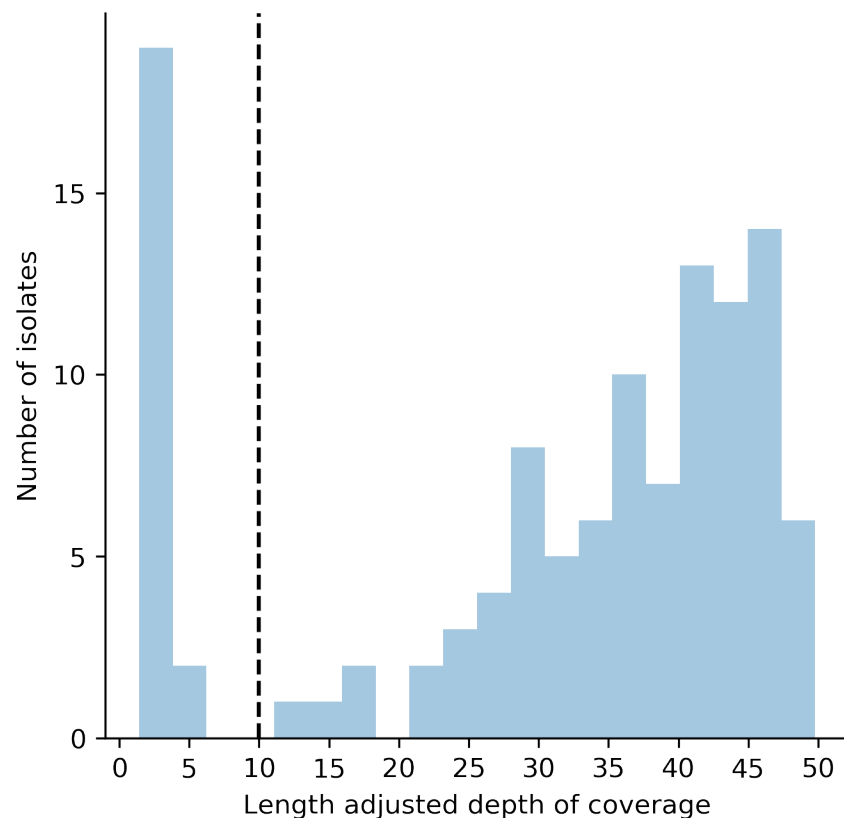


Figure 51: Length adjusted depth of coverage of transmissible beta-lactamases from ARIBA

In addition to potentially being a more robust method of resistance genotyping in Gram-negatives, ARIBA provides additional information on whether or not reads suggest the presence of heterozygous sequences (i.e. isolates containing two or more variants of a particular sequence) and assembly issues through its “flags”. This additional information was used for three purposes.

- Heterozygosity in the MLST alleles was used to identify and exclude potential mixed samples as an additional QC step. Through this eight additional samples were excluded.
- Disruption of porin assemblies was used in addition to porin mutations as a proxy for porin loss (see below). This was used to account for situations where disruption was

so severe that genes could no longer be accurately assembled to enable identify truncating mutations.

- To investigate potential discrepancies between WGS-predicted and observed phenotype, in case a low frequency variant may confer resistance for the isolate as a whole.

Of note however, additional information on heterozygotes was not used in the WGS-based prediction algorithm itself, given their significance for interpretation has, to my knowledge, not yet been established.

Beta-lactamases

Beta-lactamase sequences from the Resfinder ²⁸ database (accessed 21st August 2017) were identified from sequence data. Given that ARIBA, like ABRicate, uses nucleotide level matching, optimal sequence choice of matches was checked by using tBLASTn ¹³⁵ on ARIBA assemblies against a pre-translated database of beta-lactamases from the Resfinder database to avoid matching issues similar to those identified in section 4.3.2. For each beta-lactamase a “DNA copy number” metric was created, representing the ratio of gene depth of coverage versus an estimated mean of chromosomal depth of coverage. Gene depth of coverage was defined as the contig coverage depth of the longest ARIBA assembled contig. The estimate of chromosomal depth of coverage was obtained using mean gene depth of coverage of identified MLST alleles. DNA copy number was therefore defined as:

$$DNA\ copy\ number = gene\ depth\ of\ coverage \div \frac{\sum MLST\ allele\ depth\ of\ coverage}{No\ MLST\ alleles\ found}$$

A DNA copy number of 1 thus suggests the gene is present in the same quantities as the MLST genes (i.e. 1 per cell) and a DNA copy number greater than 1 suggests the gene is

present in higher quantities than the MLST genes (i.e. >1 per cell, for example due to multiple copies on the chromosome or on plasmids). Further, given this represents an average across multiple cells taken for sequencing as part of a single colony pick, non-whole number DNA copy numbers were possible. For example a gene could have a DNA copy number of 2.5 if some cells in the colony contained two copies of it and others contained three copies.

Repeatability of this metric was tested by comparing DNA copy numbers of 46 elements from 9 isolates which were sequenced in duplicate as part of plate quality control (with colonies taken for extraction from the same re-cultured stock). Based on this analysis, standard deviation of the DNA copy number was 0.17 (95% CI 0.14, 0.21) across repeats.

bla_{TEM} promoters

Using ARIBA, each isolate's data were searched for sequences similar to the Lartigue *bla_{TEM}* P3 promoter²²⁴. Sequences found were then compared to each of the Lartigue promoter types, and the closest match was selected. Sequence position relative to the *bla_{TEM}* gene was identified where possible using BLASTn¹³⁵ on ARIBA assembly flanking sequence and isolate Velvet¹⁶⁸ *de-novo* assemblies.

Lartigue promoter type	Difference to P3 reference
P3	-
PA/PB	C37>T
PC/PD	36_44Del, G157>T
P4	G157>T
P5	C146>G

TEM Promoters

<i>bla</i> _{TEM} P3 promoter	1	TTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTATAGGTTAATGTCATGATAATAA	63
	64	TGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAATGTGCGCGGAACCCCTATTGTTTA	126
	127	TTTTTCTAAATACAT TTCAA TATGTATCCGCTCATG GACA ATAACCCCTGGTAAATGCTTCA	189
	190	ATAATATTGAAAAAGGAAGAGT TATG	

-35 box -10 box

*bla*_{TEM} start

Figure 52: The Lartigue P3 promoter²²⁴, and sequence differences of other named *bla*_{TEM} promoters

ampC and its promoter

In *E. coli*, the mechanism of regulation of its chromosomal *ampC* gene is fairly straightforward, with its expression normally being repressed and non-inducible except in the context of promoter/attenuator mutations immediately upstream of the gene causing de-repression and leading to constitutive production of high levels of AmpC^{70,225}. To identify these mutations, each isolate's data was searched for sequences similar to the *ampC* gene and promoter present in *E. coli* ATCC 25922 (a known non-hyperproducer *E. coli*, commonly used as laboratory control strain) (Figure 53). Position of promoter sequences next to the *ampC* gene were confirmed where possible, as above, and promoter sequences found were inspected for variants, including but not restricted to those known to increase chromosomal *ampC* expression in *E. coli*²²⁶.

<i>ampC</i> promoter mutations treated as significant
-42 C>T
-32 T>A
+24 C>A
Any insertion between -35 and -10 promoter boxes

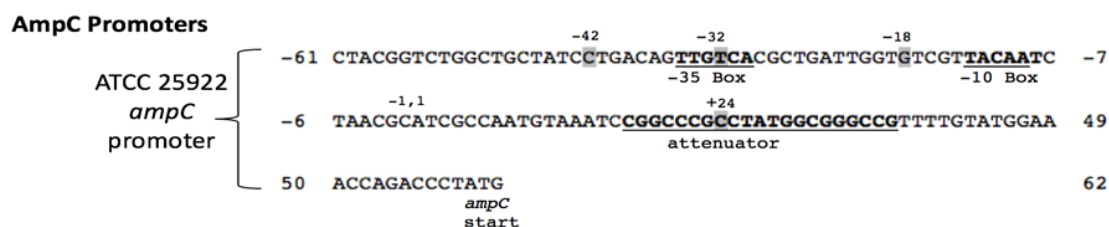


Figure 53: *ampC* promoter region and sequence mutations thought to cause constitutive AmpC hyper-expression

Porin disruption

To identify porin disruption, isolate data was searched using ARIBA for sequences similar to the porin genes *ompF* and *ompC* found in *E. coli* K-12 (RefSeq: NC_000913.3) (percentage identity cut-off 90% (default)). In cases where genes were successfully assembled by ARIBA (i.e. ARIBA flags being “unique contig”, “complete gene found”, and “assembled into one contig”), porin sequences were inspected for truncations, insertions, and deletions likely to cause loss of function. Where genes were not successfully assembled (lacking any of the successful assembly flags, or containing any of the mis-assembly flags “hit both strands”, “region assembled twice”, “scaffold graph bad”), BLASTn search of *de-novo* assemblies for porin genes was performed, to identify possible insertion sequences.

5.2.3. Resistance prediction from genotype

For each beta-lactamase, and all features in the extended genotype, the association between genotype and phenotype was examined to assess their validity for inclusion in resistance prediction. Following this, features were used to make two resistance

predictions. A basic resistance prediction was analogous to the initial resistance prediction used in section 4.2.1 in which only beta-lactamases were considered; an extended prediction used all valid features from the “extended” genotype as described above. These two predictions were compared to assess whether the additional genetic information improved WGS-based prediction of amoxicillin-clavulanate resistance.

The basic prediction identified an isolate as resistant if it contained:

1. Any Ambler class C or D beta-lactamase gene (e.g. *bla*_{OXA-1}, *bla*_{CMY-2})
2. Any inhibitor resistant Ambler class A beta-lactamase (e.g. *bla*_{TEM-30}, *bla*_{SHV-10})

(Note that this set of isolates contained no Ambler class B (metallo-)beta-lactamases.)

The extended prediction identified an isolate as resistant if it contained

1. Any Ambler class C or D beta-lactamase gene (e.g. *bla*_{OXA-1}, *bla*_{CMY-2})
2. Any inhibitor resistant Ambler class A beta-lactamase (e.g. *bla*_{TEM-30}, *bla*_{SHV-10})
3. Any beta-lactamase present at “high” DNA copy number (see section 6.3.2., DNA copy number)
4. Any mutation suggestive of ampC hyper-expression (Figure 53)
5. A *bla*_{TEM} gene with a P4, P5, PA/PB, or PC/PD promoter (Figure 52)
6. Any beta-lactamase together with evidence of porin gene disruption (either *ompF* or *ompC*) in the isolate. Porin deficiency alone was not considered to cause amoxicillin-clavulanate resistance given most studies investigating have found this mechanism involved as part of multi-mechanism resistance^{92,216}.

6.3. Results

6.3.1. Samples and phenotypes

With the additional quality control step used in this study (samples excluded if ARIBA identified heterozygous MLST alleles), the sample size of the Oxford dataset was reduced by eight, from 984 used in the initial prediction (Chapter 4) to 976 used in this chapter. The observed beta-lactam antibiogram (BD Pheonix) of these isolates is summarised in Table 20.

Table 20: AST beta-lactam antibiogram

	Number (%) of resistant isolates	
	Overall	
Antibiotic	All isolates (N=976)	Amoxicillin-clavulanate resistant (N=340)
Ampicillin	531 (54%)	333 (98%)
Amoxicillin-clavulanate	340 (35%)	340 (100%)
Piperacillin-tazobactam	50 (5%)	50 (15%)
Cefoxitin *	31 (3%)	29(9%)
Ceftazidime	53 (5%)	42 (12%)
Ceftriaxone	79 (8%)	53 (16%)
Cefepime	57 (7%)	44 (13%)
Ertapenem	3 (0.3%)	2 (0.2%)
Meropenem	0 (0%)	0 (0%)
Aztreonam	61 (6%)	44 (13%)

* Number of isolates positive for ampC production screen

6.3.2. Examining features of the resistance genotype

Beta-lactamases

530/976 (54%) of samples contained at least one non-chromosomal beta-lactamase gene, with 67 containing beta-lactamases from multiple families (55 containing two and 12 containing three) (Table 21). *bla*_{TEM-1} was by far the commonest beta-lactamase gene seen occurring in 409/976 (42%) isolates, with two next most common being *bla*_{OXA-1} and *bla*_{CTX-M-15}, occurring in 60 (6%) and 52 (5%) isolates respectively. All other beta-lactamase variants

were found in 20 or fewer isolates. Aside for *bla*_{TEM-1}, association between presence of these beta-lactamases (when in isolation) and amoxicillin-clavulanate resistance generally followed Bush-Jacoby classifications (Table 21).

In addition to the known beta-lactamases, this set also contained three isolates containing novel beta-lactamases, and 6 isolates with incomplete “unknown” beta-lactamases. The novel variants consisted of two *bla*_{CTX-M-1} like enzymes with Ser130Gly (Ambler position) mutations and one *bla*_{TEM} with a Ser243Ala mutation. All three of these were amoxicillin-clavulanate resistant. While this is expected for the *bla*_{CTX-M} enzymes given how this mutation is expected to modify the active site of the enzyme ⁶⁵, the causality of the Ser243Ala *bla*_{TEM} mutation is less clear. However, proximity to a known important mutation may be a factor. Several inhibitor resistant *bla*_{TEM} have been described ²²⁷ with arginine 244 mutations, with this site being suspected to be involved in breaking the clavulanate five-member ring (Figure 49), one of the steps of reaction between the acyl-enzyme intermediate and the enzymes eventual inactivation ²²⁸.

The unknown genes (where ARIBA assembly failed or only produced partial genes) consisted of three suspected *bla*_{TEM} and three suspected *bla*_{CMY}. Of note, all of these genes were fragmented in the SPAdes assemblies used in Chapter 4. The significance of these unknown variants for resistance prediction is unclear, as demonstrated by the fact that only 2/3 of the unknown *bla*_{CMY} containing isolates were amoxicillin-clavulanate resistant, even though these are class C beta-lactamases, which can hydrolyse amoxicillin and are invariably resistant to clavulanate inhibition.

Table 21: Beta-lactamases identified

Feature	Mutations (Ambler positions)	Number of isolates with mechanism (total sequenced [n=976])	Number of isolates with only this mechanism (total sequenced [n=976])	Number of isolates amoxicillin-clavulanate resistant by BD Phoenix	
				Total (% of no with this mechanism)	Not beta- lactam resistant by any other ^o genetic mechanism
<i>bla</i>_{TEM} beta-lactamases					
TEM-1		409	312	244 (60%)	158 (50%)
TEM-1-like*	Ser243Ala	1	1	1 (100%)	1 (100%)
TEM-30		4	3	4(100%)	3 (100%)
TEM-31		2	0	1 (100%)	-
TEM-33		1	1	2(100%)	1 (100%)
TEM-40		5	2	5(100%)	2 (100%)
TEM-135		1	1	1 (100%)	1 (100%)
TEM-190		1	1	1 (100%)	1 (100%)
TEM-unknown*		3	3	2 (67%)	2 (67%)
<i>bla</i>_{CTX-M-1} group beta-lactamases					
CTX-M-1		3	3	0 (0%)	0 (0%)
CTX-M-15		52	9	40 (77%)	0 (0%)
CTX-M-15-like*	Ser130Gly	1	1	1 (100%)	1 (100%)
CTX-M-3		1	0	1 (100%)	- ⁺
CTX-M-55		1	0	1 (100%)	- ⁺
CTX-M-55-like*	Ser130Gly	1	1	1 (100%)	1 (100%)
<i>bla</i>_{CTX-M-9} group beta-lactamases					
CTX-M-9		1	0	1 (100%)	- ⁺
CTX-M-14b		9	1	5 (56%)	1 (20%)
CTX-M-27		4	3	2 (50%)	0 (0%)
Other beta-lactamases identified					
OXA-1		60	17	60 (100%)	17 (100%)
OXA-48		1	0	1 (100%)	- ⁺
SHV-11		20	18	6 (30%)	4 (22%)
SHV-12		3	1	1 (33%)	0 (0%)
CMY-2		5	1	5 (100%)	1 (100%)
CMY-60		1	1	0 (0%)	0 (0%)
OXA-unknown*		1	1	0 (0%)	0 (0%)
CMY-unknown*		2	1	1 (50%)	0 (0%)
LAT-unknown*		1	0	1 (100%)	- ⁺

o: Other beta-lactam resistance mechanisms includes any other beta-lactamase, ampC promoter mutations suggestive of ampC hyper production and functional porin loss

-: no isolates in this category

*: "like" indicates inexact match with parent reference sequence, "unknown" indicates inability to identify a complete gene (ARIBA flag), note for unknown mutations not presented

Note: Orange colouring indicates those expected to cause resistance by their Bush-Jacoby classification.

DNA copy number

Mean copy number of transmissible beta-lactamases was 2.23 (IQR 1.73,3.31), and this was relatively constant irrespective of beta-lactamase family (Figure 54).

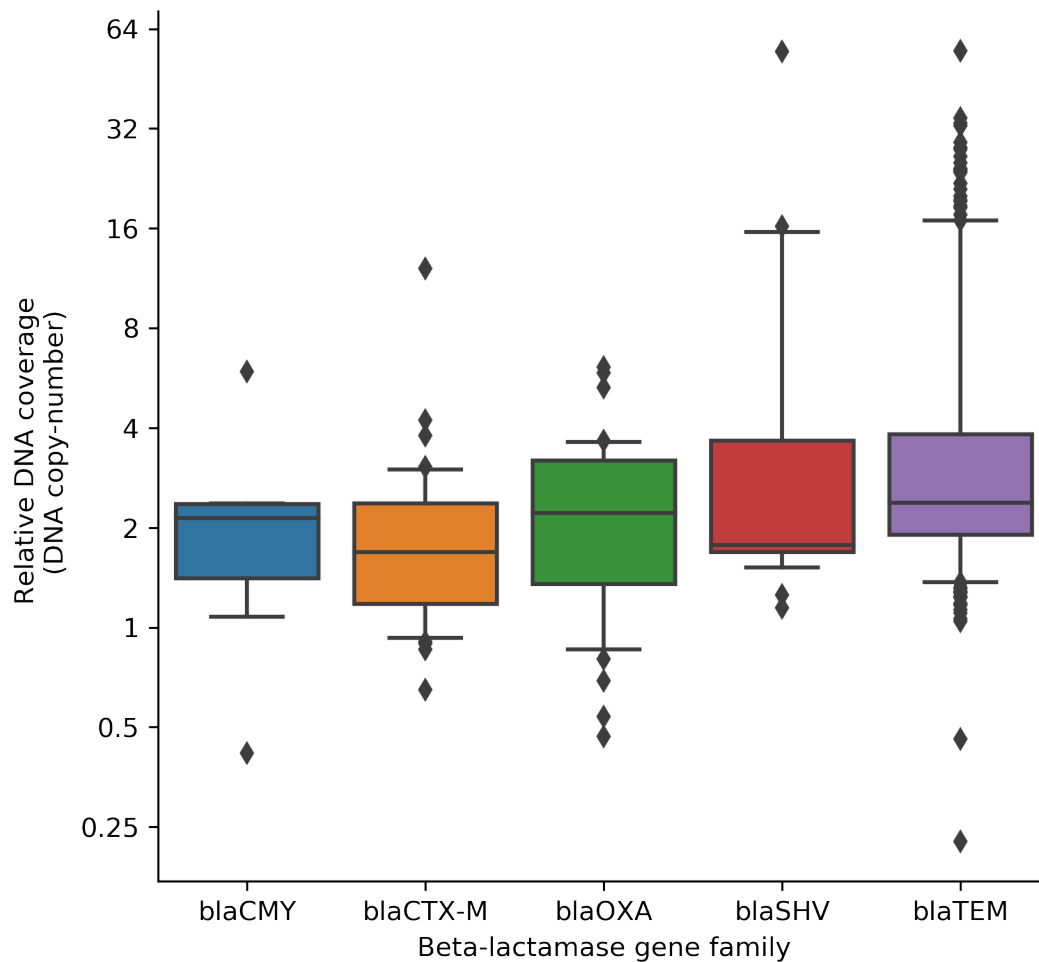


Figure 54: Beta-lactamase copy numbers

Note: Boxes showing interquartile ranges and whiskers showing 5th and 95th centiles.

Colours only used to help distinguish boxes and have no intrinsic meaning.

Comparing beta-lactamase copy number metric with observed MIC in isolates with only a single, non-inhibitor resistant beta-lactamase identified, demonstrated a strong positive correlation between the two (Figure 55). Further, beta-lactamases with gene copy number >6 were only found in amoxicillin-clavulanate resistant isolates (Figure 56).

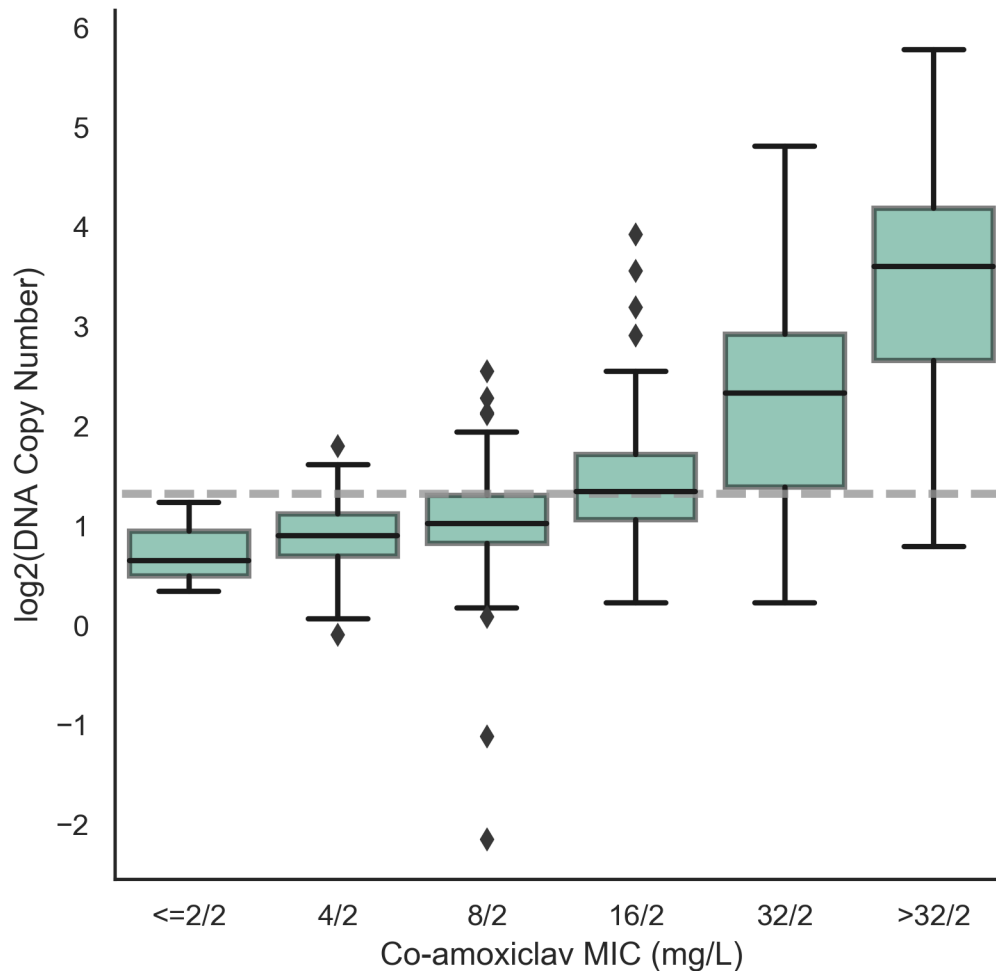


Figure 55: Association between copy number and amoxicillin-clavulanate MIC

Note: 328 isolates with only one beta-lactamase identified.: 294 had *bla*_{TEM} genes (290 with *bla*_{TEM-1}, 4 with other non-inhibitor resistant *bla*_{TEM} genes), 19 had non-inhibitor resistant *bla*_{SHV} genes and 15 had *bla*_{CTX-M} genes. Grey line indicates the copy number 2.5. This value was selected (by ROC analysis) to classify isolates as resistant in the “extended” resistance prediction algorithm.

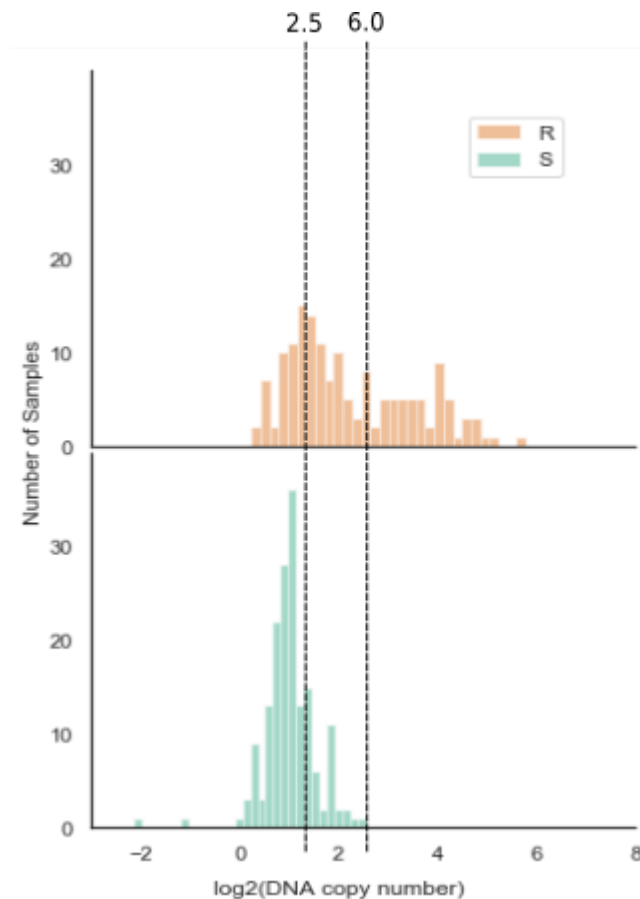


Figure 56: Copy numbers in amoxicillin-clavulanate resistant and susceptible isolates

Note: Isolates only containing one non-inhibitor resistant beta-lactamase and no other mechanisms of amoxicillin-clavulanate resistance. Orange: resistant, green: susceptible.

While the strong association between copy-number and MIC made it an important inclusion in the extended resistance prediction algorithm, there was no clear threshold which separated susceptible and resistant samples. Consequently, a receiver operator curve (ROC) analysis was used to select the threshold by which copy numbers were classified into low (not to causing resistance) or high (causing resistance categories) categories. Specifically this threshold was created by comparing copy numbers of lone beta-lactamases (i.e. no other beta-lactamases in the isolate) with observed amoxicillin-clavulanate resistance (Figure 57). All tested metrics suggested an optimal cut-off between 2.3-2.4 (with all 95% confidence intervals contained in the range 1.7-3.0). A cut-off of 2.5 was chosen for ease of interpretation. Expressing this in biological terms, a gene having high DNA copy number

(greater than 2.5) means that the mean number of copies of this gene that each cell contains is greater than 2.5. 237/594 (40%) of the beta-lactamases had DNA copy number >2.5.

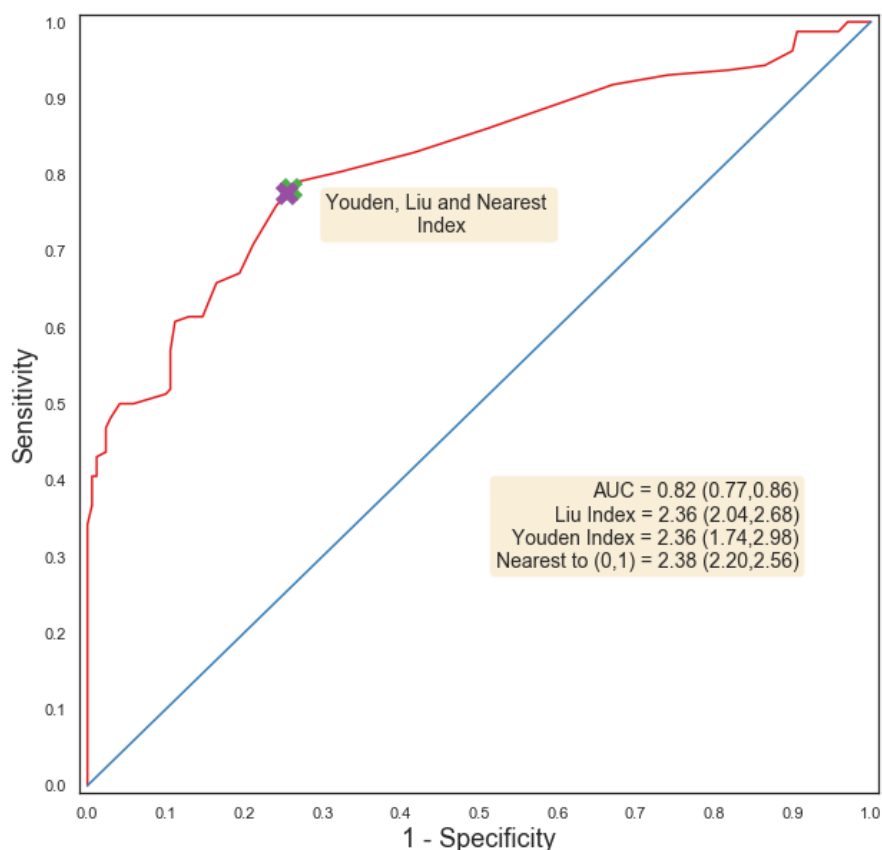


Figure 57: ROC curve comparing DNA copy number with resistance prediction

bla_{TEM} promoters

Promoter regions similar to the Lartigue *bla_{TEM}* promoter sequences were found in 401 (94%) of the 427 isolates which contained a *bla_{TEM}* gene and were never found without a concomitant *bla_{TEM}* gene. P3 was the commonest, occurring in 352/401 (88%) promotor-containing isolates, and the commonest promoter which would lead to increased *bla_{TEM}* expression was Pa/Pb, occurring in 41 (10%) isolates (Table 22). All but one of the 49 (12%)

isolates containing non-P3 like promoters were amoxicillin-clavulanate resistant, consistent with these isolates producing greater quantities of TEM beta-lactamase.

Table 22: blaTEM promoters, Lartigue types and amoxicillin-clavulanate resistance

Mutation pattern	Number of isolates with mutation pattern (total isolates with blaTEM [n=427])	Closest promoter type^x	Number of isolates co-amoxiclav resistant by BD Phoenix (% of total with this mechanism)
None (vs P3)	254	P3	122 (48%)
175 G>A	96	P3	68 (71%)
32 C>T	26	Pa/Pb	26 (100%)
32 C>T 175 G>A	15	Pa/Pb	15 (100%)
162 G>T	4	P4	4 (100%)
162 G>T 175 G>A	2	P4	1 (50%)
43 G>A 175 G>A	1	P3	0 (0%)
30_38Del	1	Pc/Pd	1 (100%)
30_38Del 175 G>A	1	Pc/Pd	1 (100%)
65 G>A 175 G>A	1	P3	0 (0%)
Unable to assemble	26	NA	13 (50%)

x: For promoter type and the effects of mutations, see (12).

For 22 of the 26 isolates where ARIBA found no sequences matching the *bla*_{TEM} P3 promoter or was unable to assemble the promoter, the assembly contig containing *bla*_{TEM} only had <250 base pairs (bp) of sequence upstream from the gene. For all of these, the first 50-150 bases upstream of this matched the P3 promoter, but the rest of the available upstream sequence, if any, matched parts of known transposases, suggesting insertion sequences may account for issues finding full promoter sequences. While for each of these, the traditional P3 promoter -35 and -10 boxes were still contained in the sequence upstream of the *bla*_{TEM} gene, whether these isolates contained further strong promoters is uncertain. Looking at the remaining four isolates, one contained a *bla*_{TEM} P3 like promoter at <10x coverage, and one contained a partial *bla*_{TEM} promoter together with a long section of sequence (120kbp) which high percentage identity agreement with a subsection of the circularised *E. coli*

genome in NCBI (CU928163.2). Complete *bla*_{TEM} genes could not be identified for both of the remaining 2 isolates, precluding the identification of upstream sequence.

ampC promoters

Isolate *ampC* promoters were found and matched to or were close variants of the *E. coli* k-12 *ampC* promoter in all but three isolates, in which only subsections of the promoter were found. However, there were a large number of variants of this promoter sequence among these matches, even in a relatively short (<100bp) section of sequence immediately upstream of the *ampC* gene (Table 23). Further, many of these variants were rare, with 32 (84%) of the 38 different variants being seen in less than 10 isolates. 20 isolates contained mutations which previous literature^{71,226} would suggest have a depressed chromosomal *ampC* gene. All 20 isolates were amoxicillin-clavulanate resistant and hence these mutations were included as part of the “extended” resistance prediction algorithm. Interestingly, only 13/20 had cefoxitin MIC > 8mg/L (a phenotypic screening test used to identify class C cephalosporinases, see Chapter 4). Further 18/31 isolates with cefoxitin MIC > 8mg/L had no significant mutation identified in their *ampC* promoter, although this test has known poor specificity¹¹⁶. However, this may represent a breakpoint issue, with the other 7/20 isolates with chromosomal *ampC* promoter mutations having cefoxitin MICs equal to 8mg/L. Hence, having a “significant” mutation was strongly associated with cefoxitin MIC being ≥ 8mg/L ($p < 0.0001$, Fisher’s exact test). Interestingly, two of the three isolates which only contained incomplete *E. coli* k-12 *ampC* promoters upstream of *ampC* were amoxicillin-clavulanate resistant and had cefoxitin MICs > 8mg/L in the absence of any other beta-lactamase. In addition to the known important promoter mutations, which were invariably amoxicillin-clavulanate resistant, there were several other mutations seen in amoxicillin-clavulanate

resistant isolates without any other cause of beta-lactam resistance. While these may lead to *ampC* de-repression, they were not included in the “extended” resistance prediction algorithm as none of these mutations were found in more than a single isolate.

Table 23: ampC promoter mutations and amoxicillin-clavulanate resistance

Mutation pattern*	Number of isolates with mutation pattern (total sequenced [n=976])	Number of isolates with mutation pattern and no beta-lactamases	Number of isolates co-amoxiclav resistant by BD Phoenix (% of total with this mechanism)	
			Total	Not beta-lactam resistant by any other mechanism ^o
Nil	368	156	120 (33%)	0 (0%)
-28 G>A	168	55	83 (49%)	0 (0%)
-28 G>A 17 C>T	168	93	39 (23%)	1 (1%)
58 C>T	125	74	36 (29%)	0 (0%)
-18 G>A* -1 C>T 58 C>T	69	32	21 (45%)	1 (3%)
22 C>T 26 T>G 27 A>T 32 G>A	17	7	9 (53%)	1 (14%)
-28 G>A 58 C>T	9	3	1 (11%)	0 (0%)
48 A>G 58 C>T	4	4	0 (0%)	0 (0%)
-42 C>T* -18 G>A -1 C>T 58 C>T	3	1	3 (100%)	1 (100%)
-18 G>A -11 C>T -1 C>T 58 C>T	3	1	3 (100%)	1 (100%)
37 G>T	3	2	1 (33%)	0 (0%)
-32 T>A	3	3	3 (100%)	3 (100%)
-32 T>A 58 C>T	3	0	3 (100%)	-
-28 G>A -24_-23InsT 17 C>T	2	2	2 (100%)	2 (100%)
34_35InsC 58 C>T	2	2	0 (0%)	0 (0%)
-28 G>A -15_-14InsT 17 C>T	2	2	2 (100%)	2 (100%)
-28 G>A -10_1Del 17 C>T	2	2	0 (0%)	0 (0%)
33 G>A	2	2	0 (0%)	0 (0%)
-15_-14InsG	1	1	1 (100%)	1 (100%)
-12_-5Del 58 C>T	1	1	1 (100%)	1 (100%)
-12_-5Del	1	1	0 (0%)	0 (0%)
-9_-8Del 58 C>T	1	1	0 (0%)	0 (0%)
-11 C>T	1	1	1 (100%)	1 (100%)
-29 C>T	1	1	0 (0%)	0 (0%)
-42 C>T -18 G>A -1 C>T 20_21InsG 58 C>T	1	1	1 (100%)	1 (100%)
20DelC	1	1	0 (0%)	0 (0%)
-28 G>A -15_-14InsGT 17 C>T 41 T>A	1	1	1 (100%)	1 (100%)
-28 G>A -12_-5Del 17 C>T	1	1	0 (0%)	0 (0%)
-28 G>A -16_-15InsC	1	1	1 (100%)	1 (100%)
-28 G>A -9_2Del 17 C>T	1	1	0 (0%)	0 (0%)
-28 G>A 17 C>T 20 C>T	1	1	0 (0%)	0 (0%)
-28 G>A 33 G>T	1	1	1 (100%)	1 (100%)
-28 G>A 34 G>A 58 C>T	1	0	1 (100%)	-
-28 G>A 26-28Del	1	1	1 (100%)	1 (100%)
30 G>A	1	0	0 (0%)	0 (0%)
-32 T>A -28 G>A	1	1	1 (100%)	1 (100%)
-32 T>A -28 G>A -21_-20InsT -11 C>T 17 C>T	1	1	1 (100%)	1 (100%)
-32 T>A -28 G>A 17 C>T	1	0	1 (100%)	-
Unable to assemble	3	3	2 (66%)	2 (66%)

^o: Other beta-lactam resistance mechanisms includes any other beta-lactamase, ampC promoter mutations suggestive of ampC hyper production and functional porin loss, see section 6.3.2.

-: no isolates in this category

*: Letters in bold represent known significant mutations

Truncated porin genes

31 (3%) isolates potentially had one non-functional porin, with 16 and 15 isolates having “functionally lost” *ompF* and *ompC* respectively. However, no isolate had lost both *ompC* and *ompF*. Frameshift mutations were the most common cause of disruption, occurring in 12/16 isolates with disrupted *ompF* genes (Table 24) and 6/15 isolates with disrupted *ompC* genes (Table 25). The remaining “functionally lost” porin genes were either not identified by ARIBA (8/13 cases) or had assembly issues (5/13 cases), and hence were assumed incomplete/functionally non-viable. However, highlighting how differences in bioinformatic approach can alter results, complete coding porin genes were found at less than 10x coverage for 8/13 cases, and for the remaining 5, porin genes were again found a low coverage, but the assembly sequence suggested multiple early frameshift mutations leading to truncation.

Due to the frequent co-occurrence of other beta-lactam resistance mechanisms (in 22/31 cases), the impact of porin disruption on phenotype was less clear. The nine isolates with porin disruption alone were universally susceptible. However, 20 (91%) of the other 22 isolates with porin disruption had other mechanisms capable of producing amoxicillin-clavulanate resistance, meaning the theory that porin loss adds to resistance caused by non-inhibitor resistant beta-lactamases could only be investigated in two isolates.

Table 24: ompF disruption mechanism, concomitant beta-lactam resistance features and beta-lactam antibiogram

Isolate	Porin gene issue ^o	Other beta-lactam mechanisms	Minimum inhibitory concentration (mg/L)						
			Ampicillin	Amoxicillin-clavulanate	Piperacillin-tazobactam	Cefoxitin	Ceftriaxone	Ceftazidime	Meropenem
OMPF1	truncation	De-repressed chromosomal <i>ampC</i> , <i>bla</i> _{TEM-1}	>8	>32/2	8/4	>16	1	8	≤0.5
OMPF2	truncation	De-repressed chromosomal <i>ampC</i>	>8	>32/2	≤4/4	>16	1	4	≤0.5
OMPF3	truncation	<i>bla</i> _{CTX-M-15} , <i>bla</i> _{OXA-1}	>8	>32/2	32/4	≤4	>4	>16	≤0.5
OMPF4	truncation	<i>bla</i> _{OXA-1} , <i>bla</i> _{TEM-1}	>8	>32/2	>64/4	≤4	≤1	≤1	≤0.5
OMPF5	truncation	<i>bla</i> _{OXA-1}	>8	>32/2	16	16	≤1	≤1	≤0.5
OMPF6	truncation	<i>bla</i> _{TEM-30}	>8	>32/2	≤4/4	8	≤1	≤1	≤0.5
OMPF7	truncation	High copy number <i>bla</i> _{TEM-1}	>8	8/2	≤4/4	≤4	≤1	≤1	≤0.5
OMPF8	truncation	High copy number <i>bla</i> _{TEM-1}	>8	32/2	≤4/4	16	≤1	≤1	≤0.5
OMPF9	truncation	High copy number <i>bla</i> _{TEM-1}	>8	8/2	≤4/4	≤4	≤1	≤1	≤0.5
OMPF10	truncation	-	≤2	4/2	≤4/4	≤4	≤1	≤1	≤0.5
OMPF11	truncation	-	≤2	4/2	≤4/4	≤4	≤1	≤1	≤0.5
OMPF12	truncation	-	4	8/2	≤4/4	>16	≤1	≤1	≤0.5
OMPF13	quality	<i>bla</i> _{CTX-M-15} , high copy number <i>bla</i> _{TEM-1}	>8	8/2	≤4/4	≤4	>4	8	≤0.5
OMPF14	quality	High copy number <i>bla</i> _{TEM-1}	>8	>32/2	≤4/4	8	≤1	≤1	≤0.5
OMPF15	quality	High copy number <i>bla</i> _{TEM-1}	>8	16/2	≤4/4	≤4	≤1	≤1	≤0.5
OMPF16	quality	High copy number <i>bla</i> _{TEM-1}	>8	32/2	≤4/4	8	≤1	≤1	≤0.5

*: Isolates predicted amoxicillin-clavulanate resistant due to non-functional porin in the presence of (any) beta-lactamase

o: Truncations occur when frameshift mutations leading to premature stop codons are found in the sequence. Quality issues are when the search method (ARIBA) failed to completely and uniquely assemble the gene (see Methods) or the gene was only found at <10x coverage.

Table 25: ompC disruption mechanism, concomitant beta-lactam resistance features and beta-lactam antibiogram

Isolate	Porin gene issue ^o	Other beta-lactam mechanisms	Minimum inhibitory concentration (mg/L)						
			Ampicillin	Amoxicillin-clavulanate	Piperacillin-tazobactam	Cefoxitin	Ceftriaxone	Ceftazidime	Meropenem
OMPC1	truncation	<i>bla</i> _{CTX-M-15} , <i>bla</i> _{OXA-1} , high copy number <i>bla</i> _{TEM-1}	>8	32/2	<=4/4	8	>4	1	<=0.5
OMPC2	truncation	High copy number <i>bla</i> _{TEM-1}	>8	>32/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC3*	truncation	<i>bla</i> _{TEM-1}	>8	>32/2	<=4/4	<=4	<=1	<=1	
OMPC4	truncation	-	4	8/2	<=4/4	8	<=1	<=1	<=0.5
OMPC5	truncation	-	<=2	4/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC6	truncation	-	<=2	<=2/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC7	quality	<i>bla</i> _{OXA-1}	>8	>32/2	64/4	8	<=1	1	<=0.5
OMPC8	quality	High copy number <i>bla</i> _{TEM-1} with a non-P3 <i>bla</i> _{TEM} promoter	>8	>32/2	>64/4	>16	1	2	
OMPC9	quality	High copy number <i>bla</i> _{TEM-1}	>8	>32/2	<=4/4	<=4	<=0.5	<=0.5	<=0.5
OMPC10	quality	High copy number <i>bla</i> _{TEM-1}	>8	>32/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC11	quality	High copy number <i>bla</i> _{TEM-1}	>8	16/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC12*	quality	<i>bla</i> _{TEM-1}	>8	16/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC13	quality	-	<=2	<=2/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC14	quality	-	<=2	<=2/2	<=4/4	<=4	<=1	<=1	<=0.5
OMPC15	quality	-	<=2	<=2/2	<=4/4	<=4	<=1	<=1	<=0.5

*: Isolates predicted amoxicillin-clavulanate resistant due to non-functional porin in the presence of (any) beta-lactamase

o: Truncations occur when frameshift mutations leading to premature stop codons are found in the sequence. Quality issues are when the search method (ARIBA) failed to completely and uniquely assemble the gene (see Methods).

6.3.3. Evaluating the impact of the extended features on resistance prediction

The positive predictive value for each of these rules was generally very high (>94%) except for DNA copy number (PPV 77%) (Table 26), which, as noted in section 6.3.2, likely reflects the lack of a perfect threshold separating low and high copy number isolates. Having demonstrated an association between each of the features identified and amoxicillin-clavulanate resistance, they were combined to generate a new extended prediction algorithm as detailed in section 6.2.3. Using these features, the extended algorithm predicted far more isolates as resistant (323/976 (33%)), increasing from only 81/976 (8%) predicted by the basic algorithm (used in the initial prediction in Chapter 4).

Table 26: Positive predictive value of individual genetic features for predicting amoxicillin-clavulanate resistance

Feature	Prediction	Number of isolates resistant by routine AST/total isolates with this feature (%)	
		PPV	PPV restricted to isolates not predicted as amoxicillin-clavulanate resistant by any other feature
Beta-lactamases			
1 Any class C or D serine beta-lactamase	Basic	66/69 (96%)	32/34 (94%)
2 Any inhibitor resistant class A beta-lactamase	Basic	12/12 (100%)	2/2 (100%)
Promoter mutations			
3 Non-P3 <i>bla</i> _{TEM} promoter associated with <i>bla</i> _{TEM} hyper-production	Extended	48/49 (98%)	29/30 (97%)
4 - <i>ampC</i> promoter mutation associated with <i>ampC</i> hyper-production	Extended	21/21 (100%)	13/13 (100%)
Increased DNA copy number			
5 Relative coverage of any transmissible beta-lactamase > 2.5*	Extended	184/227 (81%)	128/167 (77%)
Decreased permeability			
6 Features suggesting disruption of either <i>ompC</i> or <i>ompF</i> in an isolate containing an additional beta-lactamase	Extended	18/21 (86%)	2/2 (100%)

*: Cut-off chosen following a ROC analysis (section 6.3.2.)

While incorporating these additional features had little impact on the algorithm's ability to identify amoxicillin resistance (sensitivity: 523/531 [98%] extended vs 511/531 [96%] beta-lactamases only [basic]; specificity: 439/445 [99%] extended vs 441/445 [99%] basic), they proved vital for inferring amoxicillin-clavulanate resistance. By including them, algorithm sensitivity for predicting amoxicillin-clavulanate improved, with the extended algorithm's sensitivity 277/339 (82%) being significantly higher (McNemar's $p < 0.0001$) than the basic algorithm (78/339 (23%)). Increased sensitivity modestly reduced specificity (93% (591/637) extended vs 100% (634/637) basic); however, overall categorical agreement of WGS-derived with observed phenotype increased from 712/976 (73%) to 868/976 (89%) when these extended genetic features were included.

Of the remaining 108 discrepant, 46 (42%) were "major errors" (ME, predicted resistant by the extended algorithm but phenotypically susceptible), and 62 (57%) were "very major errors" (VME, predicted susceptible when resistant), giving VME and ME rates of 18% and 7% respectively. While this still falls short of standards required for regulatory approval (Table 3) it is a substantial improvement over the ~70% VME rate seen in Chapter 4. Errors were significantly more frequent among isolates with MICs within one doubling dilution of the R/S breakpoint ($p < 0.0001$, Fisher's exact test) (Figure 58). Specifically, the extended algorithm correctly identified 463/469 (99%) isolates with MIC $\leq 4/2$ mg/L as susceptible and 230/250 (92%) isolates with MIC $\geq 32/2$ mg/L as resistant (Figure 58), whereas the majority (82/108 [76%]) of the incorrect WGS resistance predictions occurred in isolates with MICs 8/2-16/2 mg/L. Tying this issue together with the poor predictive value of the copy number metric, of 79 isolates containing non-inhibitor-resistant beta-lactamases

where the algorithm incorrectly predicted susceptibility/resistance (i.e. based on DNA copy number alone), 64 (81%) had peri-breakpoint (8/2-16/2mg/L) MICs.

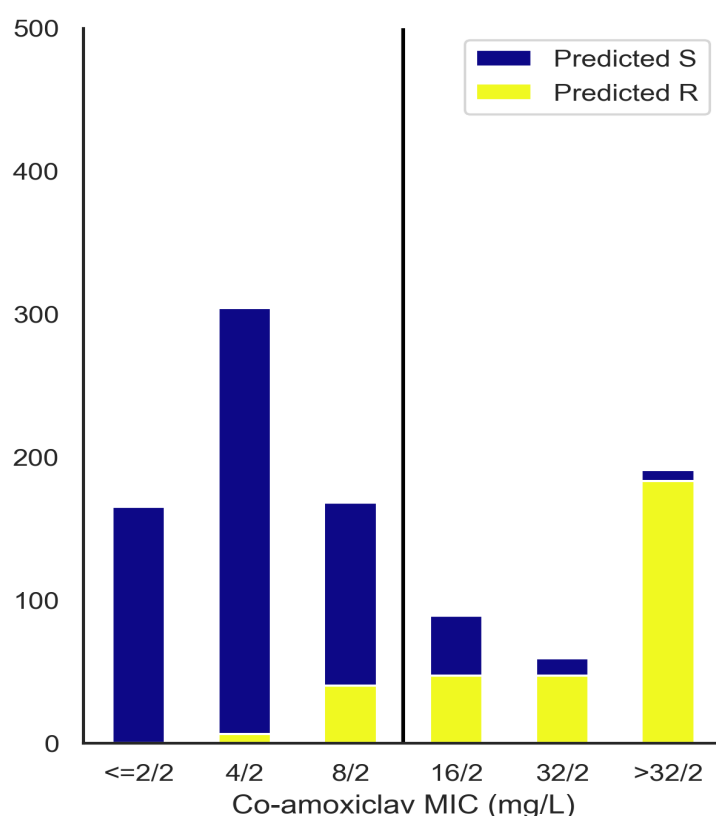


Figure 58: Extended algorithm resistance prediction by amoxicillin-clavulanate MIC

6.4. Discussion

In this study, I found that amoxicillin-clavulanate resistance prediction depended on multiple mechanisms beyond sheer presence and absence of specific beta-lactamases. However, some of these mechanisms were not obviously casual, despite a strong association with amoxicillin-clavulanate resistance. Further, despite adding these mechanisms significantly improving resistance prediction, a substantial number of discrepancies between genotypic and phenotypic prediction remained. Resistance in isolates with peri-breakpoint MICs was particularly difficult to predict.

While each of the additions to the basic prediction algorithm was generally successful, they each had their own limitations. For example, high copy number, while strongly associated with resistance, also resulted in several false positive predictions. Similarly, when found, promoter mutations were strongly predictive of resistance, however significant disruption to gene promoter regions led to false-negative predictions, as isolates lacking matching sequence were wrongly assumed susceptible. Finally, disrupted porin genes, as expected, were not capable of causing beta-lactam resistance in isolation. However their frequent co-occurrence with other resistance mechanisms capable of causing amoxicillin-clavulanate resistance in the same isolate meant the effect of this feature was far from clear.

There are several possible issues with the high copy number metric which may have contributed to discrepancies. The metric itself is defined by the relative sequencing coverage of resistance genes versus chromosomal genes, which is then used to infer the number of copies of the gene in the cell, with the assumption being that increased gene dosage leads to increased enzyme production ²²⁹. Two important assumptions are implicit in this definition. Firstly, that this metric is a good representation of the gene copy number in the isolate which underwent AST. This assumption may not hold for several reasons such as:

- different isolate sub-cultures containing different numbers of plasmids,
- variation and biases in the sequencing process leading to non-representative or poorly reproducible relative DNA coverage,
- differences in numbers of mapping reads caused by differences in bioinformatic approach (e.g. tolerance of mismatching sequence).

The second assumption is that increased gene-dosage is associated with increased gene expression for AMR genes. However, with expression also depending on other factors such as promoter strength, the relationship between gene-dosage and expression level may not be this simple. For example, if stronger promoters were commonly found with resistance genes on lower copy plasmids, lower copy number may be associated with increased expression. Reassuringly, the data suggest both assumptions are reasonable with test re-test variability being low in the limited number of isolates sequenced twice (from the same culture), suggesting reproducibility of the DNA copy number metric, and the presence of a strong dosage-response relationship of estimated copy number on MICs in isolates with a single beta-lactamase gene (Figure 55) suggesting that this proxy has some biological relevance. One further potential issue with the copy number metric is its classification into “high” and “low” categories for the purpose of resistance prediction (susceptible vs resistant). While sensible for predicting a binary phenotype, this categorization may be a poor model of the underlying biology. Correspondingly, there was no DNA copy number threshold which perfectly separated the resistant and susceptible isolates (Figure 55), and further there was no obvious way to define “high” and “low” copy numbers from their distribution which reflected a continuum rather than being bimodal (Figure 56). Like the second theoretical concern outlined in the assumptions above, this lack of a distinction may also be due to confounding of gene dosage effects by other features of beta-lactam resistance.

While in this study the promoter mutations included in the algorithm were highly predictive of resistance, these were only a small subset of all the mutations identified. Further, absence of upstream sequence similar to reference promoter sequences led to false-

negative predictions. These findings highlight several possible issues for rule-based resistance prediction. Approaches rely on searching WGS for a catalogue of known resistance sequences and then attributing phenotypic resistance directly to that sequence. Resistance that occurs by altering gene expression conforms poorly to this model, and consequently is hard to catalogue in the same way. Firstly, finding mutations outside the gene of interest which modify expression can be challenging as some genes (e.g. *ampC* in *P. aeruginosa* and *E. cloacae*) have multiple regulatory elements distantly located on the genome ^{214,230}. Secondly, as observed in isolates with non-recognisable promoter sequences in this study, reference promoter sequences may be less useful, since, in contrast to genes where large modifications typically result in gene loss of function, drastically altered promoter sequences can potentially increase expression and hence resistance. While similar issues of inferring susceptibility from unknown SNPs has been described previously ¹²¹, the potential for promoter sequences to tolerate much larger variation and still be capable of causing a significant increase in gene expression could make the eventual aim of understanding all resistance causing mutations unattainable, even in the immediately upstream region alone. Finally, effects of promoter mutations on resistance are complex to study, and resistance can only be easily attributed when they cause drastic changes to expression. In contrast to beta-lactamase proteins which can be purified and their effect on antibiotics directly observed, promoter strength is typically assessed using transfection assays, which only gives information about their level of expression. Disentangling the effects of resistance from promoter strength from other effects such as copy number, permeability and presence of other beta-lactamases can be highly challenging, particularly when effects may only cause modest changes to MIC.

The evidence of a causal association between “porin disruption” as defined in this chapter and increased amoxicillin-clavulanate resistance was limited. Consistent with previous studies, disruption of a single porin gene alone was not sufficient to cause beta-lactam resistance ⁹². However, when seen in combination with other resistance features, isolates were generally resistant (Table 24-25) to at least one beta-lactam compound. However, it was not clear if porin disruption caused any additional resistance over and above that conferred by the presence of concomitant beta-lactam resistance mechanisms. Further, although 2/2 isolates with porin disruption and low copy number beta-lactamase genes were correctly predicted amoxicillin-clavulanate resistant, 3/19 isolates (OMPF8, OMF9, and OMPF13) were phenotypically susceptible, even though they contained high copy number *bla*_{TEM} genes in addition to disrupted porins. A possible reason for these discrepant is that, again, the “porin disruption” feature used in this study is a proxy for decreased cell permeability, a quality not easily assessed by WGS. This proxy assumes that, if both (*ompF* and *ompC*) complete coding porin genes are not found in WGS data, then it is likely these genes are “functionally lost” which then in turn leads to reduced permeability to beta-lactam antibiotics (in particular for amoxicillin and clavulanate) and resistance. However, there are several ways in which these assumptions could be violated. Detecting gene loss of function from WGS data is not straightforward, as mis-assembly due to any cause can masquerade as gene disruption. Even in cases of likely genuine disruption (e.g. in the presence of high confidence frameshift mutations) of either gene, inferring reduced permeability is tenuous. *E. coli ompF* and *ompC* porin expression is jointly regulated via the *ompR* gene, and with reduced expression of one leading to increased expression of the other ⁹². Consequently, porin loss may not automatically lead to reduced cell permeability and instead only cause permeability to be “modified”, with permeability being increased

from some substances and reduced for other. Finally, different beta-lactam compounds often have multiple routes entry to the cells, and hence it may not be possible to infer resistance caused from a single “lost” porin channel at all, especially when many permeability defects in isolation seem to cause little resistance.

While the limitations of the extended prediction features may still be causing some error, results from this chapter also suggest phenotyping may play a role. Two main observations support this theory, firstly the fact that the distributions of DNA copy numbers in resistant and susceptible isolates overlapped (Figure 56) and secondly the fact that errors were more common in isolates with peri-breakpoint MICs (Figure 58). These two issues were linked, with many of the peri-breakpoint discrepancies caused by incorrect predictions for isolates with beta-lactamases present at borderline copy numbers. What made these isolates particularly problematic was the apparent similarity of the resistance genotype for isolates on either side of the breakpoint. AST claims to distinguish these isolates into well-defined resistant and susceptible groups, however, it is important to note that these groups are only a single MIC doubling dilution apart. If amoxicillin-clavulanate AST has the expected variability of MIC on repeat testing as other antibiotics (repeats within plus/minus one doubling dilution ¹⁰⁶) it is likely that AST could also be misclassifying these isolates. Further, if this were the case and the phenotype itself caused frequently classification errors in the large proportion of isolates with peri-breakpoint MICs (Figure 58), this could potentially make achieving reasonable concordance between WGS predicted and observed resistance impossible.

Amoxicillin-clavulanic acid (fixed) / *Escherichia coli*
International MIC Distribution - Reference Database 2019-04-05

MIC distributions include collated data from multiple sources, geographical areas and time periods and can never be used to infer rates of resistance

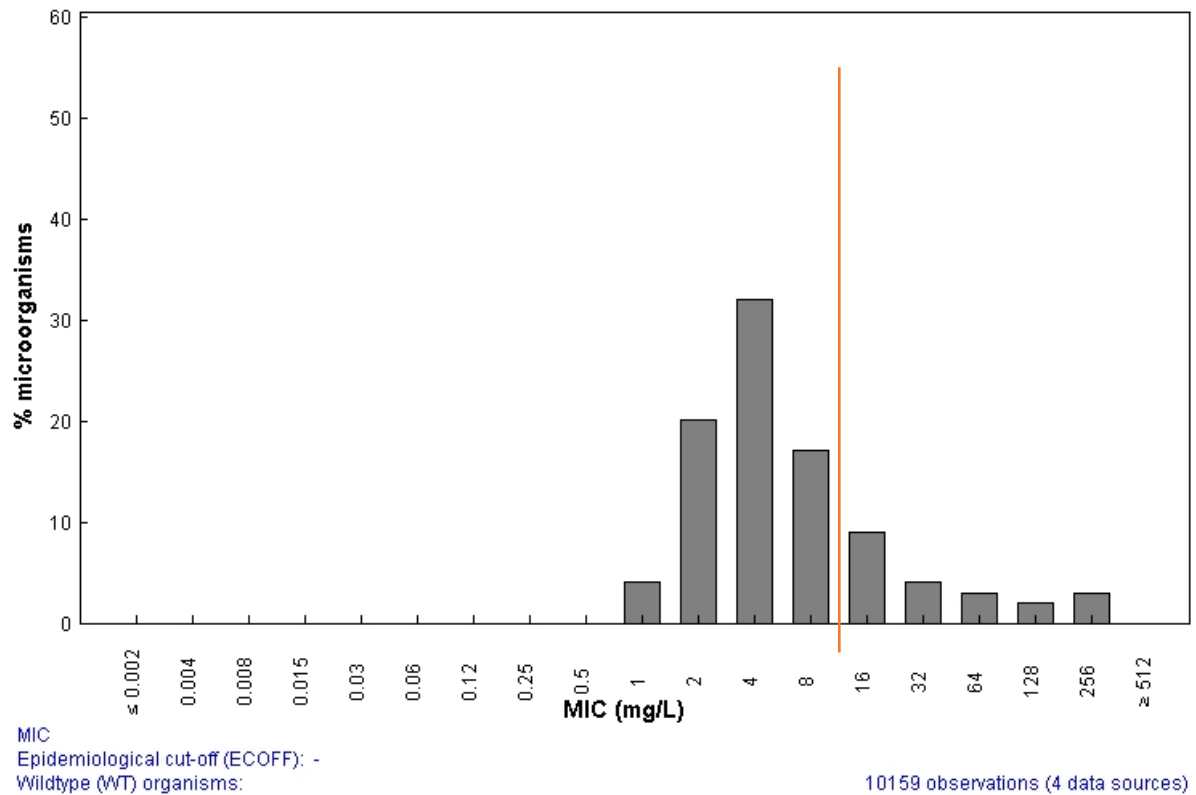


Figure 59: *E. coli* MIC distribution for amoxicillin-clavulanate

Note: reproduced from the European Committee of Antimicrobial Susceptibility Testing website, “MIC and zone diameter distributions and ECOFFs”²³¹. Current clinical breakpoint (indicated by the orange line): isolates with MIC > 8 mg/L are classified as resistant, otherwise susceptible.

Chapter 7: The problem of binary interpretation of a continuous phenotype

7.1. Introduction

While WGS-based amoxicillin-clavulanate resistance prediction was drastically improved by using a more complete reflection of the basis of beta-lactam resistance as features, VME and ME rates were still 18% and 7% respectively. Most of these remaining errors occurred in isolates with “peri-breakpoint” MICs. While they could represent a problem with the resistance prediction algorithm, expected variability of AST could also cause errors. To understand how much of the remaining error in amoxicillin-clavulanate resistance prediction was due to the phenotype, this chapter assessed amoxicillin-clavulanate testing methodology in detail:

- Firstly, assessing the accuracy of automated susceptibility phenotypes versus “gold standard” phenotypes
- Secondly, assessing the variability of the phenotype, and whether it can reliably separate isolates into resistant/susceptible categories.

This chapter then reassessed WGS-predicted/observed phenotype concordance, this time incorporating the more complete genetic description of amoxicillin-clavulanate resistance and accounting for the impact of variable AST phenotyping using multivariate mixed modelling.

7.1.1. The amoxicillin-clavulanate phenotype

Similarly to its genetic basis, amoxicillin-clavulanate phenotyping is more complex than for many beta-lactams. The fact that it consists of combination of a beta-lactam antibiotic

(amoxicillin, a broad spectrum penicillin) with a suicide beta-lactamase inhibitor (clavulanic acid, a clavam) alters both its pharmacodynamic and pharmacokinetics properties and its susceptibility testing ^{65,232}. Despite being over 40 years since its development and its status as one of the most commonly used antibiotics in Europe ²⁰¹, there are significant concerns as to the reliability of amoxicillin-clavulanate AST ^{233,234}.

Special considerations for combination antibiotics

Determining the phenotype for combination antibiotics comes with additional complexity of both testing and breakpoint determination ¹⁰¹. Firstly, as concentrations of both agents can be varied, an isolate can have multiple MICs dependent on the concentration of each agent. Secondly, at any given concentration of each component, the antibiotic combination may work additively, synergistically or antagonistically (Figure 60). Assessment of these different MICs, and how the antibiotics work together, can be done rigorously using checkerboard assays (Figure 60). However in practice these tests are more costly, difficult to perform and interpret, particularly across multiple strains, and may not reflect interactions *in vivo*. Consequently, none of the major automated AST devices support these tests. Instead, as advised by the EUCAST ¹¹⁶ and CLSI ¹¹⁵ guidelines, MICs for combination antibiotics are tested along a single axis, either by altering the concentration of one antibiotic while keeping the other fixed, or by keeping the ratio of the two antibiotics fixed (Figure 61).

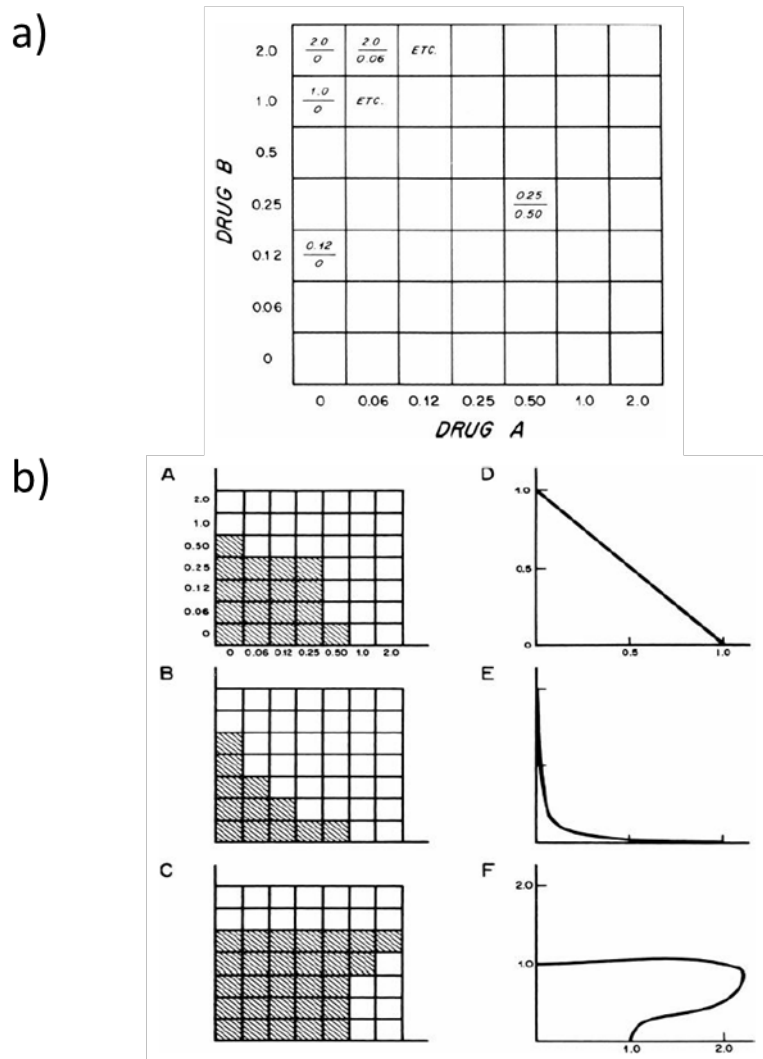


Figure 60: Checkerboard assays for combination antibiotics

Note: a) Checkerboard technique. In the checkerboard, serial dilutions of two drugs are performed using drug concentrations proportional to MICs of the drugs being tested. b) Assessment of antimicrobial combinations with the checkerboard method. A, B, and C: results of testing combinations of two drugs. Shading, visible growth. Concentrations are expressed as multiples of the MIC. D, E, and F: isobolograms (plotted on an arithmetic scale) that represent the results of checkerboards shown in A, B, and C, respectively. A and D, additive effect; B and E, synergism; C and F, antagonism. Reproduced from Antibiotics in Laboratory Medicine, Pillai S et al ¹⁰¹.

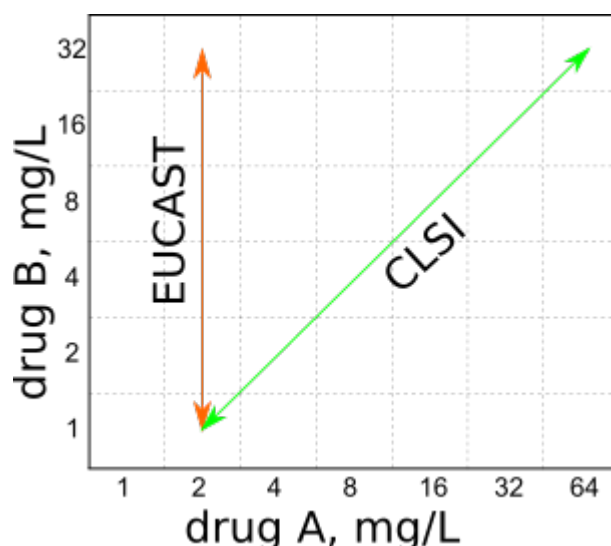


Figure 61: Typical MICs tested for combination antibiotics

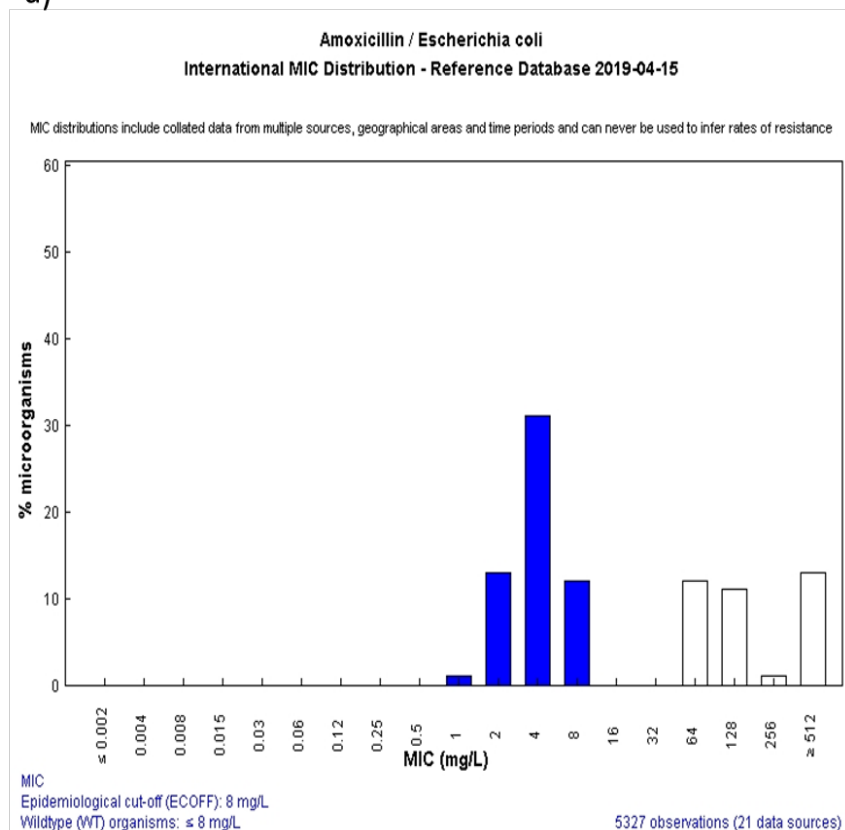
Note: orange: keeping the concentration one component of the pair fixed, this is the EUCAST guidance for testing amoxicillin-clavulanate; green: keeping the concentration ratio fixed, this is the CLSI guidance for testing amoxicillin-clavulanate.

With the increasing use of beta-lactam/beta-lactamase inhibitor combinations, due to their ability to treat highly drug-resistant organisms ^{65,86}, efforts have been made to improve clarity on how to set breakpoints for agent/inhibitor combination antibiotics ²³⁵. However, these assume that the inhibitor only acts to prevent degradation of the active compound, and that it does so completely, fully restoring the activity of the active compound. With this assumption, the active compound then operates as if no beta-lactamases were present in the isolate, and hence its original breakpoints are still valid. However, this is likely an oversimplification as the quantity of beta-lactamase inhibitor (Figure 50), and its affinity to inhibit its target (i.e. a beta-lactamase) both play a role, particularly since these affinities are specific for each beta-lactamase. Additionally, the distinction between inhibitor and active compound for beta-lactamase inhibitors and beta-lactam antibiotics is somewhat artificial, as many inhibitors (e.g. clavulanate) display low level antibiotic activity and many antibiotics also act as beta-lactamase inhibitors (e.g. carbapenems) ⁶⁵.

Known issues for amoxicillin-clavulanate AST

In addition to the complexity arising from being a combination antibiotic, there are several longstanding issues which make AST of amoxicillin-clavulanate particularly difficult. Unlike some other antibiotics, the distribution of amoxicillin-clavulanate offers no easy separation isolates into susceptible/resistant groups (Figure 62). It is perhaps for this reason, plus the fact that defining which genetic mechanisms constitute an organism being part of the non-wild-type population, that amoxicillin-clavulanate has no ECOFF defined.

a)



b)

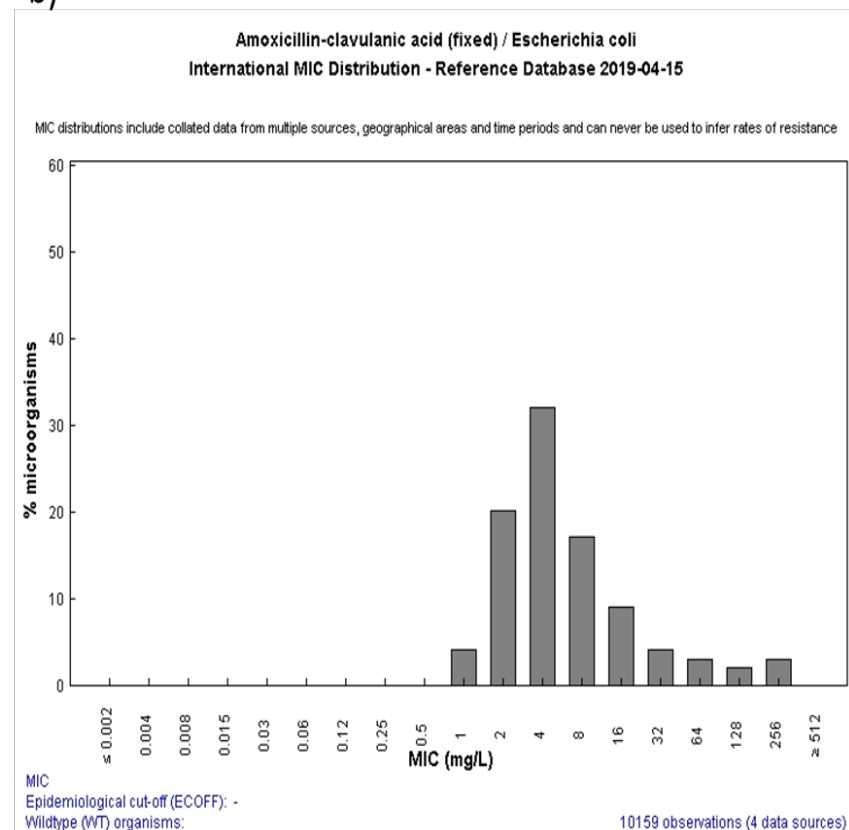


Figure 62: MIC distribution a) amoxicillin and b) amoxicillin-clavulanate.

Note: Clear natural separation of amoxicillin susceptible/resistant distributions (a), compared no obvious separation for amoxicillin-clavulanate (b). For amoxicillin, this naturally seen separation defines both the clinical and ECOFF breakpoints. Blue: below the ECOFF breakpoint, white: above the ECOFF breakpoint, grey: no clear ECOFF breakpoint. Reproduced from www.eucast.org ²³¹.

Clinical breakpoints are no less complex, with EUCAST and CLSI guidelines differing on both how to perform AST, and what the clinical breakpoint should be ^{115,116}. For AST testing, EUCAST uses a fixed clavulanate concentration of 2mg/L and labels isolates with amoxicillin-clavulanate MICs > 8/2mg/L as resistant, whereas CLSI fixes the ratio of amoxicillin: clavulanate concentration at 2:1, and its breakpoints include an intermediate zone (16/8mg/L) between susceptible and resistant isolates. Further, EUCAST breakpoints for amoxicillin-clavulanate are not consistent across methods and indications, with different breakpoints recommended for treating uncomplicated urinary tract infections and disk diffusion testing of susceptibility having “area of technical uncertainty” range which is not reflected in the broth dilution test guidelines ²³⁶. These differences in method/breakpoints are not purely semantic, with different guidelines having been shown to give opposite interpretations for the same isolates (Figure 63) ^{234,237}. Currently there is no guidance as to which breakpoint is optimal clinically, with no set of breakpoints having any strong evidence of association with clinical outcome.

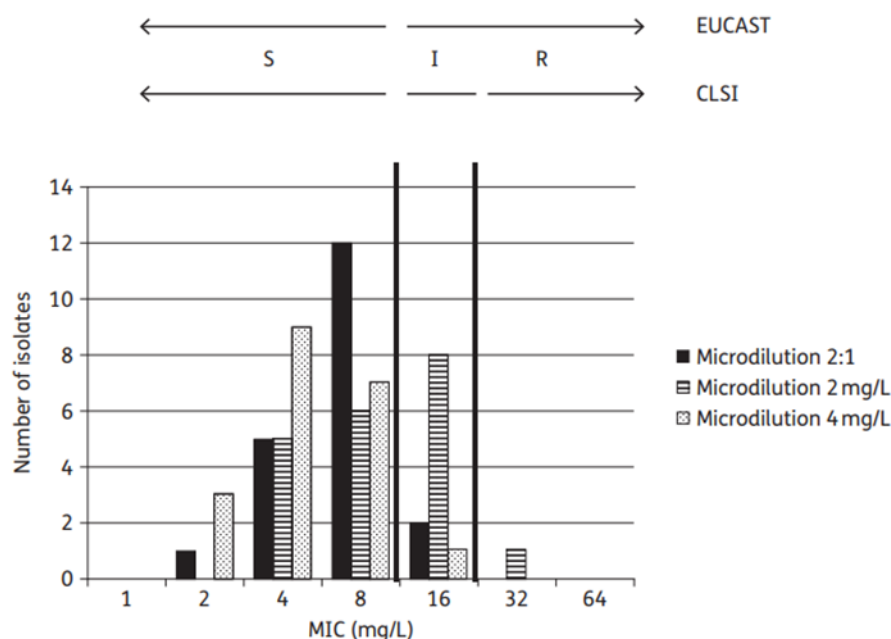


Figure 63: Amoxicillin-clavulanate MIC distribution for ESBL-producing *E. coli* tested by different methods

Note: Isolates tested by microdilution susceptibility testing using a 2:1 co-amoxiclav ratio, fixed 2 mg/L clavulanate and fixed 4 mg/L clavulanate. The corresponding interpretive categories according to EUCAST and CLSI breakpoints are shown in the upper part of the figure. Reproduced from Diez-Aguilar et al ²³⁴.

7.2. Methods

7.2.1. Sample selection

To select isolates for detailed phenotyping, a subsample was chosen from the OUH *E. coli* bloodstream infection (BSI) dataset (details in Chapter 3) by stratified random sampling (Figure 64). Samples were classified into strata based on whether they contained a beta-lactamase gene (identified using the initial prediction approach which used *de novo* assembly ¹⁶⁹ and BLASTn ¹³⁵, Chapter 4), whether they had peri-breakpoint MICs on automated AST performed in the routine service laboratory, and whether they were WGS predicted/observed phenotype discordant on initial screen (Chapter 4, Figure 65). This partitioned the available isolates into eight categories, from which isolates were selected.

Additionally, any piperacillin-tazobactam resistant isolates that were not included by any other means were also selected for detailed agar-dilution phenotyping. The primary aim of using this strategy was to over-represent “interesting” isolates in this set, which would challenge a WGS-based resistance prediction algorithm, namely:

- Isolates with peri-breakpoint MICs
- Phenotypically resistant isolates with no beta-lactam resistance mechanisms identified
- Phenotypically resistant isolates which only contained non-inhibitor resistant beta-lactamases
- Isolates with inhibitor-resistant beta-lactamases
- Isolates resistant to piperacillin-tazobactam

As a result, the subsample was enriched for amoxicillin-clavulanate resistance, and all strata excepting non-beta-lactamase containing WGS/phenotype concordantly susceptible samples, were over-represented (Figure 65).

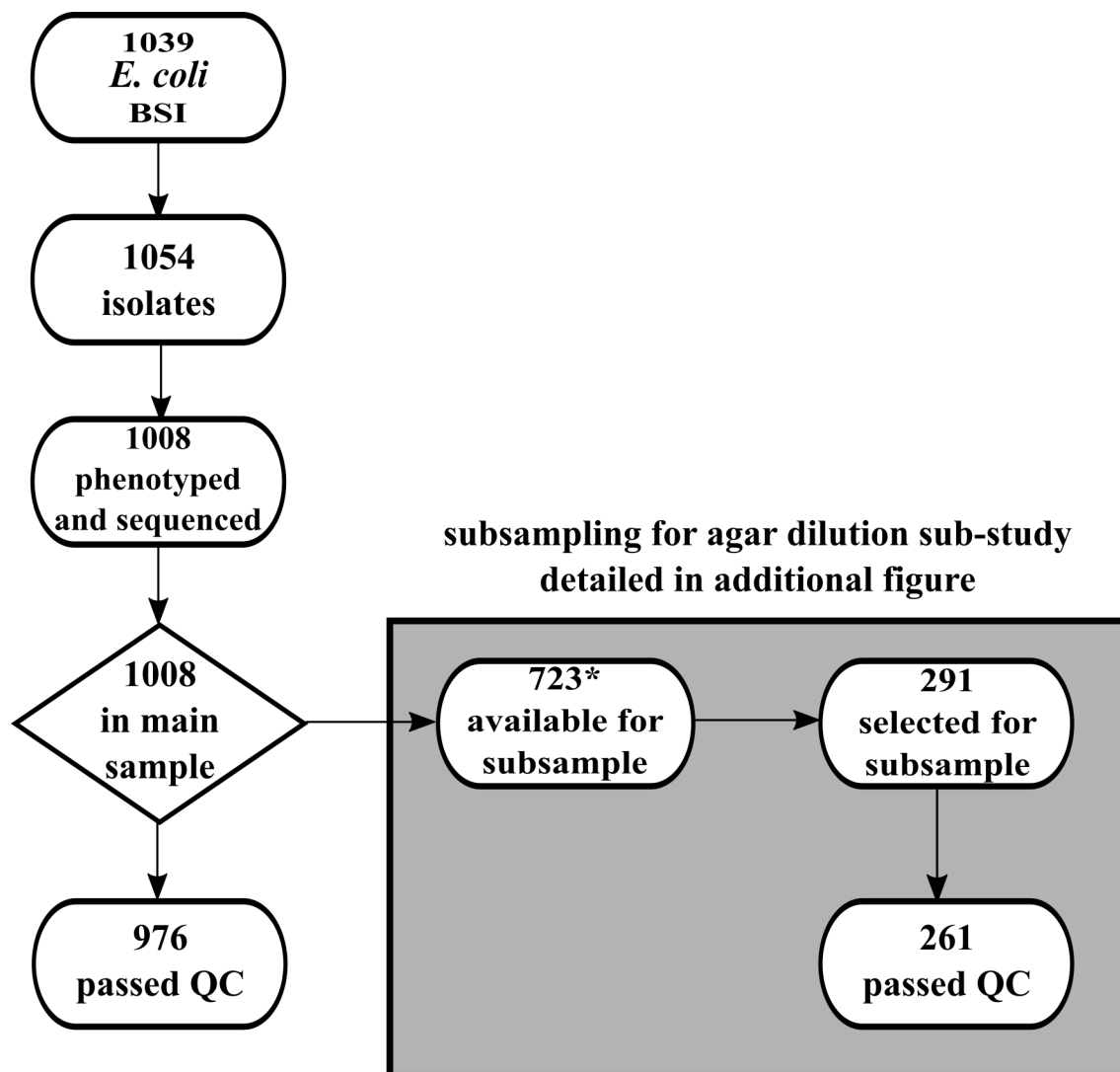


Figure 64: Sampling frame for isolates selected for detailed agar dilution phenotyping.

*: Of all 1039 *E. coli* BSI occurring between January 2013-August 2015, only 723 samples (1 January 2013 – 31 December 2014) had complete sequencing data and BD Phoenix phenotype at the time of sample selection.

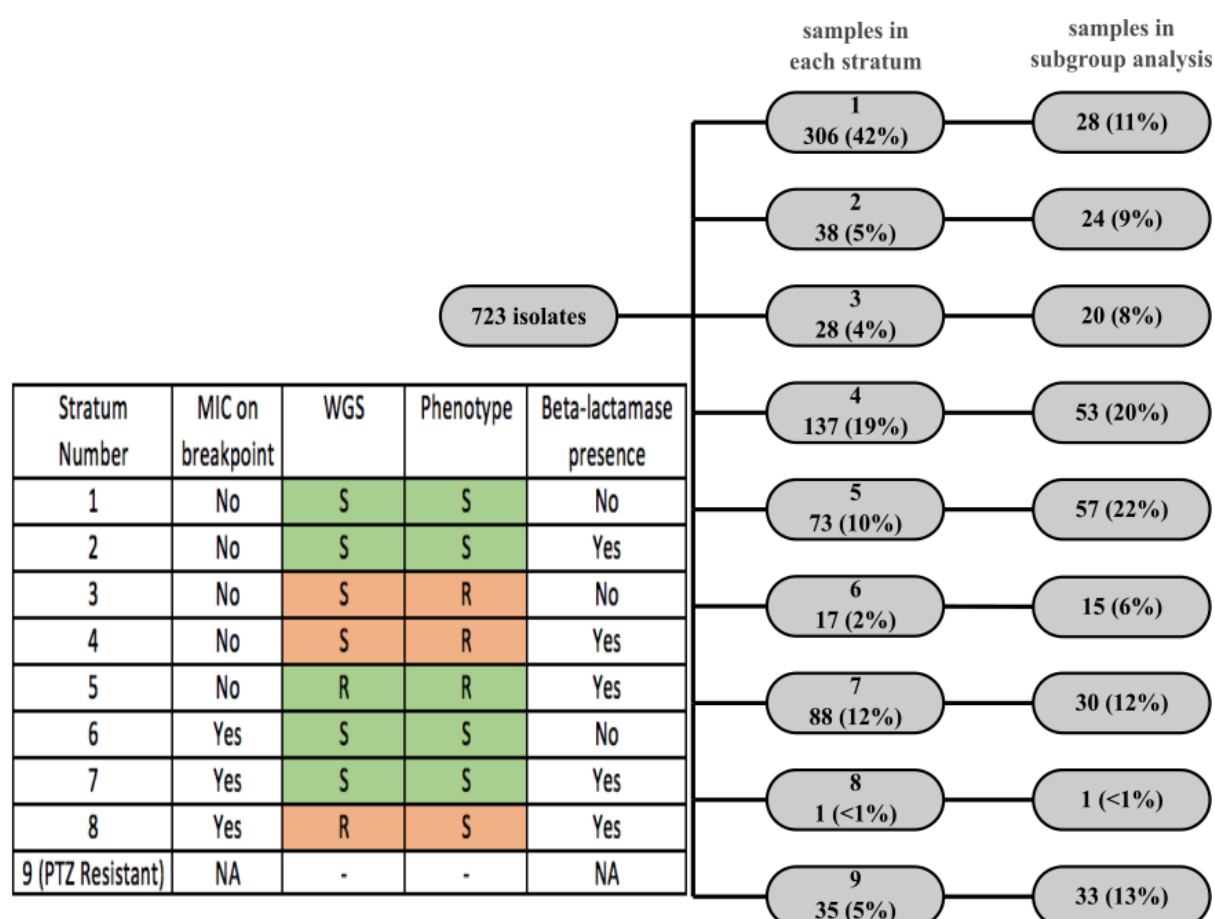


Figure 65: Stratified random sampling strategy for detailed agar dilution phenotyping

Note: samples included in subgroup analysis if in addition to being selected, they passed additional quality control checks (see section 7.2.2). Orange indicates WGS-predicted/AST observed phenotype discordance.

7.2.2. Phenotyping

Selected samples then underwent replicate agar dilution amoxicillin-clavulanate

phenotyping with clavulanate concentration and MIC interpretation according to each of

EUCAST (EUCAST-based agar dilution), and CLSI (CLSI-based agar dilution) guidelines ²³⁸.

Agar dilution testing was based on BSAC guidelines ¹⁷⁶ and followed protocols used by the

UK Animal and Plant Health Agency (APHA) ²³⁹ where the testing took place. Of note, these

methods are equivalent to the “gold-standard” endorsed by each of these regulatory

bodies. For each method, sub-cultures (from frozen stocks) were tested a minimum of three

times using ISO-Sensitest agar plates containing amoxicillin and clavulanate in a 2:1 ratio

(CLSI) or a fixed concentration of clavulanate (2 mg/L, EUCAST), with *E. coli* controls ATCC25922 (wild type) and ATCC35218 (TEM-1 beta-lactamase producer). For any individual test to be deemed satisfactory, the plate controls needed to satisfy the following:

- plate control ATCC25922 had amoxicillin MIC in range 2-8 mg/l ^{115,240}
- plate control ATCC25922 had CLSI-based agar dilution (2:1 ratio amoxicillin/clavulanate) MIC 2/1-8/4 mg/l ¹¹⁵
- plate control ATCC25922 had EUCAST-based agar dilution (fixed 2mg/l clavulanate) MIC 2/2-8/2 mg/l ²⁴⁰
- plate control ATCC35218 had CLSI-based agar dilution (2:1 ratio amoxicillin/clavulanate) MIC 4/2-16/8 mg/l ¹¹⁵
- plate control ATCC35218 had EUCAST-based agar dilution (fixed 2mg/l clavulanate) MIC 4/2-32/2 mg/l ²⁴⁰

For additional quality control, bacterial isolates were plated on sheep blood agar and incubated overnight at 37°C to check purity, with isolates excluded if multiple colony morphologies were seen. Any isolate failing any of these quality control checks for any replicate was phenotyped an additional time. Isolates were included in the final analyses if they had phenotypic results with two or more MICs in essential agreement (+/- one doubling dilution) for all of amoxicillin, EUCAST-based and CLSI-based agar dilution. The susceptibility of an isolate was defined using the “upper median” MIC (choosing the higher MIC when the median lay between two MIC readings) for phenotype category interpretation.

7.2.3. Comparison of phenotypes

For the agar dilution phenotypes, first the variability of each individual phenotype was evaluated (both in terms of MICs and resistance interpretations). Then comparisons were made between CLSI and EUCAST-based agar dilution gold standard phenotypes, and between each agar-dilution phenotype and the routine AST results generated by the BD Phoenix. The variability of each agar dilution phenotype was examined using simulation. For each individual phenotype, two simulated sets of MIC results were generated by selecting a single result from repeat tests conducted on the isolate. Sets were then assigned first/second in random fashion. Then taking the first as the reference, essential agreement (EA), categorical agreement (CA), VME and ME rates were calculated. This procedure was then repeated 10000 times to give point estimates and 95% confidence intervals for each rate.

7.2.4. Comparison with genotype

Each set of agar dilution phenotype results was compared against the basic (beta-lactamase only, Chapter 4) and extended (beta-lactamases + “extended” features, Chapter 6) resistance predictions (details in sections 6.2.2 and 6.2.3) in terms of sensitivity, specificity, CA, VME and ME.

Given the suggestion of issues with the binary interpretation of phenotype/genotype (section 6.4) which was also supported by agar dilution phenotypes (section 7.3.2), a third method used a statistical (linear mixed) model to predict actual observed log₂ MIC as a function of genetic resistance features (fitted using Stata 14.2; StataCorp LP, 2015). This model was first built using agar dilution MICs (the “training set”), and then tested on the remaining non-agar dilution (independent) OUH isolates (the “testing” set) to see if it could

predict the routine BD Phoenix MIC. At Oxford University Hospitals NHS Trust the BD Phoenix is calibrated against EUCAST guidelines, and hence these MICs “should” match EUCAST-based agar dilution. This avoided the need to define a categorical boundary separating resistant/susceptible isolates. The linear mixed model could also account for phenotypic variability by including all the individual measurements and directly estimating the variance associated with within isolate repeats, and could directly assess the impact of test method and individual WGS-identified genetic elements. Additional details regarding model construction are described within the results section, to explain decisions I made in the model building process.

Models were constructed to relate each MIC (modelled on the log₂ scale, so a 1-unit increase represents a doubling) to various genetic elements. Each test result was included as an observation, accounting for repeated tests of the same isolate using isolate-level random effects. Sets of agar plates containing varying concentrations of amoxicillin-clavulanate (fixed clavulanate 2mg/L (EUCAST) and 2:1 ratio clavulanate (CLSI)) were prepared in batches, which were then used to test 55 isolates for each batch. One source of variation in MICs is therefore the test batch. Another source of variation is the isolate itself. As not all batches contained the same isolates due to repeat testing, isolates were not nested within the same batch (so-called crossed random effects).

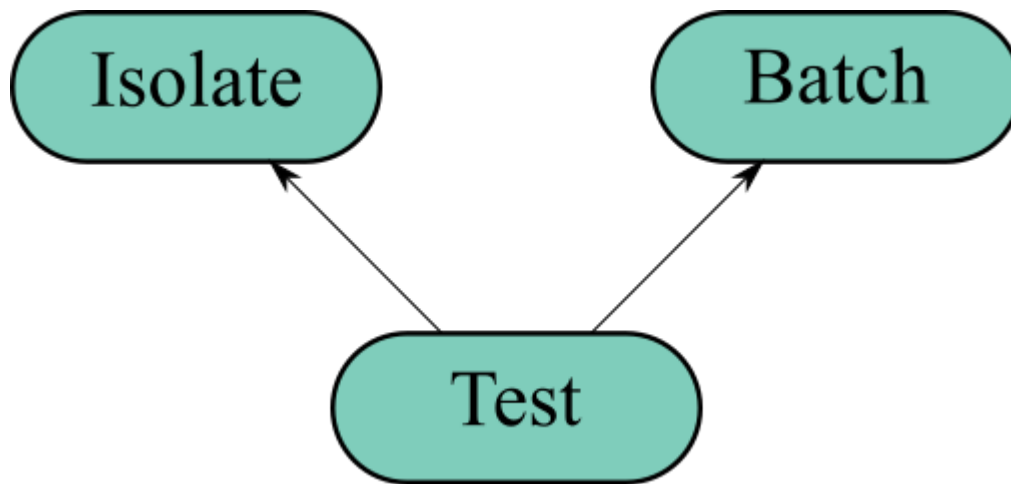


Figure 66: Inherent clustering of phenotypic tests

The model was fitted in the following stages

1. Selecting model type (linear vs ordinal logistic regression) and random effects structure (fixed vs various random effects structures) (detailed below)
2. Categorizing resistance features
3. Adding main fixed effects to the model
 - a. Selecting features for inclusion
 - b. Selecting the functional form for the relationship between each feature and outcome
 - c. Checking for strain-level effects
4. Checking interactions between fixed effects and checking for the presence of heteroscedasticity in the random effects.
5. Checking importance of components of full model for model fit

Model construction; selecting model and random effects structure

All models were fitted using maximum likelihood estimation. Given the outcome (MIC) consisted of results belonging to one of 10 ordinal categories, and that using linear models has been successful on similar datasets previously ¹²³, it was modelled using mixed linear regression (as opposed to mixed ordinal logistic regression) for simplicity and interpretability. Random effects structure was then chosen from random effects for isolate only, batch only, separate and crossed random effects for isolate and batch, selecting the model with the lowest Akaike Information criterion (AIC). Model fit was assessed by inspecting random effect residuals using QQ(quantile-quantile)-plots (Figure 67) and caterpillar plots (Figure 68) of level-specific residuals from the random-effects only models, with figures showing plots for the best-fitting (lowest AIC) random effects structure, crossed random effects for isolate and batch.

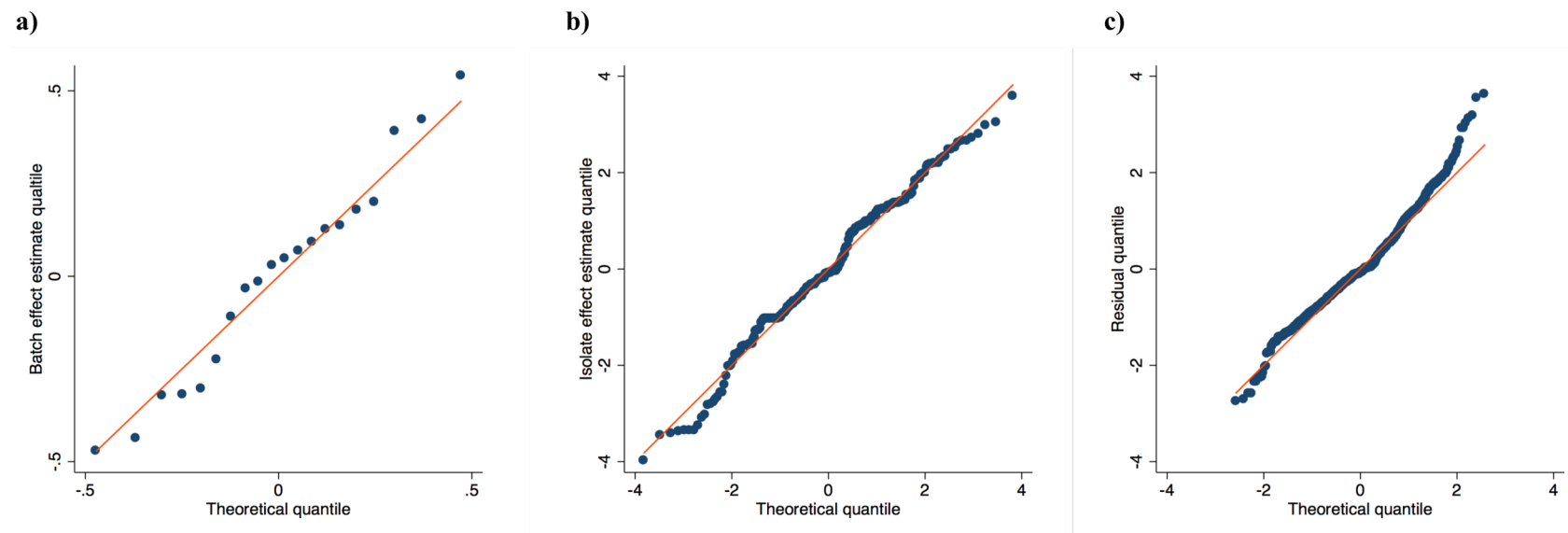


Figure 67: QQ plots checking random effect and residual normality.

Note: This shows QQ plots of batch, isolate and “random error” residuals obtained when using the crossed random effects (batch X isolate) only model a) batch level random effects, b) isolate level random effects and c) residual errors

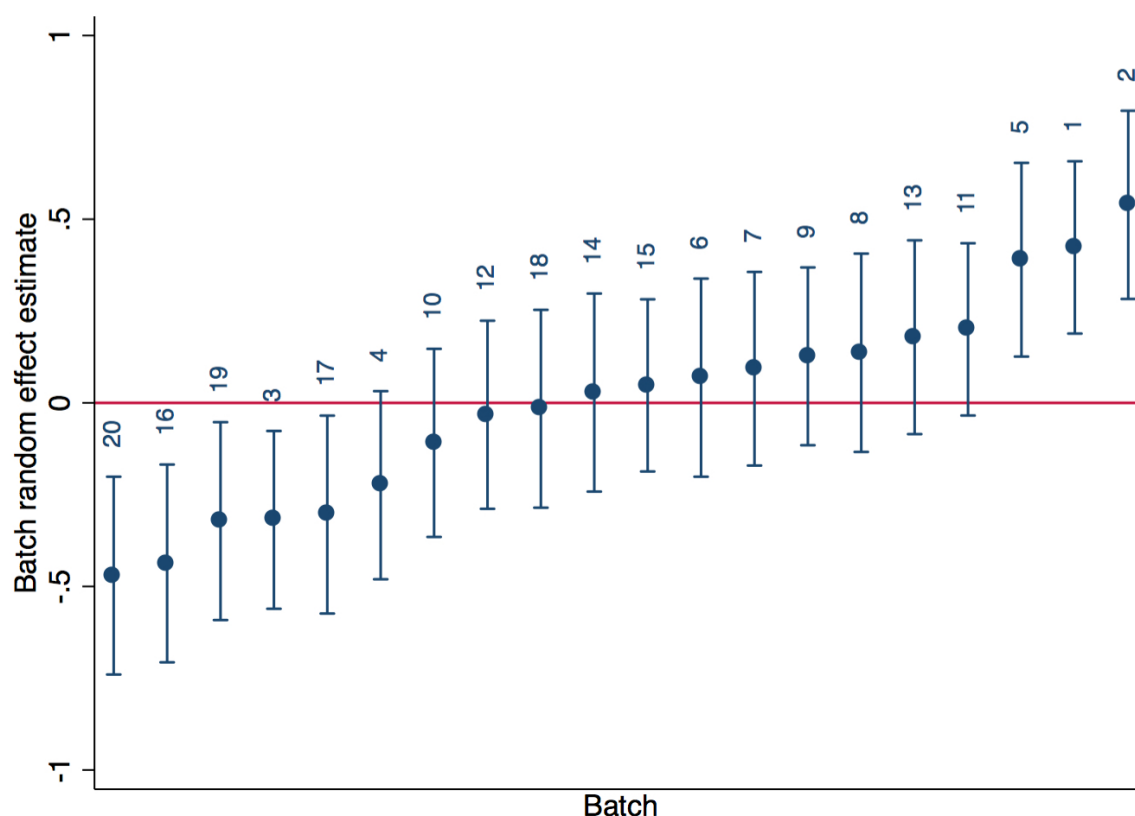


Figure 68: Size of batch errors with 95% confidence interval

Note: Batch level effects as estimated when using the crossed (batch X isolate) random effects only model. Batch numbering reflecting the ordering in which batches were done. Of note there was a small but significant trend ($p=0.003$) of later batches having lower MIC results. The cause for this was unclear but potentially reflects either reagent differences or the fact that manual AST is operator dependent, and highlights the need to control for batch effects when manually measuring MIC.

To estimate the effect of genetic features on MIC, some of which occurred in small numbers of isolates, genetic features were first categorised. Beta-lactamases were categorised by family and Bush-Jacoby classification. Allelic variants of *bla* genes in categories with fewer than 10 isolates and non-complete (but not definitively truncated) *bla* genes present in the agar dilution sub-sample were reclassified as “Other_bla” (Table 27).

Table 27: Categorization of beta-lactamases

Model Category	Number of isolates with this category (% of total 261)	Included components
blaTEM:2b	155 (59%)	<i>bla</i> _{TEM-1} (N=154), <i>bla</i> _{TEM-135} (N=1)
blaCTXM:2be	37 (14%)	<i>bla</i> _{CTX-M-15} (N=30), <i>bla</i> _{CTX-M-14b} (N=2), <i>bla</i> _{CTX-M-1} (N=2), <i>bla</i> _{CTX-M-27} (N=2), <i>bla</i> _{CTX-M-3} (N=1)
blaOXA:2d	35 (13%)	<i>bla</i> _{OXA-1} (N=35)
blaSHV:2b	10 (4%)	<i>bla</i> _{SHV-11} (N=10)
Other_bla	13* (5%)	<i>bla</i> _{TEM-40} (N=2), <i>bla</i> _{TEM-30} (N=3), <i>bla</i> _{CMY-2} (N=3), <i>bla</i> _{OXA-48} (N=1), <i>bla</i> _{TEM-190} (N=1), <i>bla</i> _{TEM-33} (N=1) 3 non-complete beta-lactamases

*: one isolate contained both *bla*_{CMY-2} and a non-complete beta-lactamase

Note: final training set included 261 isolates, see Results for details.

For both promoters in the genotype (the *ampC* and *bla*_{TEM} promoters), there were no clear association of resistance with commonly seen variants (section 6.3.2, Table 22, Table 23) and large number of individually relatively rare promoters with “significant” mutations (section 6.2.2). Therefore, for each promoter, isolates were categorised into two groups, those containing any of the pre-determined mutations known to cause hyperexpression (9 *ampC* promoters and 30 *bla*_{TEM} promoters) of their linked beta-lactamase^{71,224}, and those which did not. This avoided including categories with few isolates which would have highly variable estimates and reduce model reliability. For porins, like promoters, isolates were grouped into one of two categories depending on whether the isolate contained a beta-lactamase and was suspected of having loss of function of either *ompC* or *ompF* genes (9 isolates). Again, this was to avoid categories with too few isolates given the small numbers of disrupted porins and evidence of loss of function being taken from multiple sources (truncation mutations, quality of gene assembly, gene coverage).

Given the need to control for the effects of the different clavulanate ratios intrinsic to the two test methods (EUCAST-based vs CLSI-based agar dilution), models *a-priori* included a main fixed effect for test method and each genetic element, initially including all genetic features as binary presence/absence (or significant/non-significant for promoters and porin non-functionality). Evidence for additional effects of gene dosage of transmissible elements was then considered. Considering each beta-lactamase in turn, with the most common first (i.e. bla_{TEM} , then bla_{CTX-M} , bla_{OXA} , bla_{SHV}), each was modelled as:

- Binary (presence/absence, classical interpretation) (default)
- Binary (presence/absence) + linear effect of increasing copy number (see section 6.2.2 for how this is estimated, Figure 54) when present (each absolute unit higher copy number has the same effect on log2 MIC)
- Binary (presence absence) + log2-linear effect of increasing copy number when present (each doubling of copy number has the same effect on log2 MIC)

For each beta-lactamase, the best configuration was selected using the Akaike information criterion (AIC) (Table 28). To reduce the influence of outliers on model selection, one extreme copy number (54.8) (for bla_{SHV-1}) was truncated to the 95th centile (16.27). For all other bla_{SHV} containing isolates, median bla_{SHV} copy number was 1.77, IQR (1.70,9.05). There were no other clear copy number outliers.

Finally, as a limited assessment of strain effects, effects of the three main multi-locus sequence types (MLSTs) (present in > 10 samples) were added to the model as fixed effects.

However, none of these improved model fit ($p > 0.14$) so they were dropped and not considered further.

Table 28: Modelling methods for each genetic feature

Model category	Model component for isolate i	Modelled as (based on AIC selection)	Equation of model component
<i>bla</i> TEM:2b	<i>bla</i> TEM:2b _{i}	Binary (presence/absence) + $\log_2(\text{copy number for copy number} > 1)$	$\text{bin}(G_i) + \log_2(\text{cpn}(G_i))$ for G_i in <i>bla</i> TEM:2b
<i>bla</i> CTXM:2be	<i>bla</i> CTXM:2be _{i}	Binary (presence/absence)	$\text{bin}(G_i)$ for G_i in <i>bla</i> CTXM:2be
<i>bla</i> OXA:2d	<i>bla</i> OXA:2d _{i}	Binary (presence/absence)	$\text{bin}(G_i)$ for G_i in <i>bla</i> OXA:2d
<i>bla</i> SHV:2b	<i>bla</i> SHV:2b _{i}	Binary (presence/absence) + $\log_2(\text{copy number for copy number} > 1)$	$\text{bin}(G_i) + \log_2(\text{cpn}(G_i))$ for G_i in <i>bla</i> SHV:2b
Other_ <i>bla</i>	Other_ <i>bla</i> _{i}	Binary (presence/absence)	$\text{bin}(G_i)$ for G_i in Other_ <i>bla</i>
<i>bla</i> TEMpr	<i>bla</i> TEMpr _{i}	Binary (significant Y/N)	$\text{sig}(G_i)$ for G_i in <i>bla</i> TEMpr if isolate contains a <i>bla</i> _{TEM} gene
<i>amp</i> Cpr	<i>amp</i> Cpr _{i}	Binary (significant Y/N)	$\text{bin}(G_i)$ for G_i in <i>amp</i> Cpr
NFOMP	NFOMP _{i}	Binary (significant Y/N)	$\text{bin}(G_i)$ for G_i in NFOMP

Note: AIC = Akaike Information Criteria. For a given isolate i , and a genetic mechanism G_i , presences/absence, gene dosage and mutation effects are respectively defined as

$$\text{bin}G_i = \begin{cases} 1, & \text{if gene present in } i \\ 0, & \text{otherwise} \end{cases}$$

$$\text{cpn}G_i = x \text{ where } x \text{ is the relative coverage of the gene}$$

$$\text{sig}G_i = \begin{cases} 1, & \text{if isolate } i \text{ has a significant mutation} \\ 0, & \text{otherwise} \end{cases}$$

As defined, the effect of mutations in *bla*_{TEM} promoters already represent an interaction, since they can only have an impact in the presence of a beta-lactamase. Likewise, consistent with the resistance prediction algorithm in Chapter 6, functional porin loss was also treated as an interaction as when initially investigated it appeared to have little effect on

amoxicillin-clavulanate phenotype in isolates with no beta-lactamases. Consequently, these were omitted from further investigation of interactions. Given the importance of test method, I included all first order interactions between test method and all other main genetic factors in the model. All individually were $p < 0.05$ on a likelihood ratio test comparing multivariable models with and without each specific interaction, supporting this choice, except for the interaction between test method and functional porin loss or *bla*_{CTX-M} genes which may be due to the smaller number of isolates with these mechanisms. Other first order interactions between genetic elements included as main effects were assessed if the two genetic elements were found together in 5 or more isolates, and were selected by forward selection using likelihood ratio tests with entry $p < 0.05$. In addition to checking fixed-effect interactions, whether random effects and residual errors varied according to test methodology was investigated. Including heteroskedastic residual errors (i.e. allowing the residual error of each test MIC to vary across test methodology), and allowing the random error associated with each isolate's MIC and the random error associated with batch effects to vary according to test method improved model AIC and so were included. Covariance between test methods was only estimable on isolate level random errors due to small numbers of clusters at batch level.

To verify the importance of all fixed effect terms not included *a priori*, starting with the last term added through the model construction process above, each term was assessed by comparing AICs of models with and without the term. Terms which did not improve the AIC were removed and this process repeated until all remaining terms were either *a priori* inclusions or improved model fit (as measured by AIC), resulting in the final model:

$$\begin{aligned}
& method_{ijk} + Mec_i + method_{ijk} * Mec_i \\
& blaCTXM:2be_i * blaOXA:2d_i + blaCTXM:2be_i * blaTEM:2b_i \\
MIC_{ijk} = & isolate_i + isolate_i * method_{ijk} \\
& batch_j + batch_j * method_{ijk} \\
& error_{ijk} + error_{ijk} * method_{ijk}
\end{aligned}
\begin{array}{l}
\left. \begin{array}{l} \\ \\ \end{array} \right\} \text{Fixed effects: main effects} \\
\left. \begin{array}{l} \\ \end{array} \right\} \text{Fixed effects: saturation effects} \\
\left. \begin{array}{l} \\ \\ \end{array} \right\} \text{Random effects}
\end{array}$$

Where for an isolate i, batch j, and test k, and test methods 0 (CLSI), and 1 (EUCAST)

$$Mec_i = \frac{Other_bla_i + blaOXA:2d_i + blaCTXM:2be_i + blaTEM:2b_i + blaSHV:2b_i}{+ ampCpr_i + blaTEMpr_i + NFOMP_i}$$

$$isolate_i \sim N(0, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix})$$

$$batch_j \sim N(0, \begin{bmatrix} \omega_0^2 & 0 \\ 0 & \omega_1^2 \end{bmatrix})$$

$$error_k \sim N(0, \begin{bmatrix} \epsilon_0^2 & 0 \\ 0 & \epsilon_1^2 \end{bmatrix})$$

(Note: names in equations above represent model components.)

Model analysis and prediction

This model was used with three aims:

- To identify independent, method-specific effects of each of the identified genetic features on MIC
- To identify and quantify variability of MICs seen on repeated testing of the same isolate (within-isolate variability)
- To create method-specific WGS-based MIC predictions to compares with routine AST overall and in an independent test set.

The first two were obtained directly from model output. Model fixed components were then used to make method-specific predictions for all OUH *E. coli* BSI isolates (including those not selected for agar dilution phenotyping). As model construction only used samples with the agar dilution phenotypes (the “training set”), those without (with only BD Phoenix

MICs) are an independent “testing set” for the model. Using this “testing set”, predicted MICs (assuming EUCAST based methodology) were then compared to those found by the BD Phoenix (calibrated against EUCAST guidelines). 11 Isolates with mechanisms not seen in the training set were excluded from this analysis.

7.3. Results

7.3.1. Samples

261/291 (90%) isolates initially selected for agar dilution passed phenotyping quality control (section 7.2.2). The stratified random sampling enriching for resistant phenotypes meant that 160/261 (61%) isolates were amoxicillin-clavulanate resistant by routine AST (Table 29). All STs with >10 isolates in the full Oxfordshire BSI dataset were represented in the agar dilution set, with 51 (20%), 43 (16%) and 29 (11%) isolates being ST131, ST73 and ST69, respectively. Beta-lactamases, again through the sampling procedure, were more common in the agar dilution set than the full dataset, with over half the isolates containing *bla*_{TEM-1} (Table 29).

Table 29: Comparison of agar dilution and full OUH datasets

Feature		No of isolates with the feature (% of respective dataset)	
AST-observed resistance		Agar dilution dataset (N=261)	Full dataset (N=976)
Amoxicillin-clavulanate		160 (61)	339 (35)
Ampicillin		219 (84)	531 (54)
Aztreonam		31 (12)	61 (6)
Cefepime		32 (12)	57 (7)
Cefoxitin		19 (7)	30 (3)
Ceftazidime		32 (12)	53 (5)
Ceftriaxone		38 (15)	79 (8)
Ertapenem		1 (0.3)	3 (0.3)
Meropenem		0 (0)	0 (0)
Piperacillin-tazobactam		34 (13)	50 (5)
Sequence Type			
131		51 (20)	123 (13)
73		43 (16)	161 (17)
69		29 (11)	94 (10)
95		18 (7)	91 (9)
12		16 (6)	44 (5)
10		5 (2)	26 (3)
127		3 (1)	33 (3)
Other		96 (37)	404 (41)
Resistance Features			
Broad-spectrum penicillinase	<i>bla</i> _{TEM}	156 (60)	412 (42)
	<i>bla</i> _{SHV}	10 (4)	20 (2)
ESBL	<i>bla</i> _{CTX-M}	37 (14)	73 (7)
	<i>bla</i> _{TEM}	2 (1)	3 (0.3)
	<i>bla</i> _{SHV}	0 (0)	3 (0.3)
Inhibitor-resistant variants	<i>bla</i> _{TEM}	7 (3)	13 (1)
Oxacillinases	<i>bla</i> _{OXA}	36 (10)	62 (6)
Increased copy-number beta-lactamases		96 (37)	227 (23)
<i>ampC</i> promoter mutations		9 (3)	20 (2)
<i>bla</i> _{TEM} promoter mutations		30 (11)	49 (5)
Porin disruption		10 (4)	22 (2)

7.3.2. Evaluating phenotyping methods

As expected, choice of phenotyping methodology had significant impact on the resistance category of isolates (Figure 69).

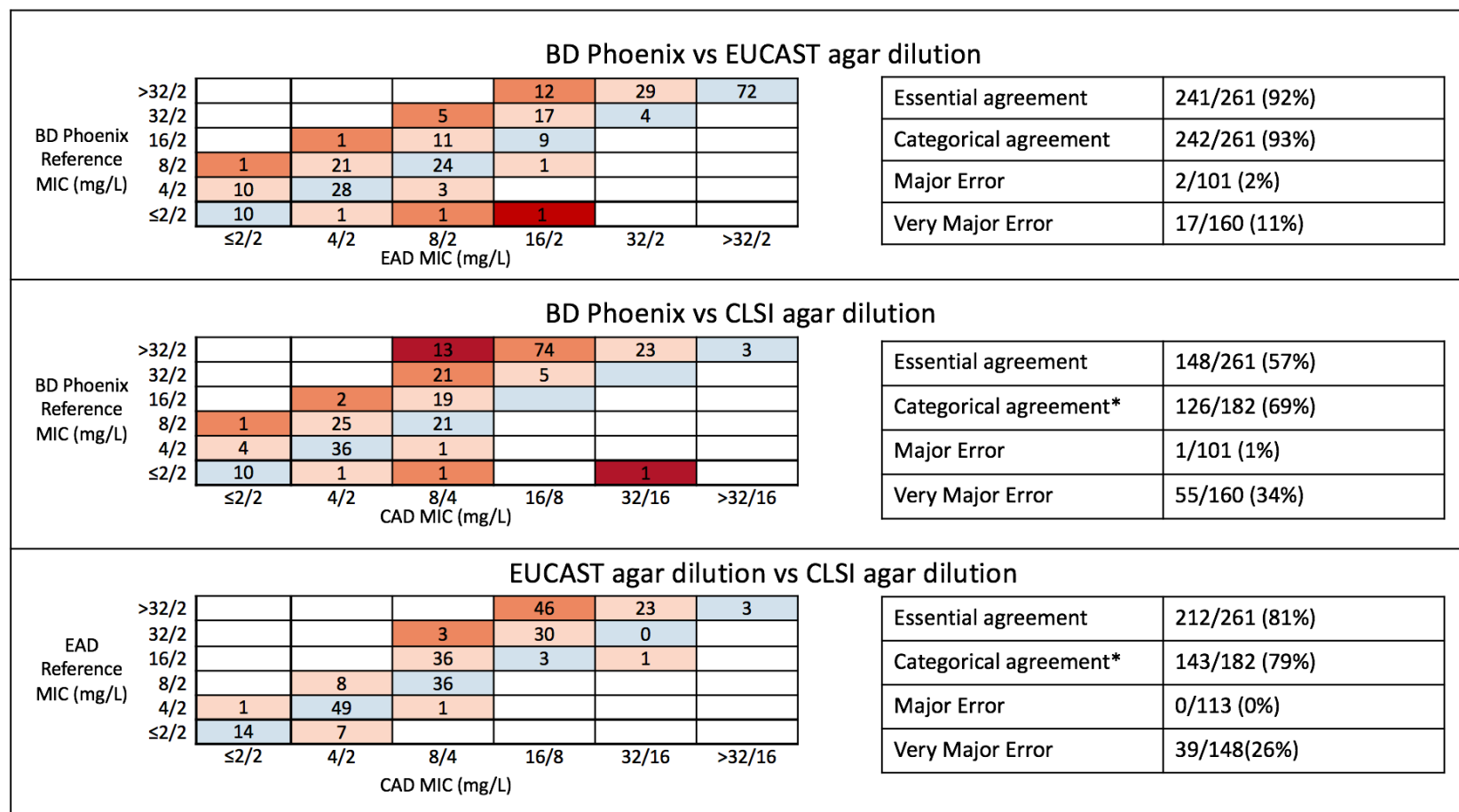


Figure 69: Comparison of MICs from different phenotypic methods

*: Categorical agreement between methods only evaluated on non-intermediate isolates as BD phoenix/EUCAST-based agar dilution have no intermediate category Note: blue: full agreement of MICs, light orange: essential agreement, dark orange: within two doubling dilutions (theoretically feasible believing both tests having an error of +/- 1 dilution) and red: disagreement

Using upper median MICs, EUCAST-based agar dilution classified 145 (56%) isolates as resistant, whereas CLSI-based agar dilution classified 27 (10%) as resistant and 79 (30%) as intermediate (total 106 [41%] CLSI non-susceptible). All CLSI resistant/intermediate isolates were EUCAST-resistant. Hence isolates could be grouped into 3 categories, 27 (10%) isolates were agreed resistant (i.e. resistant by both reference methods), 116 (44%) agreed susceptible (i.e. susceptible by both methods), and 118 'indeterminate' (EUCAST-resistant and either CLSI-susceptible (39 [15%]) or CLSI-intermediate (79 [30%])). EUCAST agar dilutions were also in closer agreement with results from the BD Phoenix (Figure 69), as expected, given the BD Phoenix is calibrated against EUCAST guidelines.

In addition to method choice, variability in MIC on repeat testing even when using the same method commonly resulted in different phenotypic classification for the same isolate (resistance category switching, Figure 70), with corresponding poor simulated within-test categorical agreement (Table 30). Although essential agreement for all tests was high, the potential for resistance category switching was greater for amoxicillin-clavulanate than for amoxicillin alone. Further, the fact that essential agreement was greater than categorical agreement for both EUCAST-based and CLSI-based agar dilution suggests that some isolates had repeat MICs which were in essential agreement but led to different resistance classification. For both CLSI-based and EUCAST-based agar dilution, VME and ME rates remained higher than FDA/ISO standards (Table 3) albeit only slightly above for CLSI. However, CLSI's apparent better performance is largely due to their intermediate category, as intermediate/resistant and Intermediate/susceptible category switches are only classed as "minor errors" for which there is no requirement in either the FDA or ISO standards (Table 3).

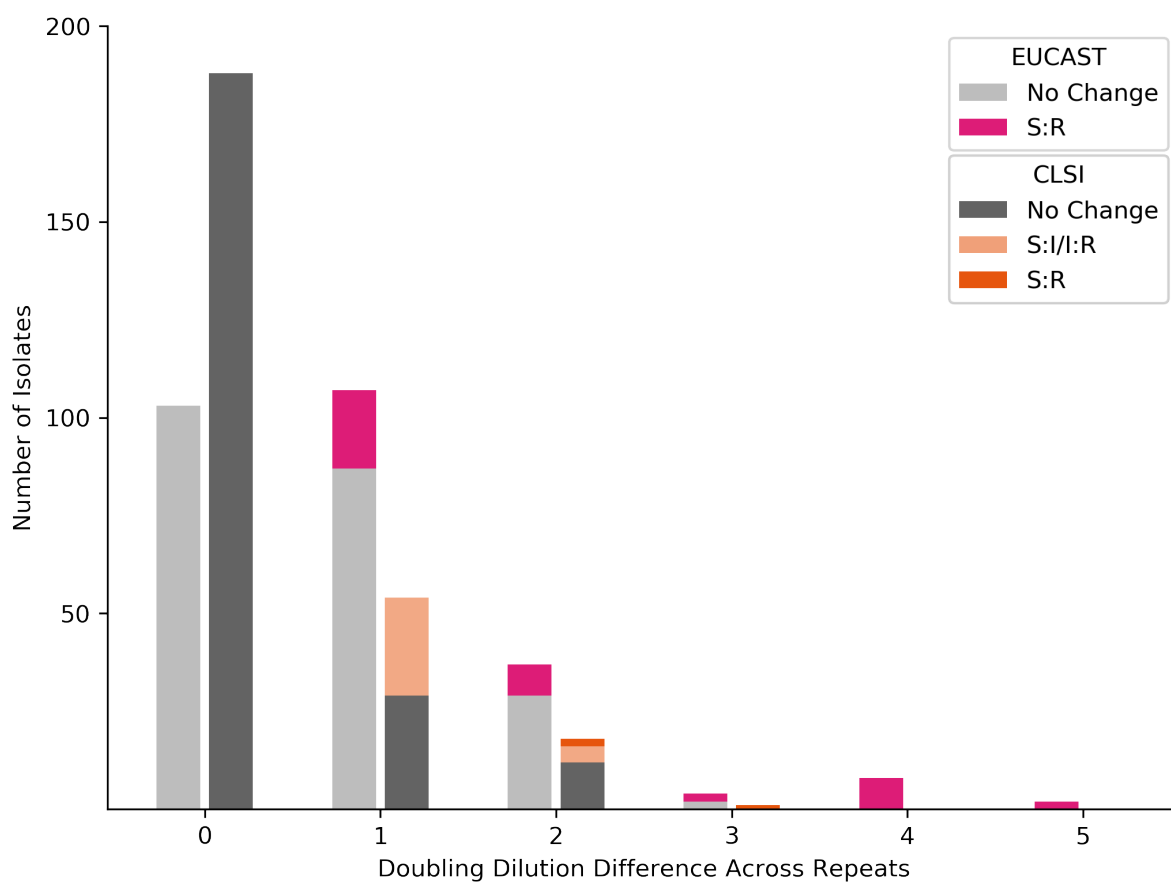


Figure 70: Resistance category switching on repeat testing

Note: Maximum MIC doubling dilution difference across repeats by method for agar dilution tested isolates

Table 30: Simulated comparisons of each test against itself as a reference

Test*	Essential agreement (95% CI)	Categorical agreement (95% CI)	Very major error (95% CI)	Major error (95% CI)
Amoxicillin	98.7 (98.1,99.6)	98.9 (98.1,99.6)	0.6 (0.0,1.3)	3.3 (0.0,6.7)
EUCAST amoxicillin-clavulanate	95.9 (94.5,97.3)	90.3 (88.5,92.3)	8.7 (5.6,12.0)	10.7 (7.0,14.6)
CLSI amoxicillin-clavulanate	98.9 (98.1,99.6)	93.2 (91.6,95.0)	2.7 (0.0,7.7)	0.4 (0.0, 1.2)

*: Agar dilution MIC values truncated to include only two MICs below breakpoint (≤ 2 mg/L) and one MIC above breakpoint (> 32 mg/L) for all tests (as per FDA guidelines ¹¹³)

7.3.3. Comparing agar dilution phenotypes with WGS-based predictions

Overall, using the extended algorithm as defined in Chapter 6, WGS classified as resistant 23/27 (85%) EUCAST/CLSI agreed resistant isolates, 107/118 (91%) indeterminate isolates (76/79 (96%) EUCAST-resistant/CLSI-intermediate, 31/39 (79%) EUCAST-resistant/CLSI-susceptible) and 17/116 (15%) agreed susceptible isolates (Figure 71). Interestingly only 8/27 (30%) of the EUCAST/CLSI agreed resistant samples contained inhibitor-resistant beta-lactamases. Conversely, 24/79 (30%) CLSI-intermediate and 10/39 (26%) CLSI-susceptible isolates contained *bla*_{OXA-1}, showing that identification of inhibitor-resistant beta-lactamases was neither necessary nor sufficient to predict CLSI resistance.

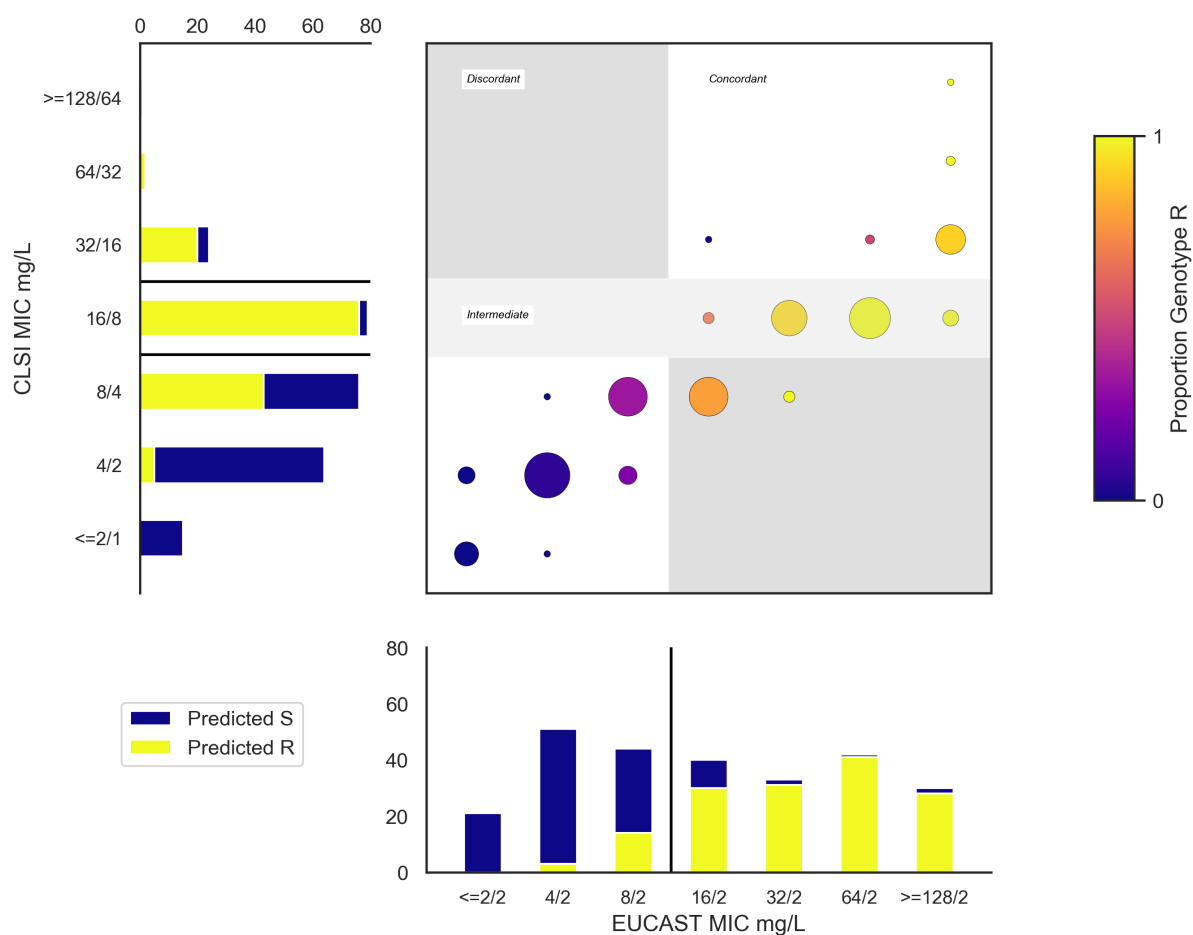


Figure 71: Proportion WGS predicted resistant (extended algorithm) by EUCAST-based and CLSI-base agar dilution MIC

Note: Sub axes show proportion of samples predicted R by MIC for each of EUCAST-based and CLSI-based agar dilution. Main axis (x,y) coordinate represents (EUCAST MIC, CLSI MIC), circle size represents the number of isolates with this combination of EUCAST-based and CLSI-based agar dilution MICs, and colour denotes proportion identified as resistant by WGS.

bla_{OXA} was not alone in having variable impact on susceptibility category. For example, *bla_{TEM-1}* also had variable impact on phenotypes even in the absence of other factors such as promoter mutations and increased gene copy number (i.e. when only the presence *bla_{TEM-1}* is thought to affect the phenotype, see section 6.2.2) (Figure 72). This variation was not solely due to strain differences, with some mechanisms having different impact on phenotype on repeat testing of the same isolate (e.g. 4/9 isolates with *ampC* promoter mutations in the agar dilution subset were both CLSI resistant and CLSI intermediate on

repeat testing). This uncertainty, combined with the fact that mechanisms commonly co-occurred in the same isolate, meant that the individual contribution of genetic features to susceptibility phenotype was often unclear from the raw data.

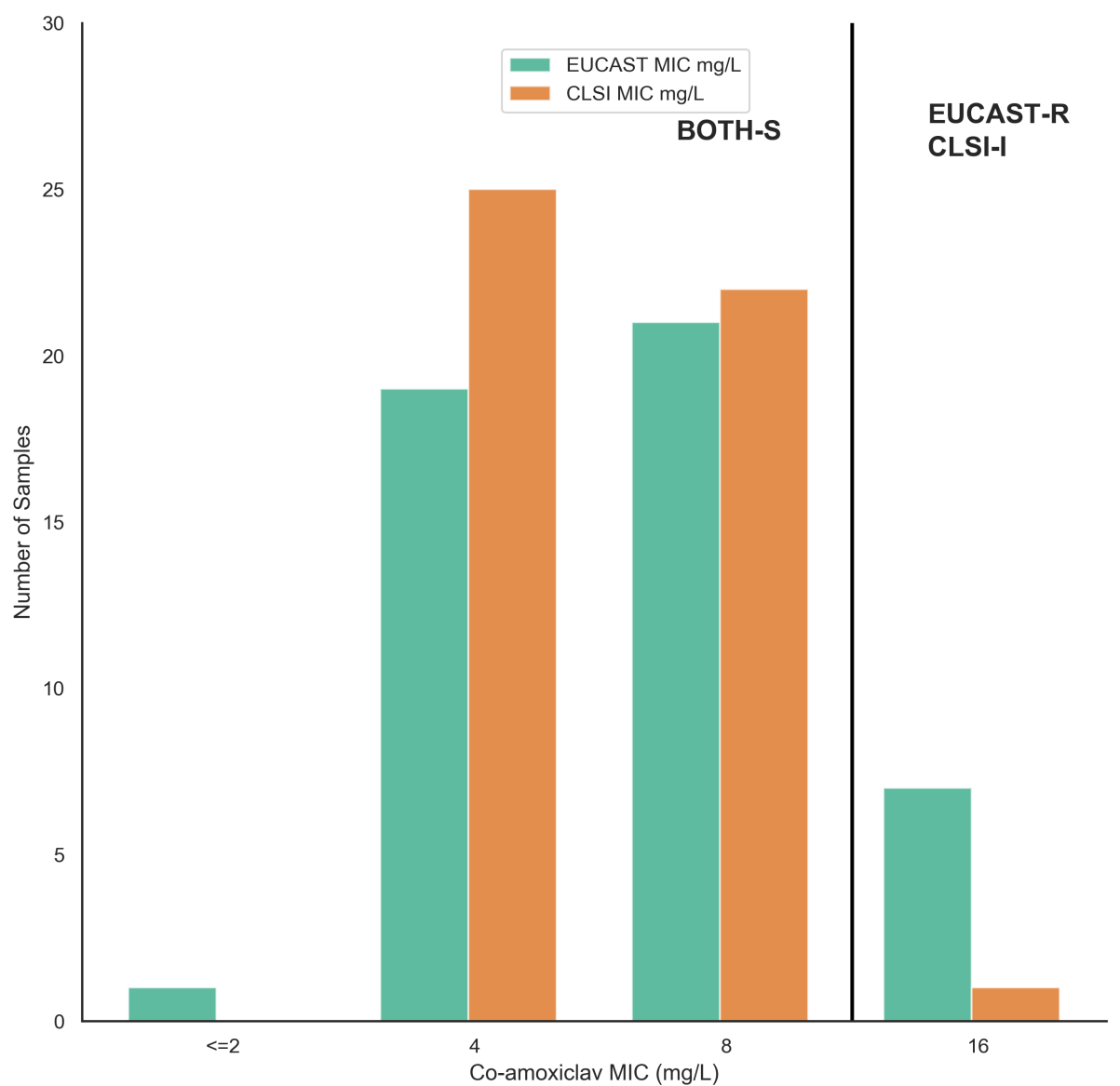


Figure 72 Variability of resistance classification of isolates containing only low* copy number *bla*_{TEM-1}

Note: *low defined as <2.5 copies, see section 6.3.2.

7.3.3. Discrepancies between WGS-predicted and observed phenotypes

As in the comparison with BD Phoenix AST results (Chapter 6), WGS predictions were more accurate for non-peri-breakpoint MICs (EUCAST $\leq 4/2$ mg/ml or $\geq 32/2$ mg/ml, CLSI $\leq 4/2$ mg/ml or $\geq 32/16$ mg/ml). For EUCAST-based agar dilution, WGS correctly identified resistance/susceptibility in 169/177 (95%) isolates with non-peri-breakpoint MICs, versus only 60/84 (71%) with peri-breakpoint MICs. Similarly, for CLSI-based agar dilution, discounting 79 intermediate isolates (16/8 mg/ml), WGS correctly predicted resistance/susceptibility in 97/106 (92%) non-breakpoint isolates, but predicted 43/76 (57%) isolates with MIC 8/4 mg/L as resistant.

Interestingly, however, there were three consistently resistant (EUCAST MIC $\geq 32/2$ mg/L, CLSI MIC $\geq 32/16$ mg/L) and three consistently susceptible (EUCAST MIC $\leq 4/2$ mg/L, CLSI MIC $\leq 4/2$ mg/L) discrepant. All three resistant discrepant were due to complexities inferring phenotype from WGS. One had a novel *bla*_{CTX-M} variant (*bla*_{CTX-M-15}-like, Ser130Gly mutation, see section 6.3.2.) which as suggested previously was likely to be inhibitor-resistant⁶⁵. The other two isolates had antibiograms consistent with *ampC* hyper-production (cefotaxime resistant, ceftazidime resistant, cefepime susceptible), with the upstream sequence only matching part of the reference (CP009072.1) promoter sequence, possibly suggesting insertion of alternative elements upstream of *ampC* which drove increased expression. All three susceptible discrepant had beta-lactamases present at mildly elevated copy numbers (2.5-3.5x relative DNA coverage, only just above the “low versus high” threshold of 2.5), leading to WGS prediction of resistance. As explained in Chapter 6, although copy number is strongly associated with resistance, there is no hard threshold perfectly delineating resistant

and susceptible isolates. Accordingly, some isolates with beta-lactamases with only mildly elevated copy number were also found susceptible.

7.3.4. Modelling MIC as a function of resistance features

Independently, all beta-lactamases were associated with increased MICs in univariable (Table 31) and multivariable models (Table 32, Figure 73). The largest effects of beta-lactamase presence/absence on MIC were for *bla*_{OXA-1} (a class 2d beta-lactamase, denoted blaOXA:2d) and members of the 'other' group of beta-lactamases, comprising either inhibitor-resistant beta-lactamases (N=10), or those with unknown impact on beta-lactam susceptibility (N=4) (Table 27). These caused two, three-fold and four-fold doubling dilution increases in EUCAST MIC respectively. Consistent with the original copy number analysis (section 6.3.2), while there was only weak evidence of presence/absence effect on MIC of class 2b *bla*_{TEM} (denoted blaTEM:2b) (CLSI p=0.01, EUCAST p=0.52) and *bla*_{SHV} genes (denoted blaSHV:2b) (CLSI p=0.92, EUCAST p=0.26), there was stronger evidence of increased copy number being associated with increased MIC for both testing methods (i.e. a gene dosage effect). Consistent with unadjusted analyses, promoter mutations were associated with increased MICs (p<0.0001). However, there was no clear increase in MIC independently associated with suspected porin loss (p≥0.06).

Table 31: Unadjusted effect of resistance features on agar dilution log₂ MIC

Element	Testing method	Mean log ₂ (MIC) (95% CI)	log ₂ (MIC) change ^o	Corresponding MIC range covering the 95% CI (mg/L)*
No element/feature	CLSI	1.75 (1.32, 2.17)	-	2-8
	EUCAST	1.71 (1.20, 2.21)	-	2-8
blaTEM:2b	CLSI	3.32 (3.17, 3.48)	1.58 (1.22,1.93)	8-16
	EUCAST	4.23 (4.02, 4.56)	2.59 (2.02,3.15)	16-32
blaOXA:2d	CLSI	3.74 (3.57, 3.92)	1.99 (1.47,2.51)	8-16
	EUCAST	5.20 (4.91, 5.49)	3.49 (2.86,4.13)	16-64
blaCTXM:2be	CLSI	3.49 (3.24, 3.73)	1.74 (1.21,2.26)	8-16
	EUCAST	4.68 (4.20, 5.15)	2.97 (2.26,3.67)	16-64
blaSHV:2b	CLSI	2.30 (1.95, 2.60)	0.55 (-0.40,1.50)	2-8
	EUCAST	2.40 (1.50, 3.30)	0.69 (0.48,1.86)	2-16
Other_bla	CLSI	4.53 (4.01,5.07)	2.79 (1.93,3.65)	16-64
	EUCAST	6.69 (5.75,7.62)	4.98 (3.91,6.05)	128-256
ampC promoter mutation	CLSI	5.11 (4.85,5.37)	3.36 (2.36,4.36)	16-64
	EUCAST	7.11 (6.65,7.57)	5.40 (4.21,6.59)	128-256
blaTEM promoter mutation	CLSI	4.27 (4.01,4.52)	2.52 (1.94,3.09)	16-32
	EUCAST	6.00 (5.58,6.42)	4.29 (3.58,5.00)	64-128
Suspected loss of function in ompC/ompF	CLSI	4.36 (3.61,5.11)	2.61 (1.66,3.57)	8-64
	EUCAST	6.27 (5.27,7.27)	4.56 (3.42,5.71)	64-256

*: MIC interpretation of the 95% CI

^o: log₂(MIC) change relative to the no-element/feature for each test method. E.g. CLSI log₂(MIC) change for CLSI is expressed relative to the CLSI no-element/feature MIC

Table 32: Independent predictors of agar dilution log₂(MIC)

Element	Modelled as	Test method	Change in log ₂ (MIC) (95% CI)	P	Heterogeneity between CLSI and EUCAST
Overall					
-		CLSI EUCAST	- 0.09 (-0.08,0.26)	- 0.29	
Transmissible					
blaTEM:2b	Presence/absence	CLSI EUCAST	0.37 (0.08,0.66) 0.15 (-0.26,0.55)	0.01 0.48	0.05
blaTEM:2b:cpn	Per doubling of copy number	CLSI EUCAST	0.38 (0.29,0.48) 0.79 (0.66,0.93)	<0.0001 <0.0001	<0.0001
blaCTXM:2be	Presence/absence	CLSI EUCAST	0.89 (0.42,1.37) 0.99 (0.41,1.57)	<0.0001 0.001	0.44
blaOXA:2d	Presence/absence	CLSI EUCAST	1.62 (1.17,2.07) 2.74 (2.17,3.31)	<0.0001 <0.0001	<0.0001
blaSHV:2b	Presence/absence	CLSI EUCAST	0.04 (-0.71,0.79) -0.60 (-1.67,0.48)	0.91 0.28	0.03
blaSHV:2b:cpn	Per doubling of copy number	CLSI EUCAST	0.28 (-0.06,0.61) 0.71 (0.22,1.19)	0.10 0.004	0.001
Other_bla	Presence/absence	CLSI EUCAST	2.13 (1.70,2.55) 3.79 (3.18,4.41)	<0.0001 <0.0001	<0.0001
Non-transmissible					
ampC promoter mutation (ampCpr)	Presence/absence	CLSI EUCAST	2.61 (2.10,3.12) 4.26 (3.52,5.00)	<0.0001 <0.0001	<0.0001
Suspected loss of function in ompC/ompF (NFOMP)	Presence/absence	CLSI EUCAST	0.26 (-0.21,0.72) 0.63 (-0.04,1.30)	0.28 0.06	0.04
Additional Interactions					
TEM promoter mutation (blaTEMpr)	Presence/absence	CLSI EUCAST	1.27 (0.97,1.57) 2.43 (2.00,2.86)	<0.0001 <0.0001	<0.0001
CTX-MxTEM	Both present		-0.89 (-1.42,-0.36)	0.001	-
CTX-MxOXA	Both present		-0.63 (-1.27,0.01)	0.05	-

Note: For categorisation groups see methods; presented visually in Figure 73.

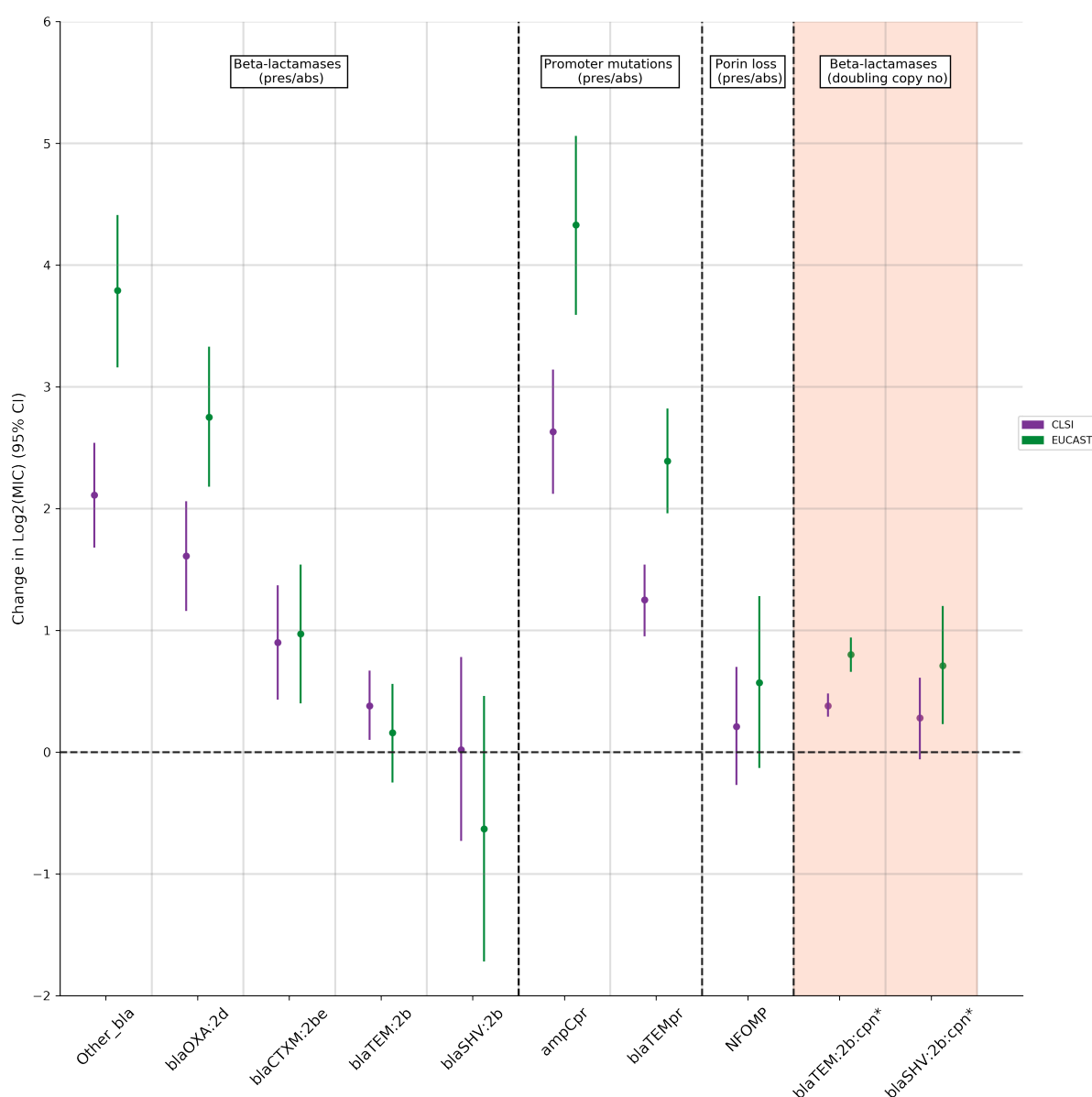


Figure 73: Changes in doubling dilution MIC independently associated with each feature/testing method

Note: Estimates obtained using a multivariable mixed model. Purple represents testing using CLSI-based agar dilution (amoxicillin/clavulanate 2:1 ratio)), and green EUCAST-based agar dilution (fixed 2mg/L clavulanate). All elements except those denoted by * and shaded in orange are modelled as binary presence vs absence effects (Table 28). Other_bla (grouped other beta-lactamase genes), includes *bla*_{TEM-40} (N=2), *bla*_{TEM-30} (N=3), *bla*_{CMY-2} (N=3), *bla*_{OXA-48} (N=1), *bla*_{TEM-190} (N=1), *bla*_{TEM-33} (N=1), (Table 27); blaOXA:2d (Bush-Jacoby 2d, *bla*_{OXA}), blaCTXM:2be (Bush-Jacoby 2be, *bla*_{CTX-M}), blaTEM:2b (Bush-Jacoby 2b, *bla*_{TEM}), blaSHV:2b, (Bush-Jacoby 2b, *bla*_{SHV}), ampCpr (*ampC* promoter mutation suggesting increased expression), blaTEMpr (*bla*_{TEM} hyper-producing promoter), NFOMP (non-functional *ompF/ompC*), blaTEM:2b:cpn (*bla*_{TEM} copy number) effect modelled as effect of doubling

copy number, blaSHV:2b:cpn (*bla*_{SHV} copy number) effect modelled as effect of doubling copy number.

In addition to showing the independent effects of genetic features, the model also highlighted the importance of testing method. Other than for suspected porin loss and *bla*_{CTX-M} genes, the impact of genetic resistance features on MIC was accentuated when using EUCAST testing methodology ($p_{\text{heterogeneity}} \leq 0.05$). However, this increased sensitivity for genetic mechanisms when using EUCAST methodology came at the expense of increased test variability (Table 33).

Table 33: Estimated variation in MIC according to different sources from the random effects model

Variance Component	Test Method	Standard Deviation in MIC (95% CI)
Between experiment run (batch)	CLSI	0.12 (0.07,0.20)
	EUCAST	0.20 (0.12,0.32)
Within isolate	CLSI	0.38 (0.36,0.40)
	EUCAST	0.68 (0.64,0.72)
Between isolate	CLSI	0.71 (0.64,0.78)
	EUCAST	0.99 (0.89,1.09)

Note: Between experiment run: variation due to differences in procedures between experimental runs (batches). Within isolate: variation intrinsic to isolates between experimental run. Between isolate: unexplained between isolate variation

7.3.5. Predictions of MIC in an independent validation set

The fixed effects from the genotype-based model could be used to predict AST MIC for 704/715 non-subsample isolates (excluding 11 isolates with genetic features not present in the training set and therefore not included in models). Predicted MIC agreed with observed AST MIC for 554/704 (79%) isolates (Figure 74) and was within one doubling dilution for 684/704 (97%) isolates. Including the 11 isolates with resistance features not present in the

agar dilution subsample (i.e. assuming these features had no effect on MIC since they were not included in the model) had little adverse impact on prediction performance, with MIC predictions being in agreement for 557/715 (78%) isolates and within one doubling dilution for 693/715 (97%). Further, although categorical agreement between WGS-predicted and observed phenotype (90%) was lower than essential agreement (97%), it was similar to categorical agreement between EUCAST and BD Phoenix (both based on EUCAST guidelines, 93% (Figure 69)) and better than categorical agreement between the EUCAST and CLSI agar dilution methods (79%). Three isolates had predicted MICs three doubling dilutions lower than observed. One had an inconsistent phenotype (ampicillin susceptible, co-amoxiclav resistant), suggesting a laboratory testing problem. The other two both had observed MIC $\geq 32/2$ mg/L but only contained a low copy number *bla*_{TEM-1} and had predicted MIC 8/2mg/L.

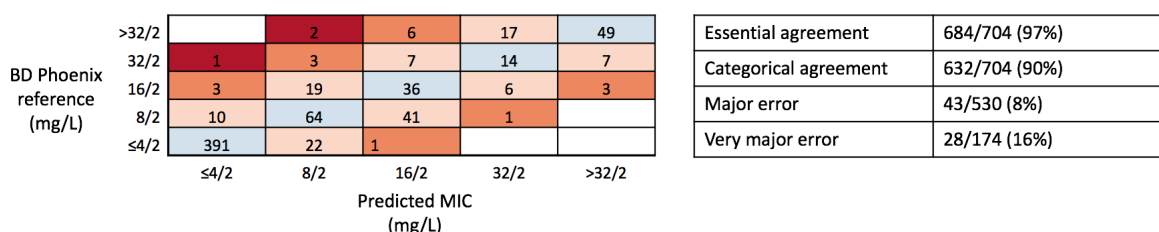


Figure 74: Model based MIC prediction for non-subsample isolates (N=704)

Note: blue shading indicates correctly predicted observed AST MIC (554/704 (79%) isolates), light pink predicted within one doubling dilution (total 684/704 (97%) isolates, essential agreement), orange within two doubling dilutions (total 701/704 (100%)) and red greater than 2 double dilutions different.

7.4. Discussion

This study highlights how differences in phenotype testing and breakpoint definition can have significant impact on the observed phenotype of isolates. For amoxicillin-clavulanate, not only were phenotypes derived using the two “gold standard” methods often discrepant,

each method was highly variable, with both often being inconsistent. Unsurprisingly, this hampered concordance between binary WGS-based predicted and observed phenotypes. However, by using a genotype-based model for MIC which accounted for these issues, WGS was able to predict observed MICs to within one doubling dilution (essential agreement) for 97% of isolates in an independent testing set of 704 population-representative *E. coli* BSI isolates. Returning to definitions in section 2.3.1, the prediction of amoxicillin-clavulanate resistance required a move away from the original “rule-based” approach to WGS-based prediction to a more “database-guided” approach, where the impact of resistance features can be more variable and potentially additive. Further, the model suggested a possible cause for the variable phenotypes. It highlighted that amoxicillin-clavulanate resistance was only rarely due to single elements. Instead, it was often multifactorial, resulting from combinations of multi-copy beta-lactamase genes, mutations in resistance gene-associated promoters, and inhibitor resistance (resistance to clavulanic acid inactivation). The impact of each of these features on MIC was often small, variable and potentially additive. However, these small changes in MIC could often be outweighed by testing variation, resulting in variable phenotypes, particularly since the median MIC in the population was peri-breakpoint.

In this work, consistent with previous studies^{233,234,237}, the amoxicillin-clavulanate AST phenotype was not only dependent on method, but was also highly variable on repeat testing. Both reference methods had significant weaknesses meaning neither could reliably be taken as a “gold standard”. While the CLSI-based method (2:1 ratio of amoxicillin and clavulanate concentrations) had better internal agreement and lower variability of classification, the clavulanate concentrations it uses are higher than physiologically seen²¹³.

Further, CLSI classifies over a third of isolates as intermediate, giving clinicians no definitive information as to the effectiveness of treatment. In contrast, the EUCAST-based method (fixed 2mg/L clavulanate concentration), while using more physiologically relevant clavulanate concentrations and classifying all isolates, was far more conservative than CLSI (with all CLSI resistant and intermediate isolate being EUCAST resistant) and its classifications were more variable on repeat testing. These issues, combined with the scale of difference in interpretations between the two methods, demonstrate definitively that there is significant uncertainty in the AST of one of the most commonly used antibiotics in Europe, as previously highlighted ²⁰¹.

The problem of phenotype interpretation varying on repeat testing was not restricted to EUCAST-based methodology but was present by both methods, and instead highlights a weakness of categorising isolates according to breakpoints. For some antibiotics, and in some sets of isolates, isolates can be categorised neatly into well-separated susceptibility categories according to their MICs (e.g. categorising *E. coli* according to trimethoprim MIC (Figure 75) or even amoxicillin (Figure 62)).

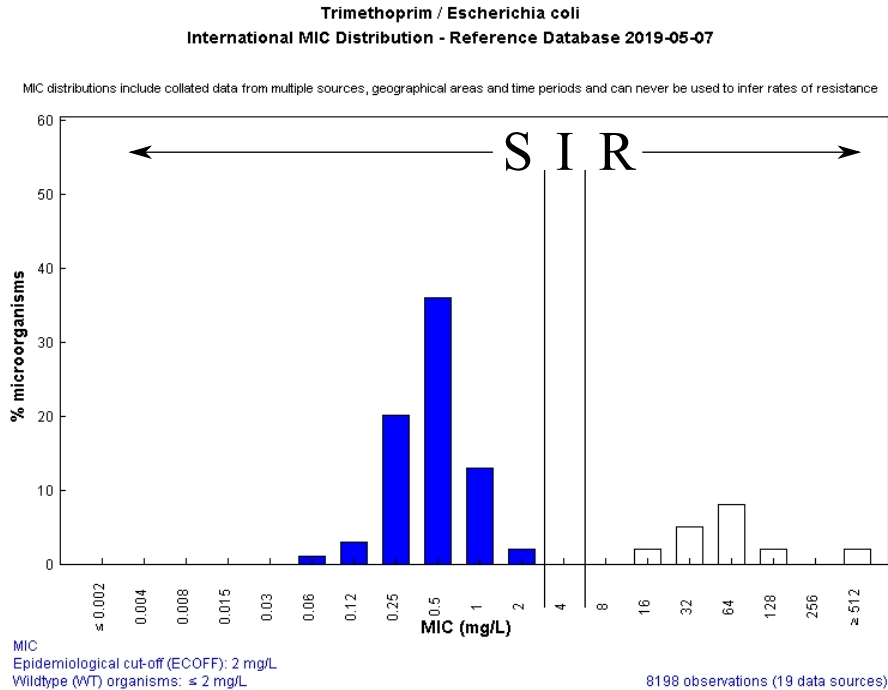


Figure 75: Distribution of MICs of *E. coli* to trimethoprim.

Note: Blue: wild-type, white: non-wild-type (i.e. containing a resistance mechanism).
Adapted from EUCAST ²³¹.

In these instances, the MIC distribution is bimodal, and isolates can be separated into two populations (the resistant (R) and susceptible (S) populations), with no observations in the interval between the MIC distributions for these two discrete populations. This space in between prevents variability associated with phenotype testing (typically quoted to be +/- one doubling dilution) from altering the phenotypic classification. However, because it is used in clinical interpretation to simplify laboratory reports for non-specialists, this resistant/susceptible nomenclature is also applied when categorisation is much less meaningful.

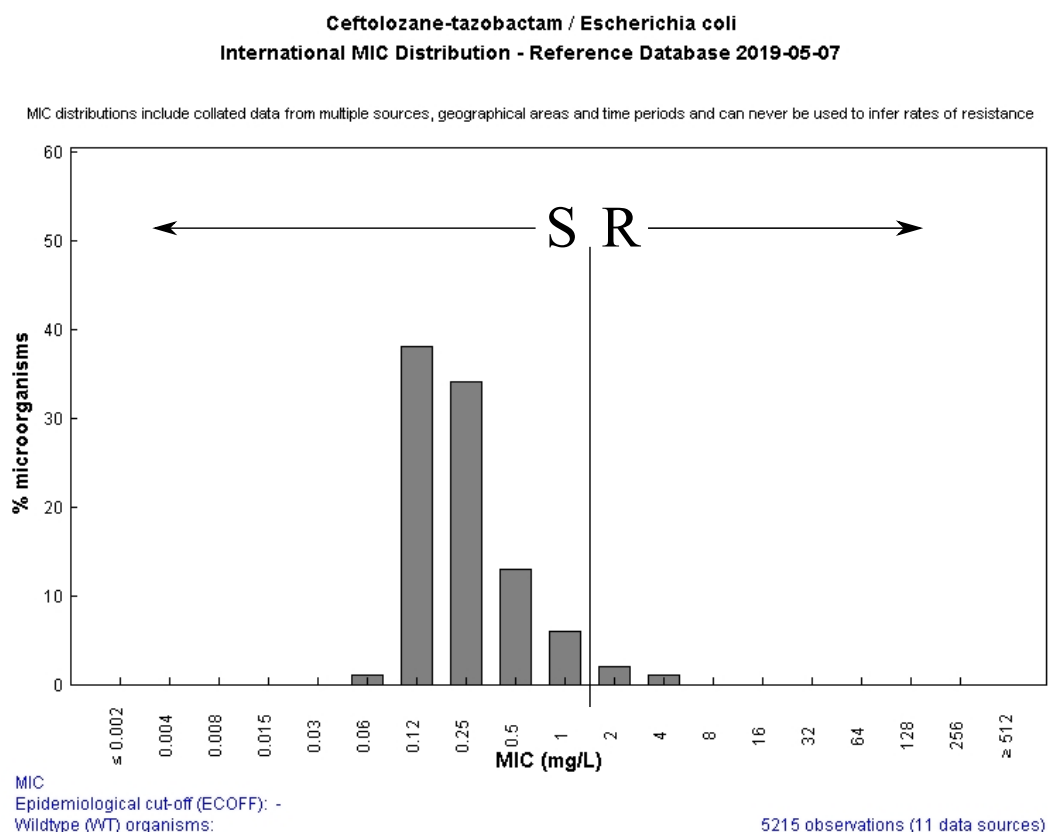


Figure 76: Distribution of MICs of *E. coli* to ceftolozane-tazobactam

Note: No split of population into wild-type/non-wild-type categories. Adapted from EUCAST 231.

In these cases, such as for amoxicillin-clavulanate (Figure 62) and ceftolozane-tazobactam (Figure 76), there is no clear separation between susceptible and resistant isolates, with the distribution instead being “heavy-tailed”, suggesting that mixtures of resistant and susceptible populations overlap. The expected test-related variation can therefore cause changes in phenotypic classification for any arbitrary threshold contained within the distribution. As was seen in this study, this results in poor categorical agreement and unacceptable VME and ME rates despite high essential agreement. Given isolates are not routinely multiply tested, it is inevitable that many isolates will be “misclassified”, reducing the clinical meaning of the categorisation. This misclassification may not be important for individual patient tests given clinical breakpoints incorporate a large margin of safety,

however this makes monitoring trends and causes of resistance difficult as it obscures which isolates are genuinely resistant. Further, this uncertainty and its resultant apparent high resistance rate likely drives switches of treatment to alternative antimicrobials, further propagating resistance. This also prevents meaningful comparison of test-systems based on categorical agreement and VME/ME rates, which further raises concerns regarding the validity of AST methods. This is particularly worrying, given these issues threaten to extend to novel BL/BLI combinations ²⁴¹, (despite the new EUCAST guidelines ²³⁵) which are some of the few antibiotics able to treat multi-drug-resistant Gram-negative bacteria ⁸².

The modelling approach used in this study had several advantages. It allowed a quantitative assessment of the impact of genetic determinants on amoxicillin-clavulanate MIC.

Explaining the issues with initial genotype-phenotype correlations, it suggested, consistent with previous studies ²⁴², that resistance to amoxicillin-clavulanate, instead of being due to large effects from a small number of individual mechanisms, was often due to combinations of effects across multiple mechanisms, each of which had smaller individual impacts on MIC. Further, with many putative mechanisms of amoxicillin-clavulanate resistance not being directly observable by DNA-sequencing (e.g. enzyme hyper-production), modelling of associations between resistance features and MIC directly avoided inferring the phenotype from the genotype using uncertain pre-specified rules. However, the complexity of the underlying associations I discovered in the isolates analysed here highlights the challenges facing standardized methods for predicting resistance for multiple drugs and species.

As my model was successful at predicting amoxicillin-clavulanate MIC from routine AST, this chapter suggests that the remaining 20% error seen in Chapter 6 was likely due to issues

with the binary phenotype, particularly in isolates with peri-breakpoint MICs. As a corollary, most of the resistance seen can be reliably attributed to known resistance factors, albeit modified by additional features such as promoters (e.g. *bla*_{TEM} promoter mutations) and DNA copy-number. Worryingly, these features have not been assessed in previous studies on WGS-based resistance prediction^{28,141,158}, likely due to the added complexity of identifying them. Consequently, the role these mechanisms play in transmissible beta-lactam resistance worldwide is unclear, leading to significant uncertainty of how resistance to some of the most commonly used antibiotics spreads. Further, as these mechanisms can impact many other beta-lactams (including new BL/BLI combinations)^{82,243}, the life-span of new beta-lactam derived antibiotics may be limited if these mechanisms are widespread.

This study demonstrates the importance of accounting for not only isolate sampling frame, as discussed in section 4.4, but also phenotyping method and breakpoint selection when assessing the ability of WGS to identify resistance. For example, a previous study of 76 *E. coli* isolated from cattle¹²⁰, which reported high sensitivity and specificity of WGS to predict amoxicillin-clavulanate resistance, as well as only containing selected more resistant isolates, also only attempted to predict CLSI-defined resistance (>32/16 mg/L). This previous study therefore adds even less evidence as to whether WGS can reliably predict important resistance in routine human isolates, as not only are mechanisms seen in its isolates highly different from population-representative studies of human clinical isolates²²², it also only compares against a breakpoint with concentrations not obtainable *in vivo*.

Consequently, some of the key strengths of this work are the unbiased, large, population-representative sampling frame and the detailed, replicated, reference-grade phenotyping. A

major limitation of this study was its size (constrained by phenotyping), resulting in limited representation of some rarer resistance elements. This meant firstly, that some infrequent features had to be categorized together for modelling, and secondly, interactions could only be assessed in a limited way. Reassuringly, however, the features causing the greatest MIC increases were those traditionally associated with co-amoxiclav resistance ⁶⁵. Another limitation is that 11 isolates had to be excluded from my “testing set”, as they contained elements not seen in my training set. While this would be helped by larger studies, similar issues will likely persist given the rate of occurrence/discovery of novel resistance features. This problem is analogous to the problem of predicting resistance caused by the novel *bla*_{CTX-M} not contained in the Resfinder database and the need to utilize knowledge about mechanisms of inhibitor resistance as opposed to relying solely on observed data when predicting resistance in isolates with previously unobserved resistance features.

While demonstrating the feasibility of WGS-based resistance prediction for amoxicillin-clavulanate once more realistic representations of resistance genotype and phenotype are used, this work suggests that the utility of WGS-based resistance prediction in Gram-negatives as a whole is far from clear. Rather, the amoxicillin-clavulanate prediction challenge represented a necessary and informative hurdle WGS-based prediction needed to clear prior to adoption in clinical practice, but it is only a single antibiotic acting on a single antimicrobial. Ongoing work will need to examine a much broader range of antibiotics and antimicrobials before WGS-based prediction can be deemed viable.

Chapter 8: Discussion

8.1. Introduction

Since the start of this doctorate, many groups have focussed on advancing WGS-based resistance prediction^{28,125,142,158}, while others have worked to improve understanding of resistance to beta-lactam/beta-lactamase inhibitor antibiotics^{93,221,243,244}. The work in this thesis is complementary to both, straddling a middle ground between genetically orientated computational biology and more classical pharmacology and microbiology. In contrast to the other groups working on WGS-based resistance prediction, it explores details of suboptimal concordance in a more in-depth and biologically minded way, investigating issues with the rule-based model of resistance and its inherent assumptions. Compared to other work on BL/BLI antibiotics, it uses larger scale sample sizes and statistical modelling to explain the poorly understood phenomenon of mechanisms giving variable AST results. Overall, the results highlight a gap between current resistance prediction performance and what is needed for regulatory approval, and suggest this is most likely due to issues with resistance prediction algorithms. More fundamentally however, they highlight issues with traditional paradigms of resistance classification. These are not only important to take into account in WGS-based resistance prediction, but also likely lead to significant uncertainty in current clinical practice.

8.2. Summary of findings

8.2.1. Predicting beta-lactam antibiograms using gene-presence absence alone

While WGS-based beta-lactam resistance prediction based on beta-lactamase presence/absence alone was sufficient for some antibiotics, for others it performed poorly. This contrasted against the initial proof-of-principle studies of WGS-based resistance prediction ^{28,141}, and may be due to selected and limited nature of the isolates they used and antibiotics they tested. More recent larger, less selected studies show findings similar to those in this thesis ²⁴⁵. Interestingly, ampicillin was among the antibiotics for which resistance prediction performed well. As its spectrum is approximately contained within that of all other antibiotics studied, detecting this resistance is almost equivalent to detecting any beta-lactam resistance in Gram-negatives. This finding, combined with the fact that beta-lactam antibiograms of a quarter of isolates tested could not possibly be predicted using the initial rules-based algorithm (established from beta-lactamase Bush Jacoby functional classes), suggested issues with the resistance prediction algorithm. Importantly, resistance prediction based on these classes is too broad to accurately reflect the complexity of beta-lactam antibiograms.

8.2.2. The choice of WGS-based method

Consistent with previous studies, WGS-based resistance prediction was relatively unchanged by the choice of bioinformatic approach ¹⁴⁰. However, looking at the phenotype level masked larger differences in reported genotype. In particular, I found that the structure and curation of sequence databases could cause common, recurring errors which could drastically disrupt which resistance genotypes were identified. Secondly, generalised

methods such as using DNA-matching and database clustering may not be appropriate for resistance genotyping. These methods have implicit assumptions, such as the optimal DNA-level match being the same as the optimal protein-level match, and that genes in a family of related alleles can be clustered using a singular sequence similarity threshold. Finally, all bioinformatics methods struggled to distinguish between two similar resistance elements within the same sample. Currently these issues have little impact on identifying the resistance prediction, but risk becoming more serious as either resistance databases or resistance genotypes encountered become more complex.

8.2.3. Accounting for alternate mechanisms of beta-lactam resistance

The beta-lactamase presence/absence only model is a poor representation of resistance for some beta-lactams in Gram-negatives (particularly for BL/BLI antibiotics¹⁵⁸ and carbapenems²⁴⁶) as it fails to account for known effects which increase beta-lactam resistance such as those which increase expression or reduce permeability. These features are complex to study as they include functional aspects of genes which are not directly identifiable from WGS data²³⁰. However, for amoxicillin-clavulanate resistance, simple additions such as comparing sequences immediately upstream of beta-lactamases with important promoters capable of gene hyper-production and using DNA coverage as an estimate of resistance gene copy number significantly improved resistance prediction. However, despite these improvements, sensitivity and specificity were still below FDA/ISO limits, and some of these new rules failed to perform perfectly. In particular, the rule based on linking resistance with plasmid copy number had moderate PPV ($\approx 80\%$), likely reflecting errors in copy number estimation. However, most errors occurred in isolates with peri-breakpoint MICs, and further, there was a strong positive association between copy number

and observed MIC. The cause of the remaining discrepancies was hypothesised to either be due to phenotyping errors in peri-breakpoint isolates, or the effects of some genetic features being too small and variable to consistently result in different observed phenotypes.

8.2.4. Phenotype method dependency of WGS versus AST comparisons

Amoxicillin-clavulanate phenotype was assessed using two supposedly equivalent “gold standard” methods. Consistent with previous studies, resistance interpretation from these two methods was markedly discrepant, and for each specific method, repeat testing often caused a change in resistance interpretation^{233,234}. However, for both reference methods, essential agreement of MICs on repeat testing was within acceptable FDA/ISO limits, suggesting resistance classification was the primary problem with phenotype, rather than the MICs themselves. I assumed that interpreting the median MIC would produce a more reliable resistance classification for isolates. However, resistance prediction remained poor even against this more reliable “binary phenotype” no matter which genetic “rules” were used. Further there was no obvious way to align the binary genotype produced by the rules-based approach with the binary amoxicillin-clavulanate phenotype and the impact of some specific genetic mechanisms on each phenotype was inconsistent. This therefore favoured the hypothesis that the effects of some genetic features were too small to have a reliable impact on binary phenotype.

8.2.5. Quantitative approaches

Linear random-effects modelling of MICs demonstrated independent effects of most beta-lactam resistance mechanisms without the use of “rule-based” inference. Importantly, as

suspected, the effects of some genetic features were small, and could be masked by the relatively larger test retest variability. This therefore explained how certain mechanisms had varying resistance classification on repeat testing. With these mechanisms being associated with MICs just above the breakpoint, the ± 1 doubling dilution error tolerated by AST was sufficient to potentially move MICs below the breakpoint. The fixed component of these linear models was highly successful at predicting MICs in an independent test set of isolates, suggesting that the model and its inferences were somewhat robust, and further, suggesting a WGS-based approach which predicted MICs could overcome issues caused by a oversimplistic binary phenotype by predicting test-specific MICs.

8.3. Interpretation

8.3.1. WGS-based resistance prediction and assumptions about resistance

While not explicitly stated by any of the previous studies of WGS-based resistance prediction, the following assumptions underpin this approach.

- For any isolate and antibiotic, the isolate can be classed uniquely into a resistance category.
- This resistance category then has several attributes
 - It is completely determined by the isolate's DNA
 - It can be accurately measured by AST
 - It determines the likelihood of treatment success when using the antibiotic to treat an infection caused by the isolate (e.g. ampicillin is unlikely to be able to treat an ampicillin 'resistant' *E. coli* infection).

Of note, these assumptions can be defined for any number of resistance categories, but for the sake of simplicity in this discussion, from here only the two category case will be discussed, i.e. isolates being resistant or susceptible to a given antibiotic.

These assumptions are highly useful when studying bacterial resistance genomics. Most importantly, the deterministic model they suggest of bacterial resistance allows for easy interpretation of genetic mechanisms (Figure 77). Further, as they allow patient outcome to be inferred from AST results, they avoid the technical and ethical complexity introduced in measuring patient outcomes directly. In addition to simplifying interpretation, these assumptions help analysis, as they imply an easy to obtain “gold standard” measure of resistance, and method performance can be assessed by easily understandable concepts such as VME/ME rates and sensitivity/specificity.

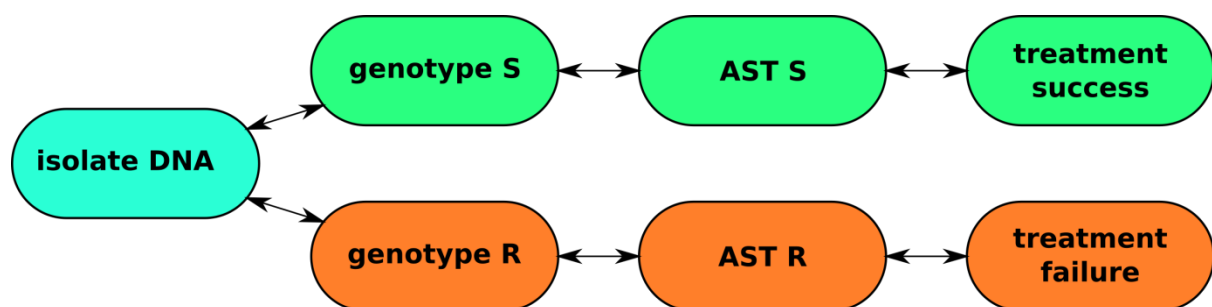
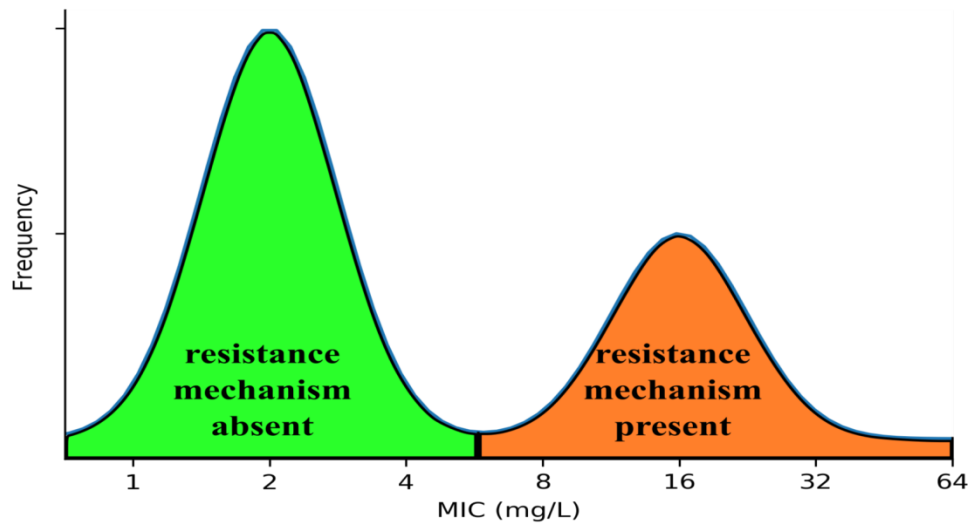


Figure 77: Assumption-based inference from genotype to outcome

While the link between AST and treatment outcome is less well understood, for many resistances, this model of the link between genotype and AST may be a good approximation of reality. For example, in some situations, such as for trimethoprim resistance (Figure 78), high level resistance is conferred by a single enzyme or mutation, with isolates otherwise

being susceptible. In this case, not only is the resistance genotype clear, but it translates directly to the phenotype.

a)



b)

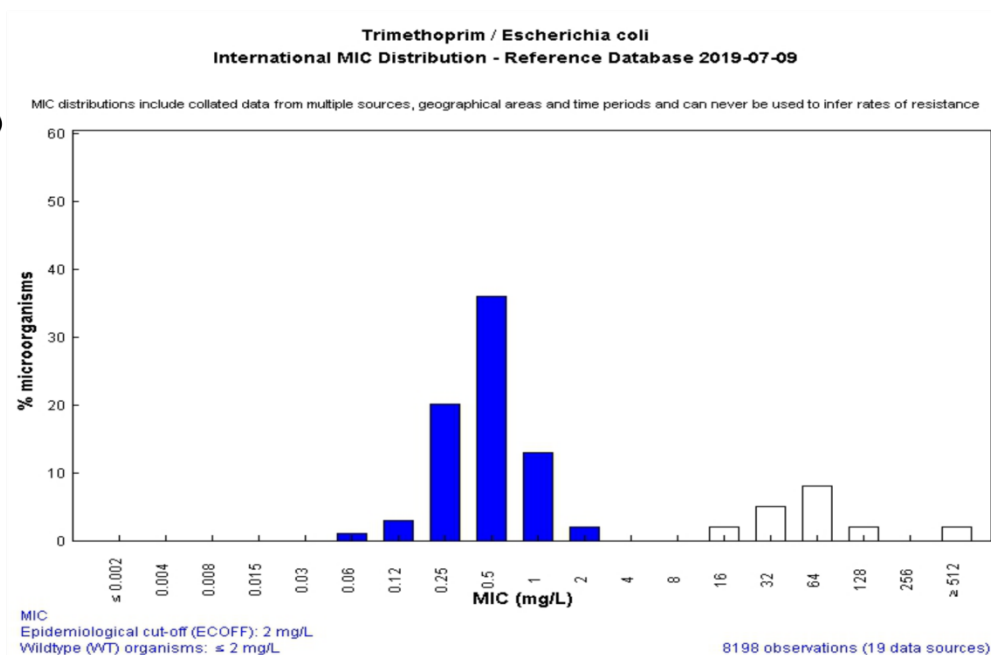


Figure 78: MIC distributions under a “single-mechanism” model.

Note: A) Theoretical separation of MIC distributions if a single mechanisms is capable of producing high level resistance. Under this model, ECOFF and likely clinical breakpoints can be easily selected just by analysis of the MIC distributions. B) The situation occurring in trimethoprim resistance, which in *E. coli* is primarily due to antibiotic target modification via presence/absence of a transmissible modified dihydrofolate reductase (*dfr*) gene^{231,247}.

Further as resistance is less likely to be misclassified in this situation, clinical effects of treatment will be easier to detect. However, while this model may be the expected “norm”,

it is not true for every case. Rule-based WGS prediction however somewhat relies on this paradigm.

8.3.2. Models of amoxicillin-clavulanate resistance

Amoxicillin-clavulanate resistance in *E. coli* clearly does not follow this expected “norm”. In this thesis I have shown that multiple genetic mechanisms contribute, in an additive fashion, to *E. coli*’s ability to evade the action of amoxicillin-clavulanate. Further, these mechanisms often have smaller, context dependent effects on isolate phenotype. This situation leads to several problems for the initial assumptions which, in turn, significantly increases complexity.

In this “multi-gene” situation, defining and categorising resistance is no longer straightforward. Resistance definitions such as wild-type and non-wild-type (and hence ECOFFs) rely on being able to define a genetic feature as causing resistance. The small effect of individual features however can make this challenging²⁴⁸. Isolates with some of these features may have MICs similar to “wild-type” isolates with above average MICs. While the cause of the higher than normal “wild-type” MICs could be identified and re-defined as “non-wild-type”, there is a risk that this process could continue indefinitely, with each step reducing the of meaning of this classification. Secondly, these resistances are likely to be harder to phenotypically classify as this situation naturally results in the loss of bimodal MIC distributions (Figure 79).

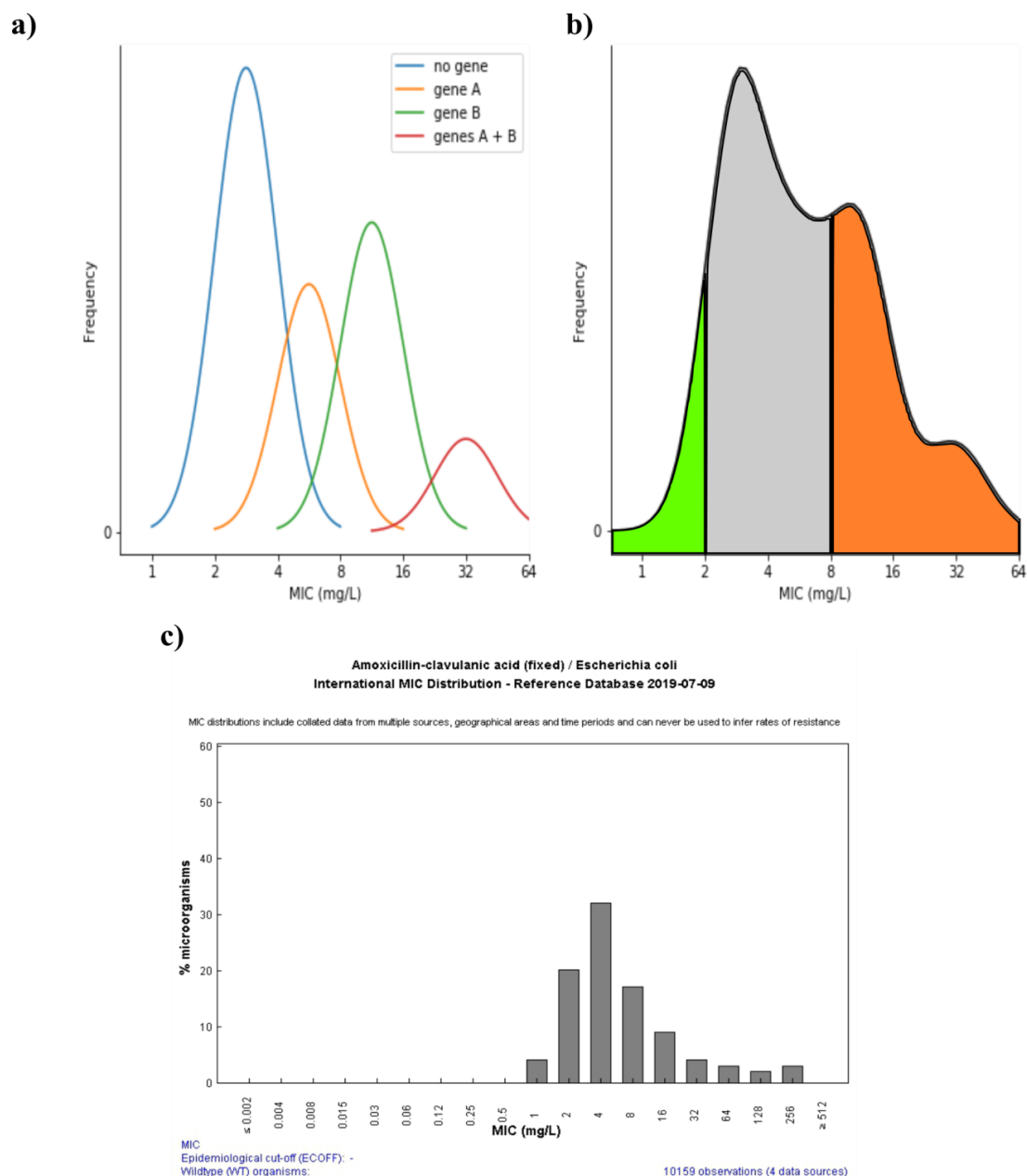


Figure 79: MICs in multi-gene resistance

Note: a) and b) theoretical explanation of how multi-gene systems lead to unimodal MIC distributions. a) shows the effect of 2 hypothetical genes, A and B, and their combination on MIC, with each gene having a relatively small effect over baseline (blue line no genes, orange line gene A present only, green line gene B present only), and then combining to produce a simplistic additive effect (red line showing both genes present). B) then shows their combined MIC distribution with no clear separation point to divide susceptible from resistant isolates; isolates in the green section definitively have no mechanism, and isolates in the orange section definitively have at least 1 if not both mechanisms, but isolates in the grey section may be either wild or non-wild-type. C) showing the MIC distribution of amoxicillin-clavulanate, reproduced from ²³¹. As demonstrated in this thesis, the resistance combines several mechanisms, some of which have a quantitative effect depending on the amount of them present in the isolate.

Any selected breakpoints then bisect this distribution, frequently with many isolates close to the breakpoint on either side. This both reduces the clear meaning of the breakpoint, and makes it more prone to classification errors resulting from expected phenotypic test errors (i.e. +/-1 doubling dilution test variation). The unclear meaning of the breakpoint and the greater risk of misclassification then likely act to weaken any hypothetical association between AST testing and clinical outcome.

The multifactorial nature of amoxicillin-clavulanate resistance not only prevents application of simplifying assumptions, but, as was seen in this thesis, causes significant issues for determining both its genotype and phenotype. Hence rather than there being problems for WGS-based resistance prediction for amoxicillin-clavulanate, this is a complex type of resistance which simplistic traditional definitions of genotype and phenotype fail to cover. Realising this may help to explain some of the difficulties describing its epidemiology ²⁴⁹ and the poor apparent link between resistance phenotypes and outcome ²⁵⁰. Further, extrapolating to other BL/BLI combinations which likely have similar issues ²⁵¹ may help explain the poor association between AST results and clinical outcome seen in clinical trials of other BL/BLI antibiotics ²⁵².

8.4. Implications

The findings in this thesis have important consequences for using amoxicillin-clavulanate in clinical practice and the reliability of its AST testing. However, outside of the use of amoxicillin-clavulanate, direct implications of these findings are less clear.

8.4.1. Importance of findings for clinical practice

The clearest implication from these findings is for clinical practice. Amoxicillin-clavulanate is the most widely used antibiotic in Europe ²⁰¹, and BL/BLIs are the second (to broad spectrum penicillins, e.g. ampicillin/amoxicillin) most widely used antibiotics worldwide ³³, yet both the definition of resistance and how to test for it are still under active debate. With many antibiotics in current development for the treatment of MDR Gram-negative infections having similar mechanisms ⁸², these uncertainties are highly concerning. In addition to the variability of individual test results, the somewhat arbitrary classifications of amoxicillin-clavulanate resistance and the difficulty defining resistance mechanisms mean the epidemiology, and impact, of resistance is poorly understood. Disputes between CLSI's and EUCAST's AST methods predominate discussion, greatly obscuring when truly rising resistance occurs ^{87,250}. One positive note however is the confirmation that MIC measurements appear more reliable than the binary phenotype. If MICs are useful for predicting clinical outcomes, binary resistance classifications could be abandoned at least in the clinical setting in favour of clinicians using MICs to direct treatment.

8.4.2. Importance of findings for different bacteria, antimicrobials and settings

To date only limited evidence, mostly coming from case reports, suggests issues similar to those of amoxicillin-clavulanate could extend to resistance to other antimicrobials ^{96,199}. Correspondingly, it is unclear whether this type of multifactorial resistance is antibiotic class, bacterial species and/or geographically specific. One plausible possibility is that it is primarily found in BL/BLI antibiotics in Gram-negative organisms, because of the non-linear reaction kinetics which result from using both active antibiotic and inhibitor concomitantly, and the limited permeability of Gram-negatives. Consistent with this, phenotyping of these

antibiotics has long been known to pose challenges ^{107,235} and genetically, there have been several reports detailing multi-mechanism resistance to BL/BLIs ^{216,251}, including for novel antibiotics still in development ²⁴³. If this view is correct, the widespread current use and expected increase in development of these antibiotics, particularly for multi-drug resistant Gram-negatives, could suggest significant problems for traditional phenotyping and resistance surveillance in future ^{82,86,201}. An alternative possibility is that this type of multifactorial resistance is likely to be a more important feature in specific Gram-negative organisms, given their more tightly controlled permeability and variably expressed chromosomal beta-lactamases. Organisms such as *P. aeruginosa*, *A. baumannii* and *S. maltophilia* have innate, complex and multifactorial beta-lactam resistance mechanisms that pre-date the use of antibiotics. Studies of these organisms have suggested that both gene presence/absence and gene expression data are important for predicting phenotype from WGS data ¹²⁴. One possibility that should not be neglected, however, is that this multifactorial resistance is simply due to the complex nature of beta-lactam resistance and treatment. The sheer number of different beta-lactam antibiotics and beta-lactamases could produce pairings where the beta-lactamase only provides weak resistance to the antibiotic, which would predispose the selection of other concomitant resistance mechanisms such as reduced permeability. While all of these are plausible hypotheses, the current limited understanding of beta-lactam resistance evolution and epidemiology prevents much meaningful extrapolation.

8.4.3. Importance of findings for applicability of WGS-based resistance prediction

Work in this thesis focussed primarily on using rule(database)-based approaches to predict antimicrobial resistance. However, to generate resistance rules, the effect of a given

mechanism needs to be both detectable and predictable. In amoxicillin-clavulanate, effects of mechanisms on MIC were often small and may have less predictable interactions with one another (e.g. permeability and beta-lactamase presence, combinations of more than one beta-lactamase) ²⁴², making rules difficult to define. As explained in the introduction, machine learning approaches, and to a lesser extent statistical models, may be better at detecting these effects. Hence, this thesis cannot answer whether WGS-based resistance prediction as a whole may be feasible in the near future, only that rule-based approaches alone would be insufficient. While still in their early stages of development, machine learning based approaches have generally been highly successful in predicting antimicrobial susceptibility, even for amoxicillin-clavulanate ^{40,125}. Although many of these results still fail to meet FDA/ISO standards, inconsistencies with established phenotypes are possibly now the greater source of error ²³³. Further, there is still significant scope for these methods to improve by either increasing the size and representativeness of model training data ¹⁴³, or by improving feature selection ²⁵³. Over time, the simple increase in availability of sequenced samples with linked phenotype data and the improved computational power enabling more complex algorithms with more input features (e.g. using reads themselves as opposed to assembly) alone will cause these methods to improve. Of note however, in some cases there will be a maximum attainable accuracy which may not be sufficient for regulatory approval given the inherent limitations of traditional AST. For example, amoxicillin-clavulanate is a case where even traditional AST methods fail to meet current acceptable standards on multiple metrics when tested using routine clinical isolates ²³³. These flaws will need to be more widely recognised before a more honest assessment of WGS's performance can take place, and may prove a bottleneck to WGS's implementation in routine clinical practice.

8.4.4. Importance of findings for resistance databases

Given the potential benefits from using machine-learning based models to improve resistance prediction, it is tempting to think the task of collecting and curating resistance databases will become obsolete. However, this is unlikely given the key role they play outside of resistance prediction, being essential to describe the epidemiology of resistance in Gram-negatives. Unlike other bacteria such as *M. tuberculosis* whose resistance is primarily driven by chromosomal mutations and can be analysed by assessing the epidemiology of strains, resistance mechanisms in Gram-negatives pass via horizontal gene transfer between both different strains and species⁴¹. To attempt to analyse this, both resistance mechanisms and their surroundings need to be identified, a task suited to using databases. Chapters 4 and 5 however highlight potential issues with using databases to identify resistance mechanisms. As more resistance mechanisms become known, standard approaches such as clustering databases and ranking matching hits according to identity may become unfeasible. The needs of nomenclature and curation of novel mechanisms will need to be balanced for resistance databases to remain usable long-term

8.4.5. Importance of findings for beta-lactam resistance epidemiology

Traditionally, beta-lactam resistance epidemiology has focussed primarily on the spread of beta-lactamase mechanisms. These are thought to both be the most important mechanism of resistance dissemination⁵⁷, and enzymes which confer high level resistance to broad-spectrum antibiotics are considered the most dangerous⁷². In this thesis however,

resistance was caused not by single enzymes causing high level resistance, but instead by multiple mechanisms individually conferring lower level resistance. The epidemiology of these effects, such as co-occurrence of beta-lactamases and high level expression of beta-lactamases, are less well understood. There is a risk, therefore, that important rises in resistance are being missed. Equally however, these mechanisms/groups of mechanisms may not be readily transferred between organisms and hence have little impact outside of locations/species where prevalent.

8.5. Strengths and limitations

8.5.1 Strengths

A major strength of this work is the samples available for study. Many previous studies of WGS-based resistance prediction, consistent with many previous studies of resistance per se, have used small sets of microbiologically “interesting” isolates^{120,141,158}. Before widespread next-generation sequencing, the high cost and technical difficulty of molecular methods to identify resistance mechanisms meant it was only applied on these kind of isolates, which had highly resistant antibiograms or where specific mechanisms were suspected. While this approach was highly informative about specific resistance mechanisms, it introduced significant bias, making findings poorly generalizable¹⁴³. For WGS-based resistance prediction, this took the form of predicting resistance in “more resistant” than average isolates, which in many cases can make the task easier. For example, predicting beta-lactam resistance in isolates with *bla*_{NDM} is easier than predicting resistance in general clinical isolates. In contrast, in this thesis, isolates were obtained sequentially and non-selectively from patients with *E. coli* BSI at a large NHS trust, using only 90-day deduplication within patient to avoid sampling effectively the same isolate repeatedly. Given the objective, to assess whether WGS-based resistance prediction is comparable to routine laboratory AST, this dataset is both more representative of the actual challenge and more generalisable, at least to settings with similar prevalence of specific resistance mechanisms.

The second major strength was the simplicity of applying database-based approaches. While machine learning may, in the long term, be a better approach to WGS-based resistance

prediction, results are based on a “black box” and thus far struggle to reach FDA/ISO standards. There are likely several reasons for these persistent errors, including poor selection of features and issues with established phenotypes. However, by their nature, machine learning approaches solely aim for prediction without using understanding of the underlying biology. Likewise, this lack of understanding is detrimental to understanding the cause of errors. As commented on above however, one aspect of translating WGS-based resistance prediction into practice will be in convincing the clinical microbiology community of any inherent weaknesses of traditional AST. By using a simpler database approach and knowledge of the relative antibiogram (specifically that ampicillin resistance can act as a sensitive marker for any beta-lactam resistance in Gram-negatives), this work found that the current biological catalogue is sufficient for identifying the beta-lactamase resistance mechanisms commonly seen in clinical practice. Then by using a simpler linear model, the effects of specific mechanisms on MIC could be visualised, which enables easy contextualisation of these results with established beta-lactamase literature. Finally, they enabled inference that multifactorial resistance leads to unavoidable AST errors. Going forward, it suggests that suboptimal specificity/sensitivity of WGS-based resistance prediction versus AST results should not merely been seen as problems with the WGS-based approach or individual phenotypic tests, but instead may point to unappreciated phenotypic complexity.

8.5.2. Limitations

This thesis assessed a complex phenotype using a combination of WGS, statistical models and phenotyping on a large dataset. While ultimately successful at understanding the causes of phenotypic uncertainty for a widely used antibiotic, I did not use more typical and

definitive laboratory techniques for understanding some of the more complex components of beta-lactam resistance. For example, signals of increased *ampC* expression were identified by comparing phenotype with mutations in the upstream region of the *ampC* gene (including looking for previously known promoter variants which lead to increased *ampC* expression). While the traditionally known mutations accounted for much of the observed resistance, there were 5 isolates which contained novel *ampC* promoter mutations (with each mutation only seen once) and were amoxicillin-clavulanate resistant without any other apparent cause. Consequently, as each mutation was only seen in a single isolate, I felt there was not enough evidence to call these isolates AmpC hyperproducers in either the resistance prediction or statistical model. A definitive traditional approach would have been to clone these mechanisms into wild-type *E. coli* and to test their effects using a combination of MICs and reverse-transcriptase PCR. Examining the effects of these mutations may prove beneficial if attempting further rule-based resistance prediction. Similarly, I did not investigate whether hypothesised markers of porin loss in WGS data had any association with isolate permeability to antibiotics. There are some case reports where truncation of porin genes has been detected from WGS data, and then porin loss confirmed using SDS-PAGE⁹⁶; however to date evidence is limited. Identifying if these effects can be reliably detected using WGS may be helpful in estimating feasibility of WGS-based resistance prediction across other Gram-negative organisms.

Another major limitation of this thesis is that whilst I began with the general aim of understanding the feasibility of WGS-based resistance prediction in Gram-negatives, in order to understand errors, I eventually focussed on two antibiotics in *E. coli* only, ampicillin and amoxicillin-clavulanate. With the good performance of standard WGS-based prediction

for many antibiotics, this attempt to understand discrepancies in detail may have been the most productive approach to identifying the feasibility of WGS-based resistance prediction. However, how useful and widely applicable (to other bacteria/antibiotics) my findings are is unknown. Machine learning methods may hold more promise than the statistical methods I used for testing the applicability of WGS-based resistance prediction across species and antibiotics given that machine learning workflows are simple to apply, reproducible and require no intricate knowledge of the mechanisms of resistance (although their caveats are listed above). More in depth and biologically orientated analysis such as the work carried out in this thesis could then be targeted specifically at understanding errors in problematic organisms/antibiotics.

Similar to other developments in the field of microbial genomics, keeping abreast of updates to methods and resources is challenging. Databases in particular have advanced rapidly over the last four years, with improved catalogues, curation and representation ¹³³.

Initiative such as CARD's "collective AMR curation effort"

(https://github.com/arpcard/amr_curation) and NCBI's "National Database of Antibiotic

Resistant Organisms" (<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/>)

are rapidly improving knowledge and curation for AMR genetics. While this work used the latest available databases at the time of writing, these do not match the quality of these modern efforts. Good database curation is vital for resistance genotyping (Chapter 5) and hence there is a risk of results changing. To counter this, I aimed to identify specifically the underlying causes of errors (e.g. by examining non-exact matching elements) and, where appropriate, correct them (e.g. for any unknown sequence checking if its protein sequence is contained in any of Resfinder, CARD, or NCBI's non-redundant nucleotide database).

Consequently, these changes are unlikely to impact inferences; however, many problems highlighted with specific databases (e.g. CARD database curation) may no longer be as significant in current releases. I still believe these problems are important to highlight however as, given the transient nature of many databases, they may be liable to cause errors again in future.

8.6. Suggested future work

8.6.1. Moving WGS into clinical practice

In spite of the challenges above, WGS still holds significant potential for predicting AMR in clinical settings. Post its initial introduction, there have been significant advances in both the prediction methods and the awareness of their pitfalls. When combined with the rapid advances in sequencing technology, particularly long-read and direct from sample sequencing, the potential benefits from over-coming problems encountered with WGS-based resistance prediction are ever increasing. However, prior to implementation of WGS, several further facts need to be established.

Robustness of WGS resistance prediction

Thus far, the feasibility of WGS-based resistance prediction has only been tested rigorously for a small number of pathogens and antibiotics. Further studies demonstrating the accuracy of WGS-based resistance prediction on large less selected sets of clinical isolates from a diverse range of pathogens will need to be carried out. Thankfully, large, unbiased sets of clinical isolates with both phenotype and genotype data are increasingly becoming available ¹⁴³, and the “big data” techniques to analyse these sets are becoming increasingly

familiar to the microbial genomics community, making these studies potentially feasible in the near future.

Importance for clinical outcome

The general route by which WGS-predicted resistance could be used to predict clinical outcome is, as discussed above, by assuming that traditional AST predicts outcome and then proving equivalence of WGS-predicted resistance with traditional AST. As seen in this thesis, some of the classical AST methods and classifications make poor gold standards and present problems for validating WGS-based prediction. Further, lack of reliability of resistance classification can call their associations with outcome into question. Importantly, if the phenotype is a poor predictor of patient outcome, aligning WGS-based predictions to it becomes pointless. An alternative approach would be to instead directly model associations between genetic determinants and outcome, conditional on antibiotic treatment received. This is attractive as the additional information given by WGS could be used to identify multiple other features outside of resistance prediction associated with clinical outcome (e.g. bacterial virulence factors), which could eventually make it a better predictor of outcome than resistance alone. However there are several issues hampering this approach. Complexity of antibiotic treatment-related confounding makes detection of genuine association difficult. Further, the assumption that AST-directed treatment improves outcome means interventional study designs comparing AST versus WGS-directed treatment have serious ethical issues preventing their use. These issues together may make determining the association between genetics and outcome extremely challenging.

How it will perform in the presence of novel resistance mechanisms

All WGS-based prediction methods (including machine learning based methods) thus far are totally reliant on whether genetic features have previously been associated with resistance. By contrast, while traditional AST cannot itself identify specific genetic mechanisms, it detects the resistance they produce, even in the case where the mechanism itself unknown. This is important as it can highlight isolates with “exceptional” phenotypes¹⁶³, which can be referred onto reference laboratories for in depth study as to the cause of the phenotype (e.g. PHE’s AMRHA1 reference unit).

Therefore, the impact of new, previously unknown, resistance variants is difficult to estimate or will even be completely missed by WGS-based resistance prediction. Current efforts to solve this problem use simulation and increased mechanistic understanding of resistance. However, work on how to identify novel mechanisms and predict the resistance they cause from WGS data alone is still far from complete. Alternative approaches would be to use early warning surveillance systems for antibiotic treatment failure as an indicator for novel types of previously unrecognised resistance.

8.6.2. Extending WGS’s utility

Direct from sample testing

Work thus far has focussed on predicting resistance of isolates obtained via pure culture. This step adds several hours to the time taken for WGS-based resistance prediction, limiting its advantages over traditional culture. DNA can be obtained directly from patient samples, but this can add significant complications. Often infections can result from a mix of bacterial species, and identifying relevant resistance can be difficult. This is particularly true for

Gram-negatives, where resistance mechanisms can be endemic in contaminating (non-infecting) organisms but also cause resistance if present in the infecting organism.

Additional features not easily detected metagenomically, such gene location ²⁴³ and gene dosage, also have effects on resistance. Short-read sequencing alone may be insufficient due to the difficulty attributing sequence to host species. With little information on the feasibility of direct from sample resistance prediction currently, future work would need to assess this. Of note however this approach is likely to require long read sequencing.

Long-read sequencing

Since the start of this work, the advent of long-read sequencing has opened up alternative approaches to predicting AMR from WGS. While not used in this thesis, the extra information it provides may be vital to accurate AMR prediction from WGS data.

Complexities such as repeat sequences, transposons and plasmids make regions of DNA difficult to resolve using short-read sequencing. These regions need longer reads which span the length of repeats to uniquely identify surrounding context. Further, because of having much longer sequences per read, species-specific reads are much easier to identify from metagenomic data. Technologies which produce longer reads have been available for some time ²⁵⁴, but the high cost and significant computational resource required to analyse their data have limited their use. However, recent advances both sequencing and data analysis ^{255,256}, as well as a significant drops in price, have enabled much more widespread adoption of this technology. As of July 2019, there are two methods in widespread use, Oxford Nanopore Technology's (ONT) nanopore based sequencing, and Pacific Biosciences (PacBio) single molecule real time sequencing. The major trade off for the long reads produced by both of these approaches, however, is their error rates, which are typically orders of

magnitude greater than Illumina short-read sequencing ²⁵⁷. However, combining short and long-read sequencing, or sequencing to increased depth can mitigate the issues caused by the long-read error rate ²⁵⁸, allowing near perfect genetic assemblies to be produced, albeit at the cost of time taken to obtain results.

Although only recently applied to bacterial resistance, these methods already represent significant advances ^{41,259,260}. In particular, their ability to resolve complete mobile genetic elements has begun to enable visualization and far greater understanding of the complex way in which resistance is acquired in Gram-negative bacteria ⁴¹. Further, the small size and portability of ONT's MinION platform have the potential to make point of care testing a reality. To my knowledge, there has been no work to date specifically examining the performance of antimicrobial resistance prediction from long read or hybrid sequencing data. Promisingly, early studies have identified some scenarios where long read sequencing enables elucidation of resistance features, such as repeated insertion of resistance genes in the bacterial chromosome, which are far more hidden in short read data ²⁵¹. However, already it is apparent that not all plasmids are reconstructable even when using long reads ²⁶¹.

Alternative sequencing applications

While not the focus of this thesis, it is important to note that in addition to whole genome sequencing, both short and long-read sequencing also have several other applications such as sequencing an organisms transcriptome or exome. Thus far these approaches have not been extensively applied in bacteria for several reasons. Firstly, their benefit was initially unclear as bacteria were believed to regulate expression primarily through the binding affinity of specific σ -factors, and consequently expression could be relatively easily derived

from genomics alone ^{262,263}. More recently however, this view has been debunked with many other methods of regulation now known ^{263,264}. Secondly, and partially as a consequence of their more limited study, obtaining and interpreting transcriptomics data is more expensive and complex. These methods require specialised library preparation and, as transcripts are produced in variable quantities, careful thought on sequencing depth ²⁶⁵. Further, with less overall study, a wide array of resources such as well curated resistance databases are less readily available ¹³⁰. Importantly however, as was seen in this thesis and is particularly the case for more complex bacteria, expression plays a key role ^{124,266,267}. Consequently, sequencing based resistance prediction will likely need to assess the bacterial transcriptome in more detail, and hence these approaches need further work to make them more widely and easily used.

8.6.3. Beta-lactam resistance epidemiology

To date, the primary focus of beta-lactam resistance epidemiology has been investigating the spread of specific enzymes present on mobile elements thought of as highly dangerous. Mechanisms outside of this have generally been felt to be less important, and only occur in sporadic cases. Work in this thesis however suggests a large degree of amoxicillin-clavulanate resistance is caused by the over-expression of commonly encountered non-inhibitor resistant *bla*_{TEM} genes. This mechanism does not conform to the typical presence/absence paradigm and cannot be monitored by simply tracking the presence/absence of relevant genes. Given the drastic rise in amoxicillin-clavulanate resistance observed over the last two decades ²⁵⁰, these non-enzyme presence/absence mechanisms may be more important than previously thought. Currently very little is known

about whether these mechanisms are increasing either over time or geographically. Further, as these mechanisms have the potential to affect some of the upcoming antibiotics designed for treating MDR Gram-negative infections ²⁴³, work needs to be done to understand if they prove a significant threat in the near future.

References

- 1 Walsh C, Wright G. Introduction: Antibiotic Resistance. *Chem Rev* 2005; **105**: 391–394.
- 2 Abraham EP, Chain E. An Enzyme from Bacteria able to Destroy Penicillin. *Nature* 1940; **146**: 837–837.
- 3 Barlow M, Hall BG. Phylogenetic Analysis Shows That the OXA b-Lactamase Genes Have Been on Plasmids for Millions of Years. *J Mol Evol* 2002; **55**: 314–321.
- 4 Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 2010; **74**: 417–33.
- 5 Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat Rev Microbiol* 2010; **8**: 260–271.
- 6 Schwartz KL, Morris SK. Travel and the Spread of Drug-Resistant Bacteria. *Curr Infect Dis Rep* 2018; **20**: 29.
- 7 Jackson N, Czaplewski L, Piddock LJ V. Discovery and development of new antibacterial drugs: learning from experience? *J Antimicrob Chemother* 2018; **73**: 1452–1459.
- 8 Reardon S. WHO warns against ‘post-antibiotic’ era. *Nature*. 2014.<http://www.nature.com/articles/nature.2014.15135> (accessed 19 Jun2019).
- 9 UK Government. Tackling antimicrobial resistance 2019-2024. 2019.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784894/UK_AMR_5_year_national_action_plan.pdf (accessed 14 Sep2019).

- 10 World Health Organisation. Global action plan on AMR.
2016.<https://www.who.int/antimicrobial-resistance/global-action-plan/en/> (accessed 25 Jun2019).
- 11 Public Health England. The UK's five-year national action plan.
2019.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784894/UK_AMR_5_year_national_action_plan.pdf (accessed 25 Jun2019).
- 12 US Centre for Communicable Disease Control. Facts about Antibiotic Resistance.
2016.<https://www.cdc.gov/antibiotic-use/community/about/fast-facts.html>
(accessed 26 Jun2019).
- 13 Shallcross LJ. Editorials: Antibiotic overuse: A key driver of antimicrobial resistance.
Br. J. Gen. Pract. 2014; **64**: 604–605.
- 14 Hulscher MEJL, Prins JM. Antibiotic stewardship: does it work in hospital practice? A review of the evidence base. *Clin Microbiol Infect* 2017; **23**: 799–805.
- 15 Mehrotra A, Linder JA. Tipping the Balance Toward Fewer Antibiotics. *JAMA Intern Med* 2016; **176**: 1649.
- 16 Caliendo AM, Gilbert DN, Ginocchio CC, Hanson KE, May L, Quinn TC *et al*. Better tests, better care: improved diagnostics for infectious diseases. *Clin Infect Dis* 2013; **57 Suppl 3**: S139-70.
- 17 The Review on Antimicrobial Resistance. Rapid Diagnostics: Stopping Unnecessary Use of Antibiotics. 2017.<https://amr-review.org/sites/default/files/Paper-Rapid-Diagnostics-Stopping-Unnecessary-Prescription.pdf> (accessed 26 Jun2019).
- 18 Doern CD. The Slow March toward Rapid Phenotypic Antimicrobial Susceptibility Testing: Are We There Yet? *J Clin Microbiol* 2018; **56**: e01999-17.

- 19 Daniels R, Nutbeam T, McNamara G, Galvin C. The sepsis six and the severe sepsis resuscitation bundle: a prospective observational cohort study. *Emerg Med J* 2011; **28**: 507–12.
- 20 Montravers P, Piednoir P, Allou N. De-escalation: in drug we trust. *Crit Care Med* 2012; **40**: 1645–6.
- 21 Baggs J, Fridkin SK, Pollack LA, Srinivasan A, Jernigan JA. Estimating National Trends in Inpatient Antibiotic Use Among US Hospitals From 2006 to 2012. *JAMA Intern Med* 2016; **176**: 1639.
- 22 van Belkum A, Dunne WM. Next-Generation Antimicrobial Susceptibility Testing. *J Clin Microbiol* 2013; **51**: 2018–2024.
- 23 Peterson LR, Dalhoff A. Towards targeted prescribing: will the cure for antimicrobial resistance be specific, directed therapy through improved diagnostic testing? *J Antimicrob Chemother* 2004; **53**: 902–905.
- 24 Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ *et al.* Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics* 2018; **19**: 714.
- 25 Tamma PD, Opene BNA, Gluck A, Chambers KK, Carroll KC, Simner PJ. Comparison of 11 Phenotypic Assays for Accurate Detection of Carbapenemase-Producing Enterobacteriaceae. *J Clin Microbiol* 2017; **55**: 1046–1055.
- 26 Eyre DW, Sheppard AE, Madder H, Moir I, Moroney R, Quan TP *et al.* A *Candida auris* Outbreak and Its Control in an Intensive Care Setting. *N Engl J Med* 2018; **379**: 1322–1331.
- 27 Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B *et al.* Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and

- Mycobacterium tuberculosis*. *Nat Commun* 2015; **6**: 10063.
- 28 Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agerso Y, Lund O *et al*. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother* 2013; **68**: 771–777.
- 29 The CRYPTIC Consortium and the 100 000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N Engl J Med* 2018; **379**: 1403–1415.
- 30 Quan TP, Bawa Z, Foster D, Walker T, del Ojo Elias C, Rathod P *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**. doi:10.1128/JCM.01480-17.
- 31 US Centre for Communicable Disease Control. Antibiotic Resistance Threats in the United States. 2013. <https://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf> (accessed 27 Jun2019).
- 32 Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C *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.
- 33 World Health Organisation. WHO Report on Surveillance of Antibiotic Consumption. 2016. <https://apps.who.int/iris/bitstream/handle/10665/277359/9789241514880-eng.pdf?ua=1> (accessed 1 Jul2019).
- 34 Laupland KB, Church DL. Population-based epidemiology and microbiology of community-onset bloodstream infections. *Clin Microbiol Rev* 2014; **27**: 647–64.

- 35 Escherich T. The intestinal bacteria of the neonate and breast-fed infant. 1885. *Rev Infect Dis*; **11**: 352–6.
- 36 Vila J, Sáez-López E, Johnson JR, Römling U, Dobrindt U, Cantón R *et al*. *Escherichia coli* : an old friend with new tidings. *FEMS Microbiol Rev* 2016; **40**: 437–463.
- 37 Foxman B. The epidemiology of urinary tract infection. *Nat Rev Urol* 2010; **7**: 653–660.
- 38 Qadri F, Svennerholm A-M, Faruque ASG, Sack RB. Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. *Clin Microbiol Rev* 2005; **18**: 465–83.
- 39 National Centre for Biotechnology Information (NCBI). *Escherichia coli* (ID 167) - Genome - NCBI. [https://www.ncbi.nlm.nih.gov/genome/?term=escherichia coli](https://www.ncbi.nlm.nih.gov/genome/?term=escherichia+coli) (accessed 5 Jul2019).
- 40 Her H-L, Wu Y-W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the *Escherichia coli* strains. *Bioinformatics* 2018; **34**: i89–i95.
- 41 Sheppard AE, Stoesser N, Wilson DJ, Sebra R, Kasarskis A, Anson LW *et al*. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene *bla*_{KPC}. *Antimicrob Agents Chemother* 2016; **60**: 3767–3778.
- 42 Fleming A. On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to their Use in the Isolation of *B. influenzae*. *Br J Exp Pathol* 1929; **10**: 226.
- 43 Chain E, Florey HW, Adelaide MB, Gardner AD, Heatley NG, Jennings MA *et al*. Penicillin as a chemotherapeutic agent. 1940. *Clin Orthop Relat Res* 1993; : 3–7.
- 44 Ligon BL. Penicillin: its discovery and early development. *Semin Pediatr Infect Dis* 2004; **15**: 52–57.

- 45 Davies J. Where have All the Antibiotics Gone? *Can J Infect Dis Med Microbiol = J Can des Mal Infect la Microbiol medicale* 2006; **17**: 287–90.
- 46 Spellberg B, Taylor-Blake B. On the exoneration of Dr. William H. Stewart: debunking an urban legend. *Infect Dis Poverty* 2013; **2**: 3.
- 47 Bush K. Past and Present Perspectives on β -Lactamases. *Antimicrob Agents Chemother* 2018; **62**. doi:10.1128/AAC.01076-18.
- 48 Kong K-F, Schneper L, Mathee K. Beta-lactam antibiotics: from antibiosis to resistance and bacteriology. *APMIS* 2010; **118**: 1–36.
- 49 Park JT, Strominger JL. Mode of Action of Penicillin: Biochemical Basis for the Mechanism of Action of Penicillin and for Its Selective Toxicity. *Science (80-)* 1957; **125**: 99–101.
- 50 Izaki K, Matshhashi M, Stromingel JI. Biosynthesis of the Peptidoglycan of Bacterial Cell Walls. *J Biol Chem* 1968; **243**: 3180–3192.
- 51 Sauvage E, Kerff F, Terrak M, Ayala JA, Charlier P. The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis. *FEMS Microbiol Rev* 2008; **32**: 234–258.
- 52 Suzuki H, Nishimura Y, Hirota Y. On the process of cellular division in Escherichia coli: a series of mutants of E. coli altered in the penicillin-binding proteins. *Proc Natl Acad Sci U S A* 1978; **75**: 664–8.
- 53 Tomasz A. The Mechanism of the Irreversible Antimicrobial Effects of Penicillins: How the Beta-Lactam Antibiotics Kill and Lyse Bacteria. *Annu Rev Microbiol* 1979; **33**: 113–137.
- 54 Cho H, Uehara T, Bernhardt TG. Beta-Lactam Antibiotics Induce a Lethal Malfunctioning of the Bacterial Cell Wall Synthesis Machinery. *Cell* 2014; **159**: 1300–

- 1311.
- 55 Zeng X, Lin J. Beta-lactamase induction and cell wall metabolism in Gram-negative bacteria. *Front Microbiol* 2013; **4**. doi:10.3389/fmicb.2013.00128.
- 56 Thornsberry C. Review of in vitro activity of third-generation cephalosporins and other newer beta-lactam antibiotics against clinically important bacteria. *Am J Med* 1985; **79**: 14–20.
- 57 Bonomo RA. β -Lactamases: A Focus on Current Challenges. *Cold Spring Harb Perspect Med* 2017; **7**. doi:10.1101/cshperspect.a025239.
- 58 McGowan SJ, Bycroft BW, Salmond GP. Bacterial production of carbapenems and clavams: evolution of beta-lactam antibiotic pathways. *Trends Microbiol* 1998; **6**: 203–8.
- 59 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
- 60 Munita JM, Arias CA. Mechanisms of Antibiotic Resistance. *Microbiol Spectr* 2016; **4**. doi:10.1128/microbiolspec.VMBF-0016-2015.
- 61 Massova I, Mobashery S. Kinship and diversification of bacterial penicillin-binding proteins and beta-lactamases. *Antimicrob Agents Chemother* 1998; **42**: 1–17.
- 62 Jacoby GA. Beta-lactamase nomenclature. *Antimicrob Agents Chemother* 2006; **50**: 1123–9.
- 63 Bush K. The ABCD's of β -lactamase nomenclature. *J Infect Chemother* 2013; **19**: 549–559.
- 64 Ambler RP. The structure of beta-lactamases. *Philos Trans R Soc Lond B Biol Sci* 1980; **289**: 321–331.
- 65 Drawz SM, Bonomo RA. Three decades of beta-lactamase inhibitors. *Clin Microbiol*

- Rev 2010; **23**: 160–201.
- 66 Bush K, Jacoby GA, Medeiros AA. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob Agents Chemother* 1995; **39**: 1211–1233.
 - 67 Livermore DM. beta-Lactamases in laboratory and clinical resistance. *Clin Microbiol Rev* 1995; **8**: 557–84.
 - 68 Vlek ALM, Cooper BS, Kypraios T, Cox A, Edgeworth JD, Auguet OT. Clustering of antimicrobial resistance outbreaks across bacterial species in the intensive care unit. *Clin Infect Dis* 2013; **57**: 65–76.
 - 69 Yang Y, Livermore DM, Williams RJ. Chromosomal β -lactamase expression and antibiotic resistance in *Enterobacter cloacae*. *J Med Microbiol* 1988; **25**: 227–233.
 - 70 Jacoby GA. AmpC beta-lactamases. *Clin Microbiol Rev* 2009; **22**: 161–82, Table of Contents.
 - 71 Caroff N, Espaze E, Bérard I, Richet H, Reynaud A. Mutations in the ampC promoter of *Escherichia coli* isolates resistant to oxyiminocephalosporins without extended spectrum β -lactamase production. *FEMS Microbiol Lett* 1999; **173**: 459–465.
 - 72 Logan LK, Weinstein RA. The Epidemiology of Carbapenem-Resistant Enterobacteriaceae: The Impact and Evolution of a Global Menace. *J Infect Dis* 2017; **215**: S28–S36.
 - 73 Paterson DL, Bonomo RA. Extended-Spectrum β -Lactamases: a Clinical Update. *Clin Microbiol Rev* 2005; **18**: 657–686.
 - 74 Gupta N, Limbago BM, Patel JB, Kallen AJ. Carbapenem-Resistant Enterobacteriaceae: Epidemiology and Prevention. *Clin Infect Dis* 2011; **53**: 60–67.
 - 75 Decraene V, Phan HTT, George R, Wyllie DH, Akinremi O, Aiken Z *et al.* A Large,

- Refractory Nosocomial Outbreak of *Klebsiella pneumoniae* Carbapenemase-Producing *Escherichia coli* Demonstrates Carbapenemase Gene Outbreaks Involving Sink Sites Require Novel Approaches to Infection Control. *Antimicrob Agents Chemother* 2018; **62**. doi:10.1128/AAC.01689-18.
- 76 McNulty CAM, Lecky DM, Xu-McCrae L, Nakiboneka-Ssenabulya D, Chung K-T, Nichols T *et al*. CTX-M ESBL-producing Enterobacteriaceae: estimated prevalence in adults in England in 2014. *J Antimicrob Chemother* 2018; **73**: 1368–1388.
- 77 Public Health England. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR). 2018.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759975/ESPAUR_2018_report.pdf (accessed 4 Dec2018).
- 78 Paterson DL, Ko W-C, Gottberg A Von, Casellas JM, Mulazimoglu L, Klugman KP *et al*. Outcome of Cephalosporin Treatment for Serious Infections Due to Apparently Susceptible Organisms Producing Extended-Spectrum β -Lactamases: Implications for the Clinical Microbiology Laboratory. *J Clin Microbiol* 2001; **39**: 2206–2212.
- 79 Paterson DL, Bonomo RA. Extended-spectrum beta-lactamases: a clinical update. *Clin Microbiol Rev* 2005; **18**: 657–86.
- 80 Woerther P-L, Burdet C, Chachaty E, Andremont A. Trends in Human Fecal Carriage of Extended-Spectrum β -Lactamases in the Community: Toward the Globalization of CTX-M. *Clin Microbiol Rev* 2013; **26**: 744.
- 81 Bush K. Beta-lactamase inhibitors from laboratory to clinic. *Clin Microbiol Rev* 1988; **1**: 109–23.
- 82 Wright H, Bonomo RA, Paterson DL. New agents for the treatment of infections with Gram-negative bacteria: restoring the miracle or false dawn? *Clin Microbiol Infect*

- 2017; **23**: 704–712.
- 83 Brooke JS. *Stenotrophomonas maltophilia*: an emerging global opportunistic pathogen. *Clin Microbiol Rev* 2012; **25**: 2–41.
 - 84 Mojica MF, Papp-Wallace KM, Taracila MA, Barnes MD, Rutter JD, Jacobs MR *et al*. Avibactam Restores the Susceptibility of Clinical Isolates of *Stenotrophomonas maltophilia* to Aztreonam. *Antimicrob Agents Chemother* 2017; **61**.
doi:10.1128/AAC.00777-17.
 - 85 Zhanel GG, Golden AR, Zelenitsky S, Wiebe K, Lawrence CK, Adam HJ *et al*. Cefiderocol: A Siderophore Cephalosporin with Activity Against Carbapenem-Resistant and Multidrug-Resistant Gram-Negative Bacilli. *Drugs*. 2019; **79**: 271–289.
 - 86 Bush K. A resurgence of beta-lactamase inhibitor combinations effective against multidrug-resistant Gram-negative pathogens. *Int J Antimicrob Agents* 2015; **46**: 483–493.
 - 87 Mouton J. Antimicrobial susceptibility testing with EUCAST breakpoints and methods. In: Mouton J (ed). *ECCMID*. Vienna, Austria, 2017.
 - 88 Finlay J, Miller L, Poupard JA. A review of the antimicrobial activity of clavulanate. *J Antimicrob Chemother* 2003; **52**: 18–23.
 - 89 Singh SB, Young K, Silver LL. What is an “ideal” antibiotic? Discovery challenges and path forward. *Biochem Pharmacol* 2017; **133**: 63–73.
 - 90 Livermore DM. beta-Lactamases: quantity and resistance. *Clin Microbiol Infect* 1997; **3 Suppl 4**: S10–S19.
 - 91 Lister PD, Wolter DJ, Hanson ND. Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clin Microbiol Rev* 2009; **22**: 582–610.

- 92 Nikaido H. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* 2003; **67**: 593–656.
- 93 Humphries RM, Hemarajata P, Nelson K, Hemarajata P, Sun D, Rubio-Aparicio D *et al*. Resistance to Ceftazidime-Avibactam in *Klebsiella pneumoniae* Due to Porin Mutations and the Increased Expression of KPC-3. *Antimicrob Agents Chemother* 2017; **61**: e00537-17.
- 94 Li X-Z, Nikaido H. Efflux-Mediated Drug Resistance in Bacteria. *Drugs* 2004; **64**: 159–204.
- 95 Nikaido H, Normark S. Sensitivity of *Escherichia coli* to various beta-lactams is determined by the interplay of outer membrane permeability and degradation by periplasmic beta-lactamases: a quantitative predictive treatment. *Mol Microbiol* 1987; **1**: 29–36.
- 96 Beceiro A, Maharjan S, Gaulton T, Doumith M, Soares NC, Dhanji H *et al*. False extended-spectrum β -lactamase phenotype in clinical isolates of *Escherichia coli* associated with increased expression of OXA-1 or TEM-1 penicillinases and loss of porins. *J Antimicrob Chemother* 2011; **66**: 2006–2010.
- 97 Moyá B, Beceiro A, Cabot G, Juan C, Zamorano L, Alberti S *et al*. Pan-Lactam Resistance Development in *Pseudomonas aeruginosa* Clinical Strains: Molecular Mechanisms, Penicillin-Binding Protein Profiles, and Binding Affinities. *Antimicrob Agents Chemother* 2012; **56**: 4771–8.
- 98 Sumita Y, Fakasawa M. Potent activity of meropenem against *Escherichia coli* arising from its simultaneous binding to penicillin-binding proteins 2 and 3. *J Antimicrob Chemother* 1995; **36**: 53–64.
- 99 Zapun A, Contreras-Martel C, Vernet T. Penicillin-binding proteins and β -lactam

- resistance. *FEMS Microbiol Rev* 2008; **32**: 361–385.
- 100 Zhang Y, Kashikar A, Brown CA, Denys G, Bush K. Unusual *Escherichia coli* PBP 3 Insertion Sequence Identified from a Collection of Carbapenem-Resistant Enterobacteriaceae Tested In Vitro with a Combination of Ceftazidime-, Ceftaroline-, or Aztreonam-Avibactam. *Antimicrob Agents Chemother* 2017; **61**: e00389-17.
 - 101 Lorian V, Pillai S, Moellering R, Eliopoulos G. *Antibiotics in Laboratory Medicine*. Wolters Kluwer Health: Philadelphia, United States, 2015.
 - 102 Wheat PF. History and development of antimicrobial susceptibility testing methodology. *J Antimicrob Chemother* 2001; **48**: 1–4.
 - 103 Fleming A. A Comparison of the Activities of Antiseptics on Bacteria and on Leucocytes. *Proc R Soc B Biol Sci* 1924; **96**: 171–180.
 - 104 Jorgensen JHH, Ferraro MJ. Antimicrobial Susceptibility Testing: A Review of General Principles and Contemporary Practices. *Clin Infect Dis* 2009; **49**: 1749–55.
 - 105 U.S. Food & Drug Administration. 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 Apr2019).
 - 106 International Organisation for Standardization. ISO 20776-2:2007 - Part 2: Evaluation of performance of antimicrobial susceptibility test devices. 2007.<https://www.iso.org/standard/41631.html> (accessed 5 Apr2019).
 - 107 European Committee on Antimicrobial Susceptibility Testing. EUCAST: Warnings! 2019.http://www.eucast.org/ast_of_bacteria/warnings/ (accessed 11 Apr2019).
 - 108 Poulter MD, Hindler JF. Challenges in Antimicrobial Susceptibility Testing and Reporting. *Lab Med* 2002; **33**: 877–884.

- 109 Cunha BA. Problems arising in antimicrobial therapy due to false susceptibility testing. *J Chemother* 1997; **9 Suppl 1**: 25–35.
- 110 Jorgensen JH. Selection criteria for an antimicrobial susceptibility testing system. *J Clin Microbiol* 1993; **31**: 2841–4.
- 111 Turnidge J, Paterson DL. Setting and Revising Antibacterial Susceptibility Breakpoints. *Clin Microbiol Rev* 2007; **20**: 391–408.
- 112 van der Bij AK, van Dijk K, Muilwijk J, Thijsen SFT, Notermans DW, de Greeff S *et al*. Clinical breakpoint changes and their impact on surveillance of antimicrobial resistance in *Escherichia coli* causing bacteraemia. *Clin Microbiol Infect* 2012; **18**: E466–E472.
- 113 U.S. Food & Drug Administration. Antibacterial Susceptibility Test Interpretive Criteria. 2019.<https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm575163.htm> (accessed 18 Apr2019).
- 114 Menozzi MG, Eigner U, Covan S, Rossi S, Somenzi P, Dettori G *et al*. Two-center collaborative evaluation of performance of the BD phoenix automated microbiology system for identification and antimicrobial susceptibility testing of gram-negative bacteria. *J Clin Microbiol* 2006; **44**: 4085–94.
- 115 Clinical Laboratory Standards Institute. CLSI supplement M100. In: *Performance Standards for Antimicrobial Susceptibility Testing, 28th ed*. Wayne, PA, 2018.
- 116 European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. 2019.http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_9.0_Breakpoint_Tables.pdf (accessed 2 Oct2019).

- 117 Pankhurst LJ, del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J *et al.* Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. *Lancet Respir Med* 2016; **4**: 49–58.
- 118 Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V *et al.* Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res* 2017. doi:10.1101/gr.216606.116.
- 119 Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.* 2019; **57**. doi:10.1128/JCM.01405-18.
- 120 Tyson GH, McDermott PF, Li C, Chen Y, Tadesse DA, Mukherjee S *et al.* WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *J Antimicrob Chemother* 2015; **70**: 2763–2769.
- 121 Walker TM, Kohl TA, Omar S V, Hedge J, Del Ojo Elias C, Bradley P *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect Dis* 2015; **15**: 1193–1202.
- 122 Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J *et al.* Prediction of *Staphylococcus aureus* Antimicrobial Resistance by Whole-Genome Sequencing. *J Clin Microbiol* 2014; **52**: 1182–1191.
- 123 Eyre DW, Silva D De, Cole K, Peters J, Cole MJ, Grad YH *et al.* WGS to predict antibiotic MICs for *Neisseria gonorrhoeae*. *J Antimicrob Chemother* 2017; **72**. doi:10.1093/jac/dkx067.
- 124 Khaledi A, Weimann A, Schniederjans M, Asgari E, Kuo T-H, Oliver A *et al.* Fighting antimicrobial resistance in *Pseudomonas aeruginosa* with machine learning-enabled molecular diagnostics. *bioRxiv* 2019; : 643676.

- 125 Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL *et al.* Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal Salmonella. *J Clin Microbiol* 2019; **57**: e01260-18.
- 126 Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R *et al.* Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci Rep* 2018; **8**: 421.
- 127 Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ *et al.* Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. *Bioinformatics* 2018; **34**: 1666–1671.
- 128 McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. *Ann N Y Acad Sci* 2017; **1388**: 78–91.
- 129 Crofts TS, Gasparrini AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. *Nat. Rev. Microbiol.* 2017; **15**.
doi:10.1038/nrmicro.2017.28.
- 130 Boolchandani M, D’Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet* 2019 2019; : 1.
- 131 Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E *et al.* MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res* 2017; **45**: D574–D580.
- 132 Bush K, Palzkill T. β -Lactamase Classification and Amino Acid Sequences for TEM, SHV and OXA Extended-Spectrum and Inhibitor Resistant Enzymes.
<https://www.lahey.org/Studies/> (accessed 17 Jan2019).
- 133 Jia B, Raphenya AR, Alcock B, Wagglechner N, Guo P, Tsang KK *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance

- database. *Nucleic Acids Res* 2017; **45**: D566–D573.
- 134 National Centre for Biotechnology Information. AMRFinder.
<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>
(accessed 17 Jan2019).
- 135 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.
- 136 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–9.
- 137 Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T *et al*. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014; **6**: 90.
- 138 Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb genomics* 2017; **3**: e000131.
- 139 Clausen PTLC, 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.
- 140 Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC *et al*. Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and Virulence Factors from *Staphylococcus aureus* Whole-Genome Sequences. *J Clin Microbiol* 2018; **56**. doi:10.1128/JCM.01815-17.
- 141 Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C *et al*. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *J Antimicrob Chemother* 2013; **68**: 2234–2244.

- 142 Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham CAD *et al.*
Evaluation of machine learning and rules-based approaches for predicting
antimicrobial resistance profiles in gram-negative bacilli from whole genome
sequence data. *Front Microbiol* 2016; **7**. doi:10.3389/fmicb.2016.01887.
- 143 Monk JM. Predicting Antimicrobial Resistance and Associated Genomic Features from
Whole-Genome Sequencing. *J Clin Microbiol* 2019; **57**: e01610-18.
- 144 Siedner MJ, Galar A, Guzmán-Suarez BB, Kubiak DW, Baghdady N, Ferraro MJ *et al.*
Cefepime vs other antibacterial agents for the treatment of Enterobacter species
bacteremia. *Clin Infect Dis* 2014; **58**: 1554–63.
- 145 Bush K, Jacoby GA. Updated functional classification of beta-lactamases. *Antimicrob
Agents Chemother* 2010; **54**: 969–976.
- 146 Jacoby GA, Bonomo RA, Bradford PA, Bush K, Doi Y, Feldgarden M *et al.* Comment on:
Resistance gene naming and numbering: is it a new gene or not? *J Antimicrob
Chemother* 2016; **71**: 2677–2678.
- 147 Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of
reconstructing plasmids from whole-genome short-read sequencing data. *Microb
genomics* 2017; **3**: e000128.
- 148 Navon-Venezia S, Chmelnitsky I, Leavitt A, Carmeli Y. Dissemination of the CTX-M-25
family -lactamases among Klebsiella pneumoniae, Escherichia coli and Enterobacter
cloacae and identification of the novel enzyme CTX-M-41 in Proteus mirabilis in Israel.
J Antimicrob Chemother 2008; **62**: 289–295.
- 149 McGowan JE. Resistance in Nonfermenting Gram-Negative Bacteria: Multidrug
Resistance to the Maximum. *Am J Med* 2006; **119**: S29–S36.
- 150 Su M, Satola SW, Read TD. Genome-Based Prediction of Bacterial Antibiotic

- Resistance. *J Clin Microbiol* 2018; **57**. doi:10.1128/JCM.01405-18.
- 151 Coll F, McNerney R, Preston MD, Guerra-Assunção JA, Warry A, Hill-Cawthorne G *et al*. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med* 2015; **7**. doi:10.1186/s13073-015-0164-0.
 - 152 Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C, Olson R *et al*. Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci Rep* 2016; **6**. doi:10.1038/srep27930.
 - 153 Stoesser N, Sheppard AE, Peirano G, Anson LW, Pankhurst L, Sebra R *et al*. Genomic epidemiology of global *Klebsiella pneumoniae* carbapenemase (KPC)-producing *Escherichia coli*. *Sci Rep* 2017; **7**: 5917.
 - 154 Neuert S, Nair S, Day MR, Doumith M, Ashton PM, Mellor KC *et al*. Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*. *Front Microbiol* 2018; **9**: 592.
 - 155 Day MR, Doumith M, Do Nascimento V, Nair S, Ashton PM, Jenkins C *et al*. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Salmonella enterica* serovars Typhi and Paratyphi. *J Antimicrob Chemother* 2018; **73**: 365–372.
 - 156 Moradigaravand D, Palm M, Farewell A, Mustonen V, Warringer J, Parts L. Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLOS Comput Biol* 2018; **14**: e1006258.
 - 157 Pesesky MW, Hussain T, Wallace M, Patel S, Andleeb S, Burnham C-AD *et al*. Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. *Front Microbiol* 2016; **7**: 1887.
 - 158 Shelburne SA, Kim J, Munita JM, Sahasrabhojane P, Shields RK, Press EG *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.
- 159 Kim J, Greenberg DE, Pfifer R, Jiang S, Xiao G, Xie Y *et al*. VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *bioRxiv* 2019; : 537381.
- 160 Finney JM, Walker AS, Peto TEA, Wyllie DH. An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med Inform Decis Mak* 2011; **11**: 7.
- 161 Butler CC, Dunstan F, Heginbotham M, Mason B, Roberts Z, Hillier S *et al*. Containing antibiotic resistance: decreased antibiotic-resistant coliform urinary tract infections with reduction in antibiotic prescribing by general practices. *Br J Gen Pract* 2007; **57**: 785–92.
- 162 Davies TJ, Stoesser N, Sheppard AE, Abuoun M, Fowler P, Swann J *et al*. Reconciling the potentially irreconcilable? Genotypic and phenotypic amoxicillin-clavulanate resistance in *Escherichia coli*. *bioRxiv* 2019; : 511402.
- 163 Phe. Bacteriology Reference Department User Manual PHE Bacteriology Reference Department User Manual. www.facebook.com/PublicHealthEngland (accessed 2 Aug2019).
- 164 AbuOun M, Stubberfield EJ, Duggett NA, Kirchner M, Dormer L, Nunez-Garcia J *et al*. mcr-1 and mcr-2 (mcr-6.1) variant genes identified in *Moraxella* species isolated from pigs in Great Britain from 2014 to 2015. *J Antimicrob Chemother* 2017; **72**: 2745–2749.
- 165 Bushnell B. BBMap download | SourceForge.net.

- 2013.<https://sourceforge.net/projects/bbmap/> (accessed 11 Dec2019).
- 166 Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: R46.
 - 167 Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 2011; **21**: 936–9.
 - 168 Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; **18**: 821–829.
 - 169 Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinforma* 2010; **Chapter 11**: Unit 11.5.
 - 170 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
 - 171 Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; **29**: 1072–1075.
 - 172 Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006; **60**: 1136–1151.
 - 173 Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S *et al*. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016; **17**. doi:10.1186/s13059-016-0997-x.
 - 174 Koeth LM, King A, Knight H, May J, Miller LA, Phillips I *et al*. Comparison of cation-adjusted Mueller-Hinton broth with Iso-Sensitest broth for the NCCLS broth microdilution method. *J Antimicrob Chemother* 2000; **46**: 369–376.
 - 175 van den Bijllaardt W, Buiting AG, Mouton JW, Muller AE. Shortening the incubation time for antimicrobial susceptibility testing by disk diffusion for Enterobacteriaceae: how short can it be and are the results accurate? *Int J Antimicrob Agents* 2017; **49**:

- 631–637.
- 176 Andrews JM. Determination of minimum inhibitory concentrations. *J Antimicrob Chemother* 2001; **48**: 5–16.
- 177 Ison CA, Lewis DA. Gonorrhea. In: Morse SA, Ballard RC, Holmes KK, Moreland AABT (eds). *Atlas of Sexually Transmitted Diseases 4th edition*. W.B. Saunders: London, 2010, pp 24–39.
- 178 Carroll KC, Borek AP, Burger C, Glanz B, Bhally H, Henciak S *et al*. Evaluation of the BD Phoenix automated microbiology system for identification and antimicrobial susceptibility testing of staphylococci and enterococci. *J Clin Microbiol* 2006; **44**: 2072–7.
- 179 Junkins AD, Arbefeville SS, Howard WJ, Richter SS. Comparison of BD Phoenix AP Workflow with Vitek 2. *J Clin Microbiol* 2010; **48**: 1929–31.
- 180 US Food and Drug Administration. Substantial Equivalence Determination Decision Summary - K163637.
https://www.accessdata.fda.gov/cdrh_docs/reviews/K163637.pdf (accessed 7 Aug2019).
- 181 Woodford N, Eastaway AT, Ford M, Leanord A, Keane C, Quayle RM *et al*. Comparison of BD Phoenix, Vitek 2, and MicroScan automated systems for detection and inference of mechanisms responsible for carbapenem resistance in Enterobacteriaceae. *J Clin Microbiol* 2010; **48**: 2999–3002.
- 182 Mittman SA, Huard RC, Della-Latta P, Whittier S. Comparison of BD Phoenix to Vitek 2, MicroScan MICroSTREP, and Etest for antimicrobial susceptibility testing of *Streptococcus pneumoniae*. *J Clin Microbiol* 2009; **47**: 3557–3561.
- 183 Bobenchik AM, Deak E, Hindler JA, Charlton CL, Humphries RM. Performance of Vitek

- 2 for Antimicrobial Susceptibility Testing of Enterobacteriaceae with Vitek 2 (2009 FDA) and 2014 CLSI Breakpoints. 2015. doi:10.1128/JCM.02697-14.
- 184 Perez F, Granger BE. IPython: A System for Interactive Scientific Computing. *Comput Sci Eng* 2007; **9**: 21–29.
- 185 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009; **25**: 1422–1423.
- 186 McKinney W. Data Structures for Statistical Computing in Python. 2010; : 51–56.
- 187 van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput Sci Eng* 2011; **13**: 22–30.
- 188 Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.
- 189 Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* 2007; **9**: 90–95.
- 190 Ewers C, Bethe A, Semmler T, Guenther S, Wieler LH. Extended-spectrum β -lactamase-producing and AmpC-producing *Escherichia coli* from livestock and companion animals, and their putative impact on public health: A global perspective. *Clin. Microbiol. Infect.* 2012; **18**: 646–655.
- 191 Bettelheim KA, Bushrod FM, Chandler ME, Cooke EM, O’Farrell S, Shooter RA. *Escherichia coli* serotype distribution in man and animals. *J Hyg (Lond)* 1974; **73**: 467–471.
- 192 Bélanger L, Garenaux A, Harel J, Boulianne M, Nadeau E, Dozois CM. *Escherichia coli* from animal reservoirs as a potential source of human extraintestinal pathogenic *E. coli*. *FEMS Immunol Med Microbiol* 2011; **62**: 1–10.
- 193 Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *Clin Microbiol Rev* 2019; **32**: e00135-18.

- 194 van Saene R, Fairclough S, Petros A. Broad- and narrow-spectrum antibiotics: a different approach. *Clin Microbiol Infect* 1998; **4**: 56–57.
- 195 Williamson DA, Roberts SA, Smith M, Heffernan H, Tiong A, Pope C *et al*. High rates of susceptibility to ceftazidime among globally prevalent CTX-M-producing *Escherichia coli*: potential clinical implications of the revised CLSI interpretive criteria. *Eur J Clin Microbiol Infect Dis* 2012; **31**: 821–824.
- 196 Gupta G, Tak V, Mathur P. Detection of AmpC β Lactamases in Gram-negative Bacteria. *J Lab Physicians* 2014; **6**: 1–6.
- 197 Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG *et al*. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015; **31**. doi:10.1093/bioinformatics/btv421.
- 198 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–77.
- 199 Siu LK, Lu P-L, Chen J-Y, Lin FM, Chang S-C. High-level expression of ampC beta-lactamase due to insertion of nucleotides between -10 and -35 promoter sequences in *Escherichia coli* clinical isolates: cases not responsive to extended-spectrum-cephalosporin treatment. *Antimicrob Agents Chemother* 2003; **47**: 2138–44.
- 200 Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics* 2015; **31**: 3350–3352.
- 201 Plachouras D, Kärki T, Hansen S, Hopkins S, Lyytikäinen O, Moro ML *et al*. Antimicrobial use in European acute care hospitals: results from the second point prevalence survey (PPS) of healthcare-associated infections and antimicrobial use, 2016 to 2017. *Eurosurveillance* 2018; **23**: 1800393.

- 202 Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L *et al.* ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob Agents Chemother* 2014; **58**: 212–220.
- 203 Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**: 1658–1659.
- 204 Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016; **32**: 2103–2110.
- 205 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C *et al.* Versatile and open software for comparing large genomes. *Genome Biol* 2004; **5**: R12.
- 206 Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012; **28**: 593–4.
- 207 Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. *J Open Source Softw* 2016. doi:10.21105/joss.00027.
- 208 Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970; **48**: 443–453.
- 209 Dhara L, Tripathi A, Pal A. Molecular characterization and in silico analysis of naturally occurring TEM beta-lactamase variants among pathogenic Enterobacteriaceae infecting Indian patients. *Biomed Res Int* 2013; **2013**: 783540.
- 210 Moran RA, Hall RM. pBuzz: A cryptic rolling-circle plasmid from a commensal *Escherichia coli* has two inversely oriented oriTs and is mobilised by a B/O plasmid. *Plasmid* 2019; **101**: 10–19.
- 211 Schmidt K, Mwaigwisya S, Crossman LC, Doumith M, Munroe D, Pires C *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017;

- 72:** 104–114.
- 212 Slocombe B, Beale AS, Boon RJ, Griffin KE, Masters PJ, Sutherland R *et al.* Antibacterial activity in vitro and in vivo of amoxicillin in the presence of clavulanic acid. *Postgr Med* 1984; **29**: 49.
 - 213 Adam D, de Visser I, Koeppe P. Pharmacokinetics of amoxicillin and clavulanic acid administered alone and in combination. *Antimicrob Agents Chemother* 1982; **22**: 353–357.
 - 214 Jacoby GA. AmpC -Lactamases. *Clin Microbiol Rev* 2009; **22**: 161–182.
 - 215 Harder,' Hiroshi Nikaido,' KJ, Matsushashi2 M. Mutants of Escherichia coli That Are Resistant to Certain Beta-Lactam Compounds Lack the ompF Porin. 1981<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC181743/pdf/aac00010-0131.pdf> (accessed 22 Mar2019).
 - 216 Nelson K, Hemarajata P, Sun D, Rubio-Aparicio D, Tsivkovski R, Yang S *et al.* Resistance to Ceftazidime-Avibactam Is Due to Transposition of KPC in a Porin-Deficient Strain of Klebsiella pneumoniae with Increased Efflux Activity. *Antimicrob Agents Chemother* 2017; **61**. doi:10.1128/AAC.00989-17.
 - 217 Haidar G, Clancy CJ, Chen L, Samanta P, Shields RK, Kreiswirth BN *et al.* Identifying Spectra of Activity and Therapeutic Niches for Ceftazidime-Avibactam and Imipenem-Relebactam against Carbapenem-Resistant Enterobacteriaceae. *Antimicrob Agents Chemother* 2017; **61**: e00642-17.
 - 218 Nikaido H, Rosenberg EY, Foulds J, Rosenberg,' And EY, Foulds2 J. Porin channels in Escherichia coli: studies with beta-lactams in intact cells. *J Bacteriol* 1983; **153**.
 - 219 Livermore DM. β -Lactamases: quantity and resistance. *Clin Microbiol Infect* 1997; **3**: 4S10-4S19.

- 220 French GL, Shannon KP, Simmons N. Hospital outbreak of *Klebsiella pneumoniae* resistant to broad-spectrum cephalosporins and beta-lactam-beta-lactamase inhibitor combinations by hyperproduction of SHV-5 beta-lactamase. *J Clin Microbiol* 1996; **34**: 358–63.
- 221 Livermore DM, Day M, Cleary P, Hopkins KL, Toleman MA, Wareham DW *et al.* OXA-1 β -lactamase and non-susceptibility to penicillin/ β -lactamase inhibitor combinations among ESBL-producing *Escherichia coli*. *J Antimicrob Chemother* 2018; : dky453–dky453.
- 222 Ortega A, Oteo J, Aranzamendi-Zaldumbide M, Bartolome RM, Bou G, Cercenado E *et al.* Spanish multicenter study of the epidemiology and mechanisms of amoxicillin-clavulanate resistance in *Escherichia coli*. *Antimicrob Agents Chemother* 2012; **56**: 3576–3581.
- 223 Seetulsingh PS, Hall LMC, Livermore DM. Activity of clavulanate combinations against TEM-1 β -lactamase-producing *Escherichia coli* isolates obtained in 1982 and 1989. *J Antimicrob Chemother* 1991; **27**: 749–759.
- 224 Lartigue MF, Leflon-Guibout V, Poirel L, Nordmann P, Nicolas-Chanoine M-H. Promoters P3, Pa/Pb, P4, and P5 upstream from bla(TEM) genes and their relationship to beta-lactam resistance. *Antimicrob Agents Chemother* 2002; **46**: 4035–4037.
- 225 Jaurin B, Grundström T, Edlund T, Normark S. The *E. coli* β -lactamase attenuator mediates growth rate-dependent regulation. *Nature* 1981; **290**: 221–225.
- 226 Peter-Getzlaff S, Polsfuss S, Poledica M, Hombach M, Giger J, Bottger EC *et al.* Detection of AmpC beta-lactamase in *Escherichia coli*: comparison of three phenotypic confirmation assays and genetic analysis. *J Clin Microbiol* 2011; **49**: 2924–

- 2932.
- 227 Wang X, Minasov G, Blázquez J, Caselli E, Prati F, Shoichet BK. Recognition and Resistance in TEM β -Lactamase . *Biochemistry* 2003; **42**: 8434–8444.
- 228 Imtiaz U, Billings E, Knox JR, Manavathu EK, Lerner SA, Mobashery S. Inactivation of class A β -lactamases by clavulanic acid: the role of arginine-244 in a proposed nonconcerted sequence of events. *J Am Chem Soc* 1993; **115**: 4435–4442.
- 229 Friebs K. Plasmid Copy Number and Plasmid Stability BT - New Trends and Developments in Biochemical Engineering. In: Scheper T (ed). . Springer Berlin Heidelberg: Berlin, Heidelberg, Heidelberg, 2004, pp 47–82.
- 230 de Vooght KMK, van Wijk R, van Solinge WW. Management of gene promoter mutations in molecular diagnostics. *Clin Chem* 2009; **55**: 698–708.
- 231 European Committee on Antimicrobial Susceptibility Testing. Antimicrobial wild type distributions of microorganisms. 2019.<https://mic.eucast.org/Eucast2/> (accessed 5 Apr2019).
- 232 Crass RL, Pai MP. Pharmacokinetics and Pharmacodynamics of β -Lactamase Inhibitors. *Pharmacother J Hum Pharmacol Drug Ther* 2019; **39**: 182–195.
- 233 María D-A, María-Isabel M, María-Carmen C, Álvaro P, Jorge C, Luis M-M *et al*. Establishing the validity of different susceptibility testing methods to evaluate the in vitro activity of amoxicillin-clavulanate against Escherichia coli. *Diagn Microbiol Infect Dis* 2016; **84**: 334–336.
- 234 Díez-Aguilar M, Morosini M-I, López-Cerero L, Pascual A, Calvo J, Martínez-Martínez L *et al*. Performance of EUCAST and CLSI approaches for co-amoxiclav susceptibility testing conditions for clinical categorization of a collection of Escherichia coli isolates with characterized resistance phenotypes. *J Antimicrob Chemother* 2015; **70**: 2306–

- 2310.
- 235 European Committee on Antimicrobial Susceptibility Testing. Guidance document on beta-lactamase inhibitor combinations. 2017.http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/General_documents/Inhibitor_combinations_-_Guidance_for_drug_developers_20171001.pdf (accessed 12 Apr2019).
- 236 European Committee on Antimicrobial Susceptibility Testing. Area of Technical Uncertainty (ATU) in antimicrobial susceptibility testing. 2019.http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/Area_of_Technical_Uncertainty_-_guidance_2019-1.pdf (accessed 15 Apr2019).
- 237 Leverstein-van Hall MA, Waar K, Muilwijk J, Cohen Stuart J. Consequences of switching from a fixed 2 : 1 ratio of amoxicillin/clavulanate (CLSI) to a fixed concentration of clavulanate (EUCAST) for susceptibility testing of *Escherichia coli*. *J Antimicrob Chemother* 2013; **68**: 2636–2640.
- 238 European Committee on Antimicrobial Susceptibility Testing (EUCAST). Determination of minimum inhibitory concentrations (MICs) of antibacterial agents by agar dilution. *Clin Microbiol Infect* 2000; **6**: 509–515.
- 239 Stubberfield E, AbuOun M, Sayers E, O'Connor HM, Card RM, Anjum MF. Use of whole genome sequencing of commensal *Escherichia coli* in pigs for antimicrobial resistance surveillance, United Kingdom, 2018. *Eurosurveillance* 2019; **24**. doi:10.2807/1560-7917.ES.2019.24.50.1900136.
- 240 European Committee on Antimicrobial Susceptibility Testing. Routine and extended internal quality control for MIC determination and disk diffusion as recommended by EUCAST. 2019.http://www.eucast.org/ast_of_bacteria/qc_tables/ (accessed 23

- Apr2019).
- 241 Humphries R. Mechanisms of resistance to ceftazidime- avibactam. In: *ECCMID*. Madrid, Spain, 2018.
 - 242 Livermore DM, Day M, Cleary P, Hopkins KL, Toleman MA, Wareham DW *et al*. OXA-1 beta-lactamase and non-susceptibility to penicillin/beta-lactamase inhibitor combinations among ESBL-producing *Escherichia coli*. *J Antimicrob Chemother* 2018. doi:10.1093/jac/dky453.
 - 243 Cheung CHP, Findlay J, Heesom KJ, Avison M. Resistance to aztreonam, in combination with a bicyclic boronate β -lactamase inhibitor in *Escherichia coli* identified following mixed culture selection. *bioRxiv* 2019; : 615336.
 - 244 Mojica MF, Ouellette CP, Leber A, Becknell MB, Ardura MI, Perez F *et al*. Successful Treatment of Bloodstream Infection Due to Metallo- β -Lactamase-Producing *Stenotrophomonas maltophilia* in a Renal Transplant Patient. *Antimicrob Agents Chemother* 2016; **60**: 5130–4.
 - 245 Spafford K, MacVane S, Humphries R. Evaluation of empiric beta-lactam susceptibility prediction among Enterobacteriaceae by molecular beta-lactamase gene testing. *J Clin Microbiol* 2019; : JCM.00674-19.
 - 246 Martínez-Martínez L, Pascual A, Hernández-Allés S, Alvarez-Díaz D, Suárez AI, Tran J *et al*. Roles of beta-lactamases and porins in activities of carbapenems and cephalosporins against *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 1999; **43**: 1669–73.
 - 247 Eliopoulos GM, Huovinen P. Resistance to Trimethoprim-Sulfamethoxazole. *Clin Infect Dis* 2001; **32**: 1608–1614.
 - 248 Card RM, Stubberfield E, Rogers J, Nunez-Garcia J, Ellis RJ, AbuOun M *et al*.

- Identification of a New Antimicrobial Resistance Gene Provides Fresh Insights Into Pleuromutilin Resistance in *Brachyspira hyodysenteriae*, Aetiological Agent of Swine Dysentery. *Front Microbiol* 2018; **9**: 1183.
- 249 Rodríguez-Baño J, Oteo J, Ortega A, Villar M, Conejo MC, Bou G *et al.* Epidemiological and clinical complexity of amoxicillin-clavulanate-resistant *Escherichia coli*. *J Clin Microbiol* 2013; **51**: 2414–7.
- 250 Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ *et al.* Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998-2016: a study of electronic health records. *Lancet Infect Dis* 2018; **18**: 1138–1149.
- 251 Shropshire WC, Aitken SL, Pifer R, Kim J, Bhatti MM, Li X *et al.* β -lactamase Amplification and Porin Loss Drive Progressive β -lactam Resistance in Recurrent ESBL Enterobacteriaceae Bacteremia. *bioRxiv* 2019; : 616961.
- 252 Harris PNA, Tambyah PA, Lye DC, Mo Y, Lee TH, Yilmaz M *et al.* Effect of Piperacillin-Tazobactam vs Meropenem on 30-Day Mortality for Patients With *E coli* or *Klebsiella pneumoniae* Bloodstream Infection and Ceftriaxone Resistance. *JAMA* 2018; **320**: 984.
- 253 Carter JJ, Walker TM, Walker AS, Whitfield MG, Morlock GP, Peto TE *et al.* Prediction of pyrazinamide resistance in *Mycobacterium tuberculosis* using structure-based machine learning approaches. *bioRxiv* 2019; : 518142.
- 254 Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* 2016; **107**: 1–8.
- 255 Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* 2018; **19**: 90.

- 256 Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol* 2016; **34**: 518–524.
- 257 Maio N De, Shaw LP, Hubbard A, George S, Sanderson N, Swann J *et al.* Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *bioRxiv* 2019; : 530824.
- 258 Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput Biol* 2017; **13**: e1005595.
- 259 Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015; **12**: 733–735.
- 260 Decraene V, Phan HTT, George R, Wyllie DH, Akinremi O, Aiken Z *et al.* A Large, Refractory Nosocomial Outbreak of *Klebsiella pneumoniae* Carbapenemase-Producing *Escherichia coli* Demonstrates Carbapenemase Gene Outbreaks Involving Sink Sites Require Novel Approaches to Infection Control. *Antimicrob Agents Chemother* 2018; **62**: e01689-18.
- 261 George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE *et al.* Resolving plasmid structures in enterobacteriaceae using the MinION nanopore sequencer: Assessment of MinION and MinION/illumina hybrid data assembly approaches. *Microb Genomics* 2017; **3**. doi:10.1099/mgen.0.000118.
- 262 Balleza E, López-Bojorquez LN, Martínez-Antonio A, Resendis-Antonio O, Lozada-Chávez I, Balderas-Martínez YI *et al.* Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiol. Rev.* 2009; **33**: 133–151.
- 263 Cho S, Cho Y, Lee S, Kim J, Yum H, Kim SC *et al.* Current Challenges in Bacterial Transcriptomics. *Genomics Inform* 2013; **11**: 76.

- 264 Hö J, Gorski SA, Rg Vogel J. Molecular Cell Technology Review Bacterial RNA Biology on a Genome Scale. 2018. doi:10.1016/j.molcel.2017.12.023.
- 265 Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* 2012; **13**. doi:10.1186/1471-2164-13-734.
- 266 Khaledi A, Schniederjans M, Pohl S, Rainer R, Bodenhofer U, Xia B *et al*. Transcriptome profiling of antimicrobial resistance in *Pseudomonas aeruginosa*. *Antimicrob Agents Chemother* 2016; **60**: 4722–4733.
- 267 Li Z, Xu M, Wei H, Wang L, Deng M. RNA-seq analyses of antibiotic resistance mechanisms in *Serratia marcescens*. *Mol Med Rep* 2019; **20**: 745–754.

Appendices

Chapter 2: Literature search strategy

There were two main methods which I used to identify relevant literature for my thesis. The first aimed to identify papers on WGS-based resistance prediction in bacteria. This method used a PubMed search with the key terms Antimicrobial Susceptibility Testing, Bacteria, Sequencing and their synonyms.

```
(((((Phenotype[MeSH Terms]) OR Microbial Sensitivity Tests[MeSH Terms])) OR
((((((((((((((((phenotyp*[Title/Abstract]) OR "minimum inhibitory concentration"[Title/Abstract]) OR
MIC[Title/Abstract]) OR breakpoint[Title/Abstract]) OR AST[Title/Abstract]) OR "disc diffusion"[Title/
Abstract]) OR "agar dilution"[Title/Abstract]) OR "broth dilution"[Title/Abstract]) OR microdilution[Title/
Abstract]) OR "gradient diffusion"[Title/Abstract]) OR "agar diffusion"[Title/Abstract]) OR "e test"[Title/
Abstract]) OR EUCAST[Title/Abstract]) OR CLSI[Title/Abstract]) OR BSAC[Title/Abstract]) OR "Clinical
Laboratory Standards Institute"[Title/Abstract]) OR "European Committee on Antimicrobial Susceptibility
Testing"[Title/Abstract]) OR "British Society for Antimicrobial Chemotherapy"[Title/Abstract]))) AND
((((((((((((((((Drug Resistance[MeSH Terms]) OR Bacterial/genetics*[MeSH Terms]) OR Bacteria/
genetics*[MeSH Terms]) OR Genome, Bacterial[MeSH Terms]) OR Drug Resistance, Bacterial/
genetics*[MeSH Terms]) OR Escherichia coli/genetics*[MeSH Terms]) OR Mycobacterium tuberculosis/
genetics*[MeSH Terms]) OR Staphylococcus aureus/genetics*[MeSH Terms]) OR Klebsiella Infections/
microbiology*[MeSH Terms]) OR Klebsiella pneumoniae/genetics*[MeSH Terms]) OR Salmonella paratyphi
A/genetics*[MeSH Terms]) OR Salmonella typhi/genetics*[MeSH Terms]) OR Drug Resistance, Multiple,
Bacterial/genetics*[MeSH Terms]) OR Shigella sonnei/classification[MeSH Terms]) OR Shigella sonnei/drug
effects*[MeSH Terms]) OR Shigella sonnei/genetics*[MeSH Terms]) OR Gram-Negative Bacteria* AND /
genetics[MeSH Terms]) OR Salmonella/genetics*[MeSH Terms]) OR Drug Resistance, Bacterial*[MeSH
Terms]) OR Genes, Bacterial*[MeSH Terms]) OR Genome, Bacterial*[MeSH Terms]) OR Pseudomonas
aeruginosa/genetics[MeSH Terms])) OR (((((((((((((((((((("micro organism"[Title/Abstract])
OR microorganism*[Title/Abstract]) OR bacteri*[Title/Abstract]) OR microb*[Title/Abstract]) OR
enterobacteria*[Title/Abstract]) OR "gram negative"[Title/Abstract]) OR "gram positive"[Title/Abstract]) OR
klebsiella[Title/Abstract]) OR pneumoniae[Title/Abstract]) OR oxytoca[Title/Abstract]) OR escherichia[Title/
Abstract]) OR shigella[Title/Abstract]) OR salmonella[Title/Abstract]) OR typhi[Title/Abstract]) OR
enterobacter[Title/Abstract]) OR cloacae[Title/Abstract]) OR acinitobacter[Title/Abstract]) OR
baumannii[Title/Abstract]) OR pseudomonas[Title/Abstract]) OR aeruginosa[Title/Abstract]) OR
mycobacterium[Title/Abstract]) OR tuberculosis[Title/Abstract]) OR abscessus[Title/Abstract]) OR
staphylococc*[Title/Abstract]) OR aureus[Title/Abstract]) OR "coagulase negative"[Title/Abstract]) OR
streptococc*[Title/Abstract]) OR pneumococc*[Title/Abstract]) OR enterococc*[Title/Abstract]) OR
clostridia*[Title/Abstract]) OR difficile[Title/Abstract]) OR neisseria[Title/Abstract]) OR gonorrhoeae[Title/
Abstract]) OR meningitidis[Title/Abstract]) OR chlamydia[Title/Abstract]) OR trachomatis[Title/Abstract])
OR campylobacter[Title/Abstract])))) AND (((((((Whole genome sequencing[MeSH Terms]) OR Sequence
Analysis, DNA[MeSH Terms]) OR Molecular Typing/methods*[MeSH Terms]) OR Sequence Analysis, DNA/
methods*[MeSH Terms]) OR High-Throughput Nucleotide Sequencing/methods*[MeSH Terms])) OR
((((((((("genom* sequenc*[Title/Abstract]) OR genotyp*[Title/Abstract]) OR sequenc*[Title/Abstract]) OR
"next generation"[Title/Abstract]) OR WGS[Title/Abstract]) OR NGS[Title/Abstract]) OR "pacific
biosciences"[Title/Abstract]) OR pacbio[Title/Abstract]) OR illumina[Title/Abstract]) OR nanopore[Title/
Abstract]) OR roche[Title/Abstract]))
```

Chapter 2 Supplementary Figure 1: Primary PubMed search strategy

The secondary method of identifying new papers was via automated alerts using Amedeo (<https://amedeo.com/>) and twitter updates from individual journals and the BioRxiv repositories

Chapter 3 Supplementary Table 1: Transmissible aminoglycoside resistance genes

Gene	Mechanism	No Isolates			
		OUH	PHE	APHA	Total
aac(3)-II	Antibiotic modification	52	62	22	136
aac(3)-IV	Antibiotic modification	1	0	22	23
aac(3)-VI	Antibiotic modification	1	1	0	2
aac(6')-I	Antibiotic modification	47	44	0	92
ant(2'')	Antibiotic modification	3	5	1	9
ant(3'')	Antibiotic modification	163	112	191	466
aph(3')-I	Antibiotic modification	304	129	206	639
aph(3')-II	Antibiotic modification	1	0	5	6
aph(3')-VI	Antibiotic modification	0	1	0	1
aph(4)-I	Antibiotic modification	1	0	24	25
aph(6)-I	Antibiotic modification	237	111	189	537
armA	Target modification	0	3	0	3
rmtC	Target modification	0	1	0	1

Chapter 3 Supplementary Table 2: Transmissible quinolone, trimethoprim and sulphonamide resistance

Gene	Mechanism	No Isolates			
		OUH	PHE	APHA	Total
aac(6')-Ib-cr	Quinolone: Antibiotic modification	47	43	0	90
QnrB	Quinolone: Target protection	3	1	5	9
QnrS	Quinolone: Target protection	9	3	63	75
QepA	Quinolone: Target protection	0	1	0	1
dfrA	Trimethoprim: Target replacement	305	178	287	770
dfrB	Trimethoprim: Target replacement	1	1	0	2
sul1	Sulphonamides: Target replacement	211	123	76	410
sul2	Sulphonamides: Target replacement	272	116	165	553
sul3	Sulphonamides: Target replacement	4	4	149	157

Chapter 3 Supplementary Table 3: Aminoglycoside phenotypes

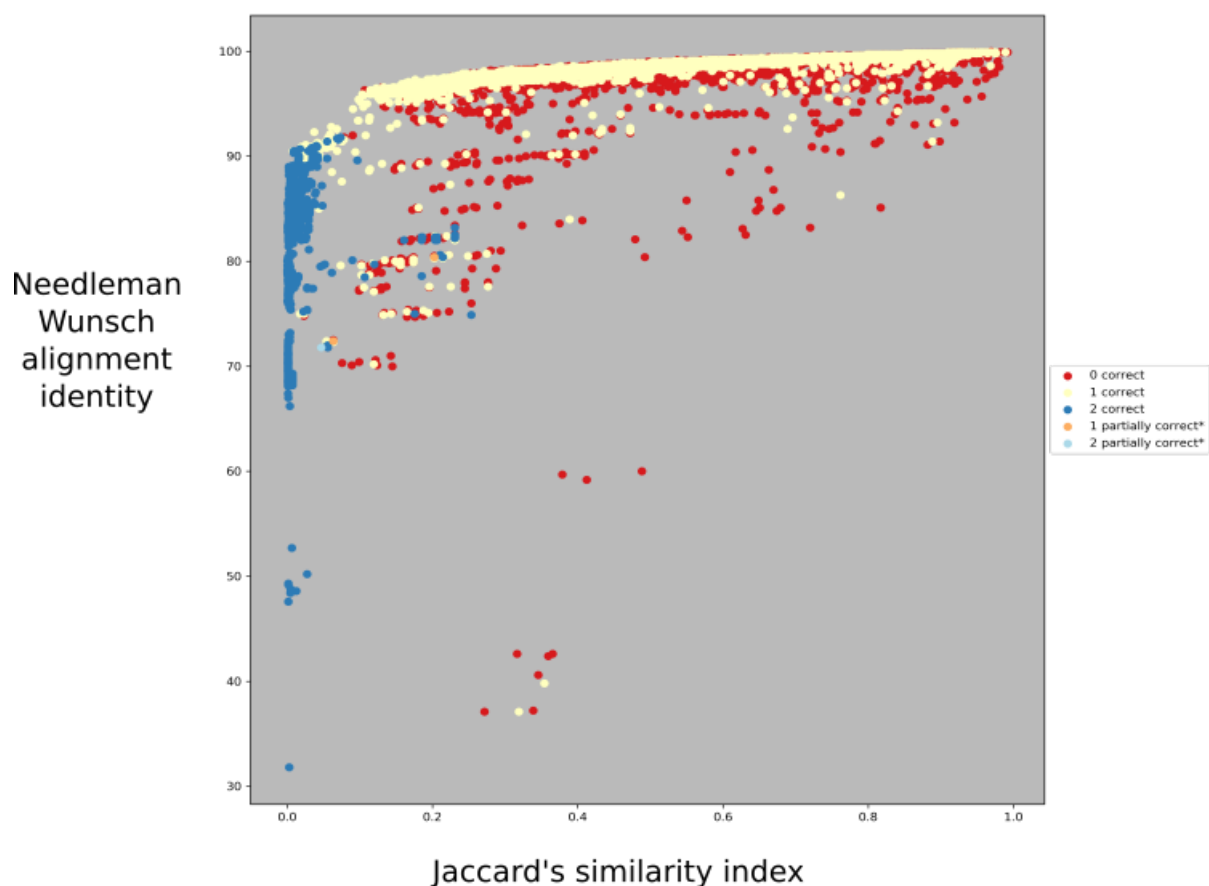
Dataset	Antibiotic	No isolates tested	Phenotype		
			S	I	R
OUH	Amikacin	982	975	7	0
	Gentamicin	982	912	11	59
	Tobramycin	981	893	9	79
APHA	Gentamicin	497	450	5	42
PHE	Amikacin	337	325	6	6
	Gentamicin	337	266	4	67
	Tobramycin	337	244	4	89

Chapter 3 Supplementary Table 4: Fluoroquinolone phenotypes

Dataset	Antibiotic	No isolates tested	Phenotype		
			S	I	R
OUH	Ciprofloxacin	981	818	25	138
	Levofloxacin	980	835	19	126
APHA	Ciprofloxacin	497	347	44	106
PHE	Ciprofloxacin	337	201	6	130
	Levofloxacin	337	206	3	128

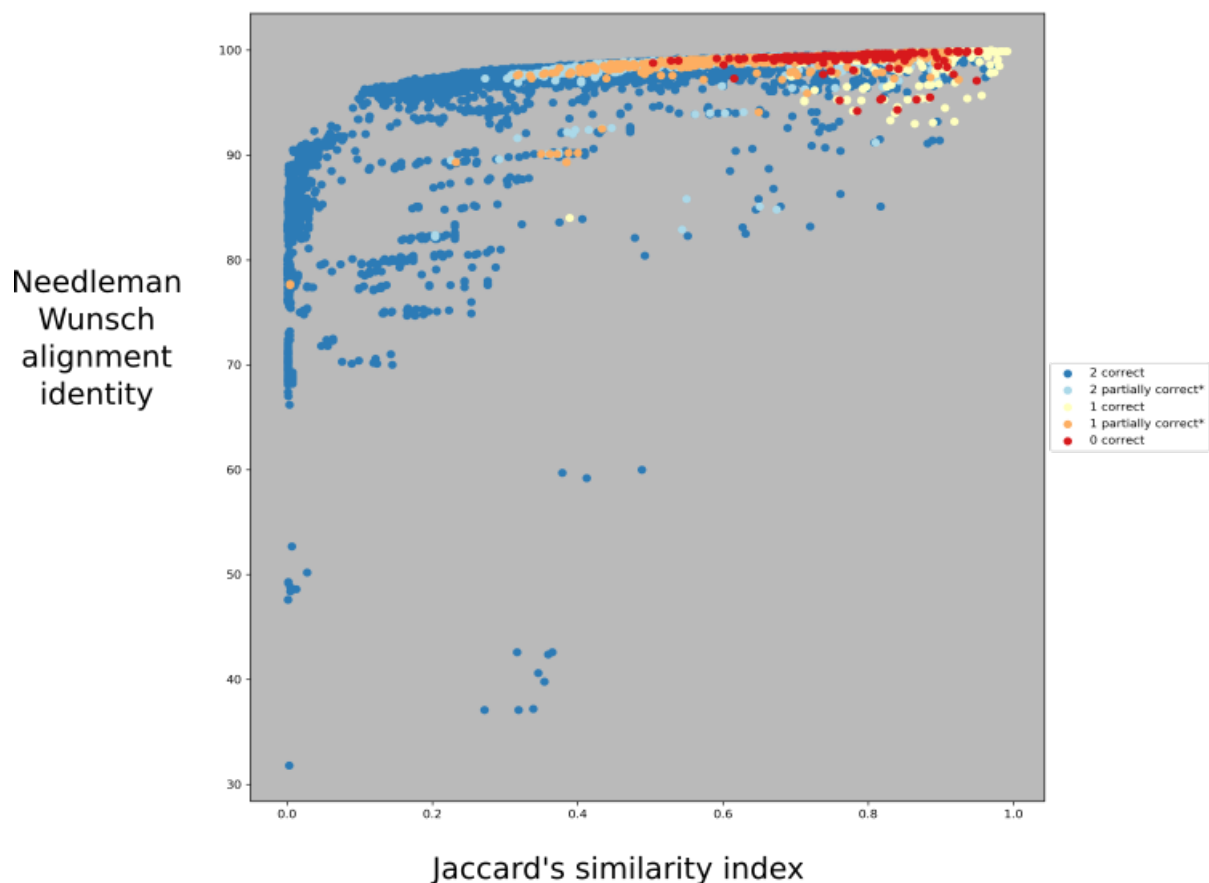
Chapter 3 Supplementary Table 5: Other antibiograms

Dataset	Antibiotic	No isolates tested	Phenotype		
			S	I	R
OUH	Colistin	982	979	-	3
	Trimethoprim	901	613	4	284
	Trimethoprim-sulfamethoxazole	980	707	5	268
	Nitrofurantoin	902	884	-	18
APHA	Trimethoprim-sulfamethoxazole	497	218	6	273
PHE	Colistin	337	337	-	0
	Trimethoprim	337	168	1	168
	Trimethoprim-sulfamethoxazole	337	177	0	160
	Nitrofurantoin	337	325	-	12



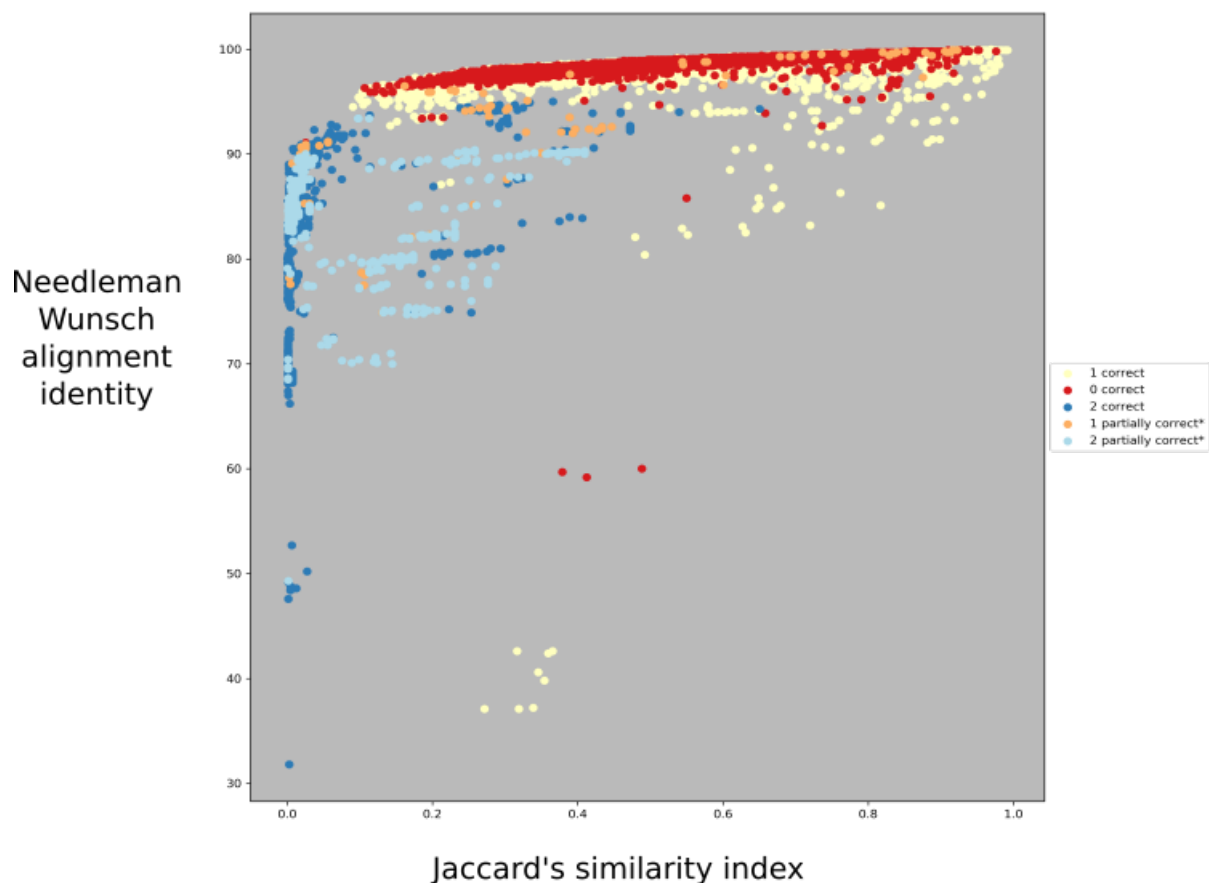
Chapter 5 Supplementary Figure 1: Similarity scores by pair similarity for ARIBA

Note: Scores obtained by ARIBA for each pair are signified by colour. To assess the impact of similarity, colours are imposed on previous plot of pair similarity Figure 36b where similarity according to two metrics is used to determine the x,y coordinates. Dark Blue = 2 correct alleles identified, no incorrect alleles identified. Light blue = 2 correct alleles identified, at least one incorrect allele identified. Yellow = 1 correct allele identified, no incorrect alleles identified. Orange = 1 correct allele identified, at least one incorrect allele identified. Red = 0 correct alleles identified.



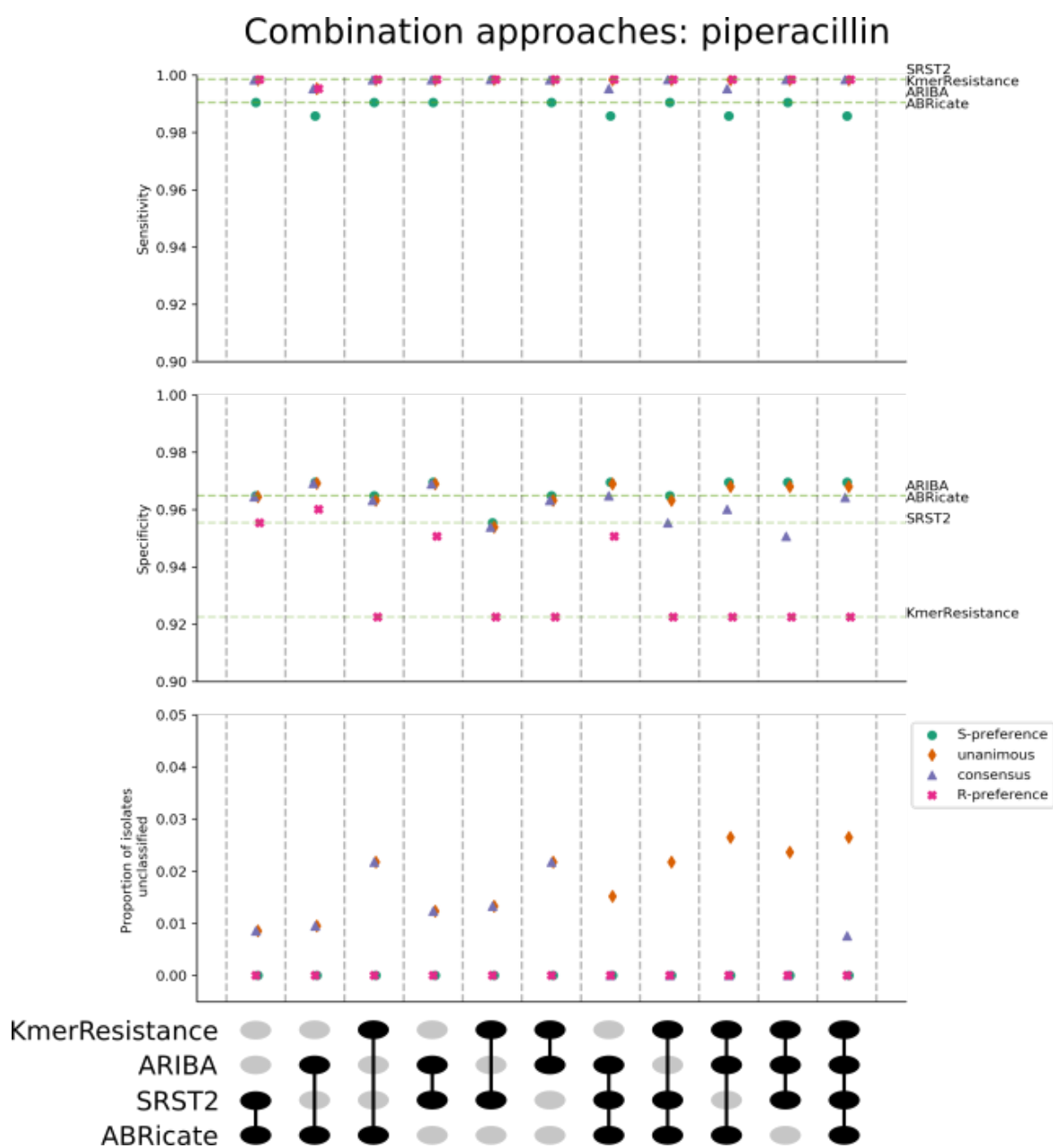
Chapter 5 Supplementary Figure 2: Similarity scores by pair similarity for KmerResistance

Note: Scores obtained by KmerResistance for each pair are signified by colour. To assess the impact of similarity, colours imposed on previous plot of pair similarity Figure 36b where similarity according to two metrics is used to determine the x,y coordinates. Dark Blue = 2 correct alleles identified, no incorrect alleles identified. Light blue = 2 correct alleles identified, at least one incorrect allele identified. Yellow = 1 correct allele identified, no incorrect alleles identified. Orange = 1 correct allele identified, at least one incorrect allele identified. Red = 0 correct alleles identified.



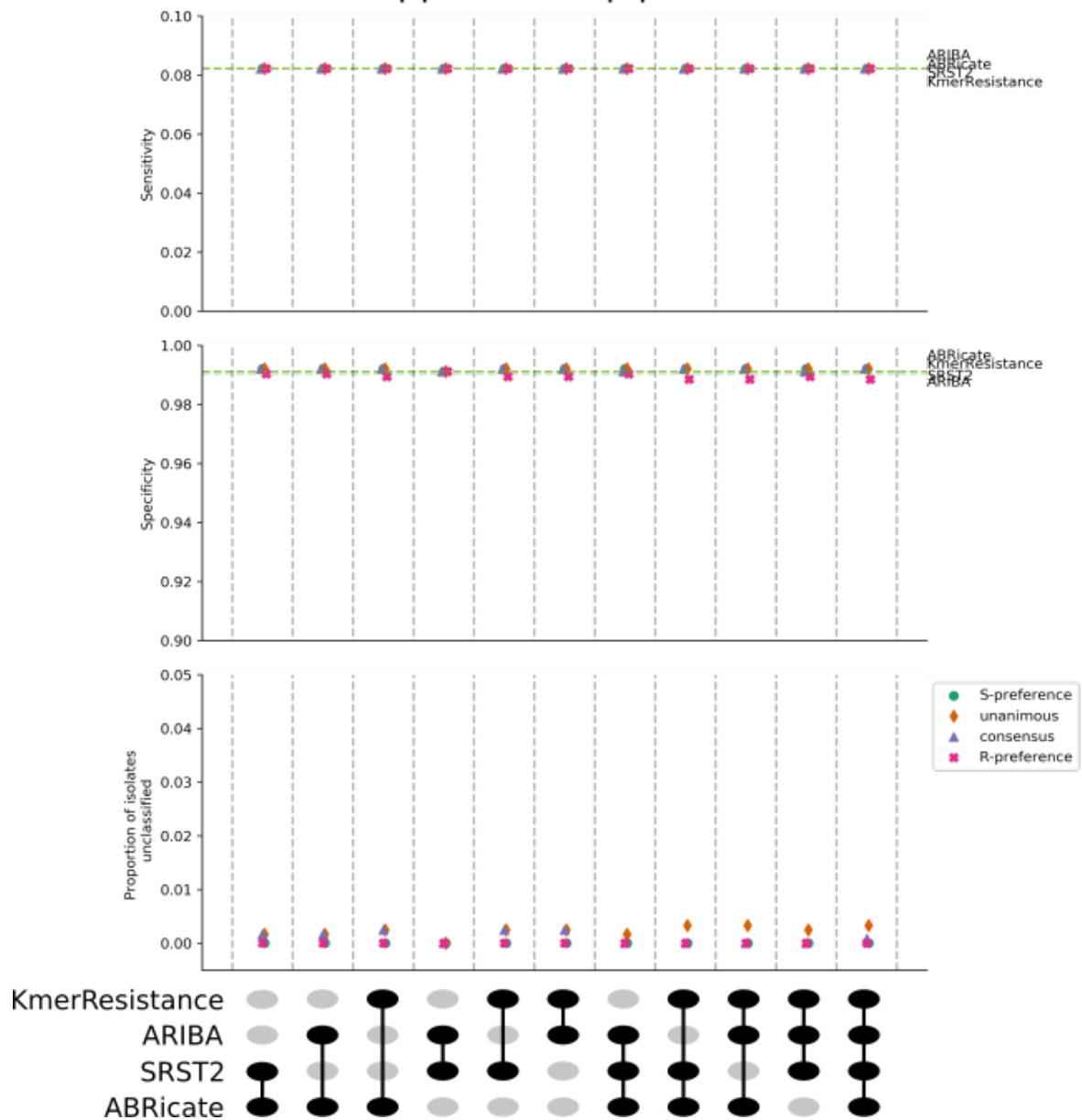
Chapter 5 Supplementary Figure 3: Similarity scores by pair similarity for SRST2

Note: Scores obtained by SRST2 for each pair are signified by colour. To assess the impact of similarity, colours imposed on previous plot of pair similarity Figure 36b where similarity according to two metrics is used to determine the x,y coordinates. Dark Blue = 2 correct alleles identified, no incorrect alleles identified. Light blue = 2 correct alleles identified, at least one incorrect allele identified. Yellow = 1 correct allele identified, no incorrect alleles identified. Orange = 1 correct allele identified, at least one incorrect allele identified. Red = 0 correct alleles identified.

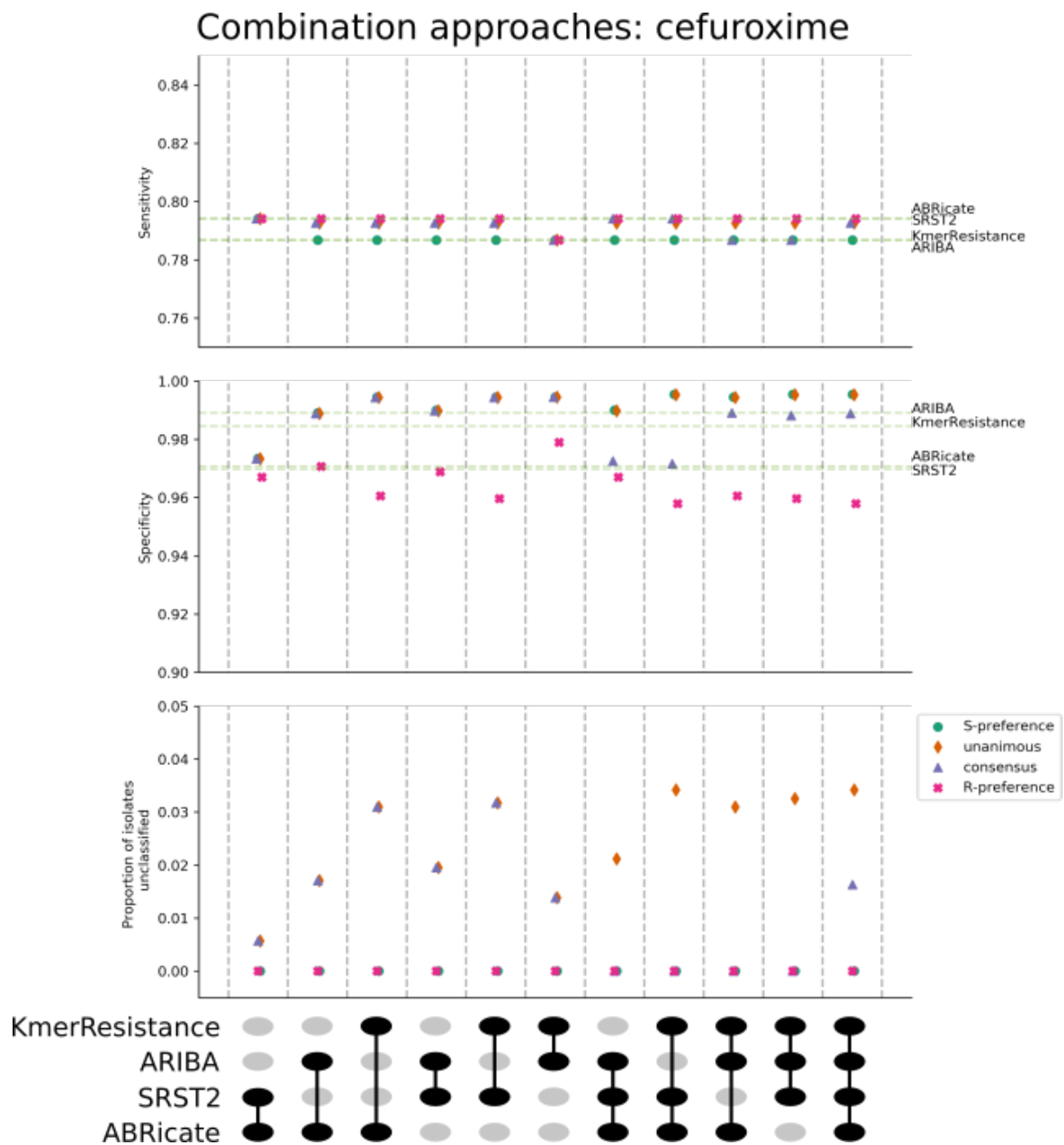


Chapter 5 Supplementary Figure 4: Combination approaches to WGS-based resistance prediction for piperacillin

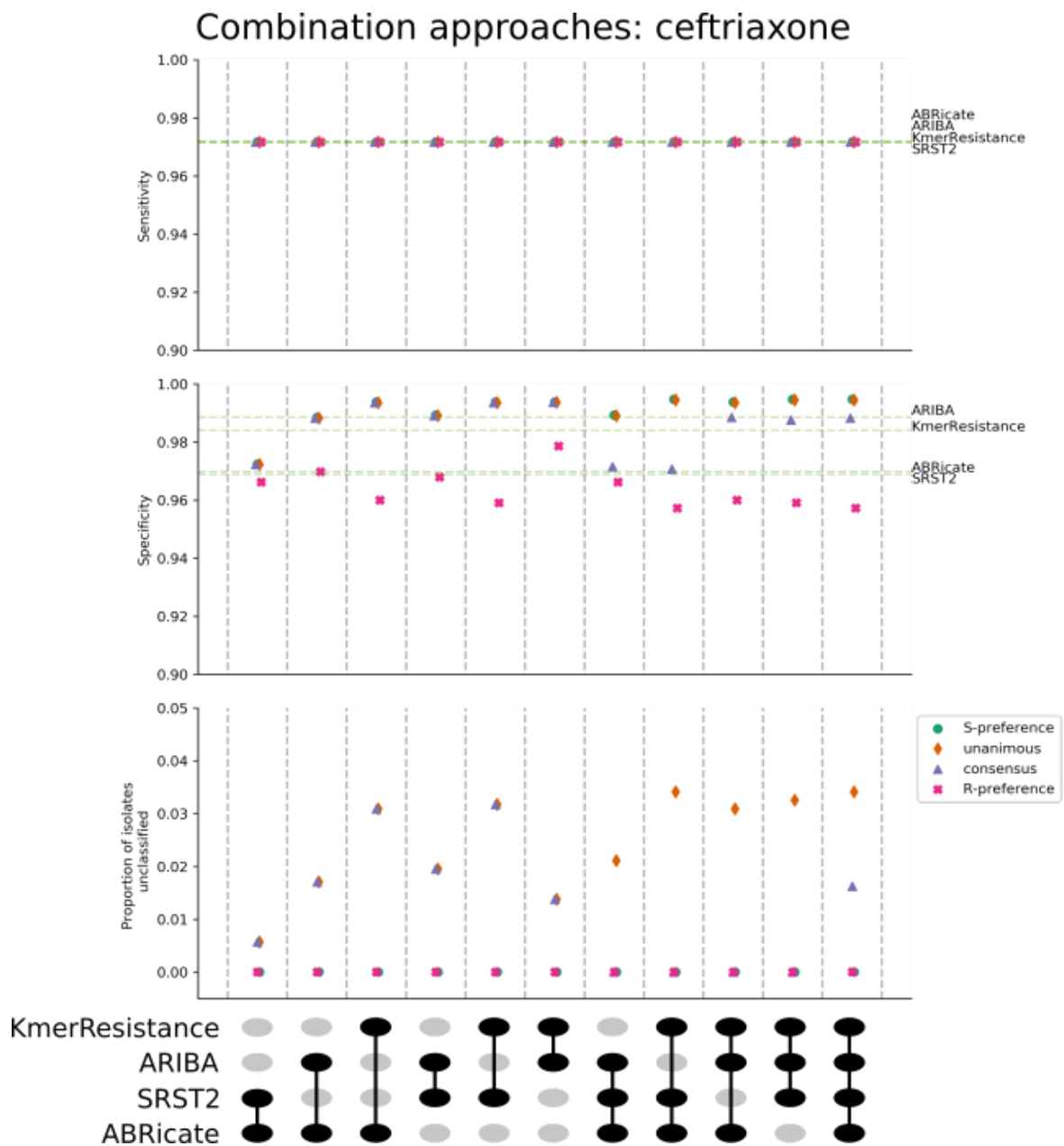
Combination approaches: piperacillin-tazobactam



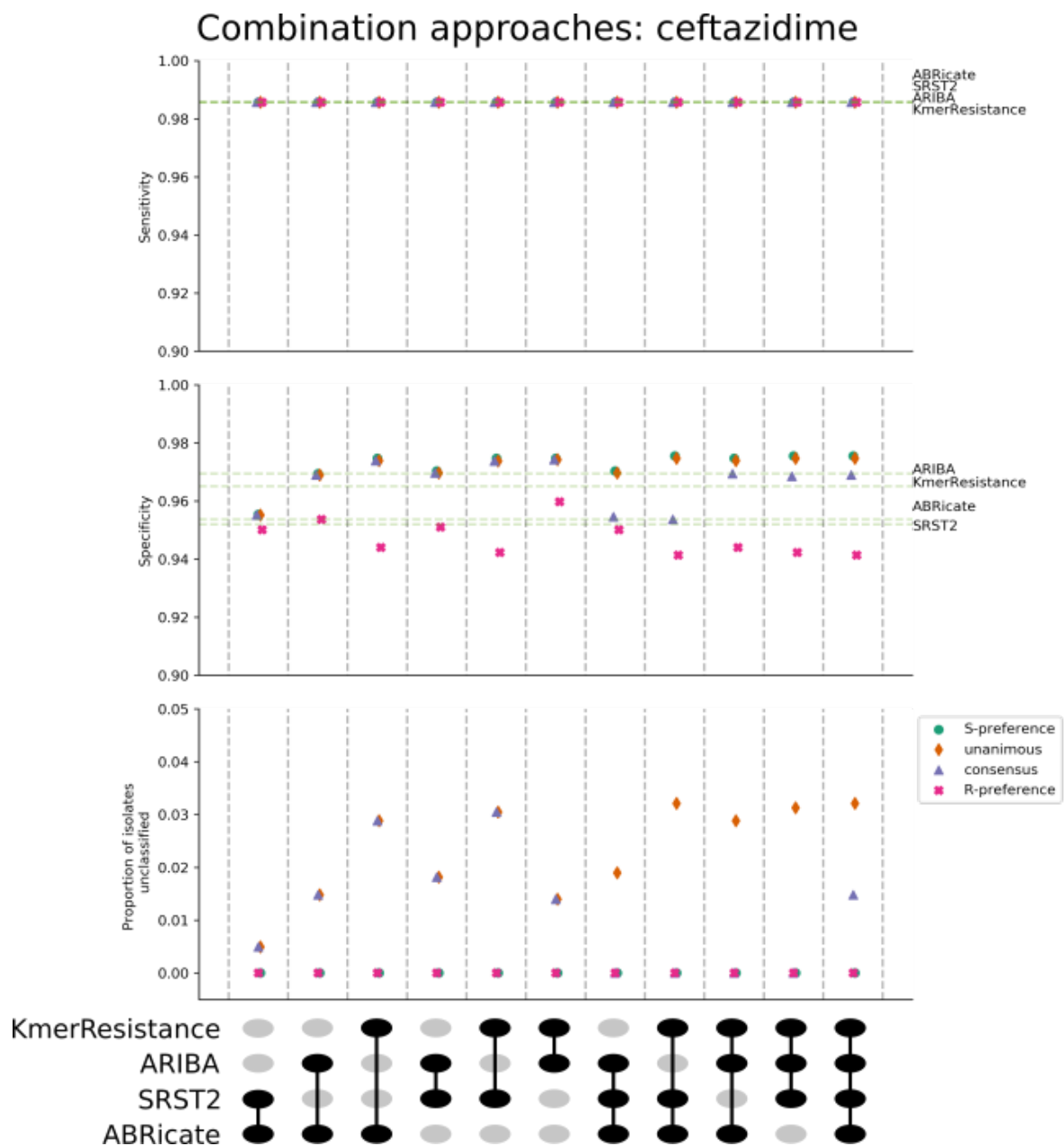
Chapter 5 Supplementary Figure 5: Combination approaches to WGS-based resistance prediction for piperacillin-tazobactam



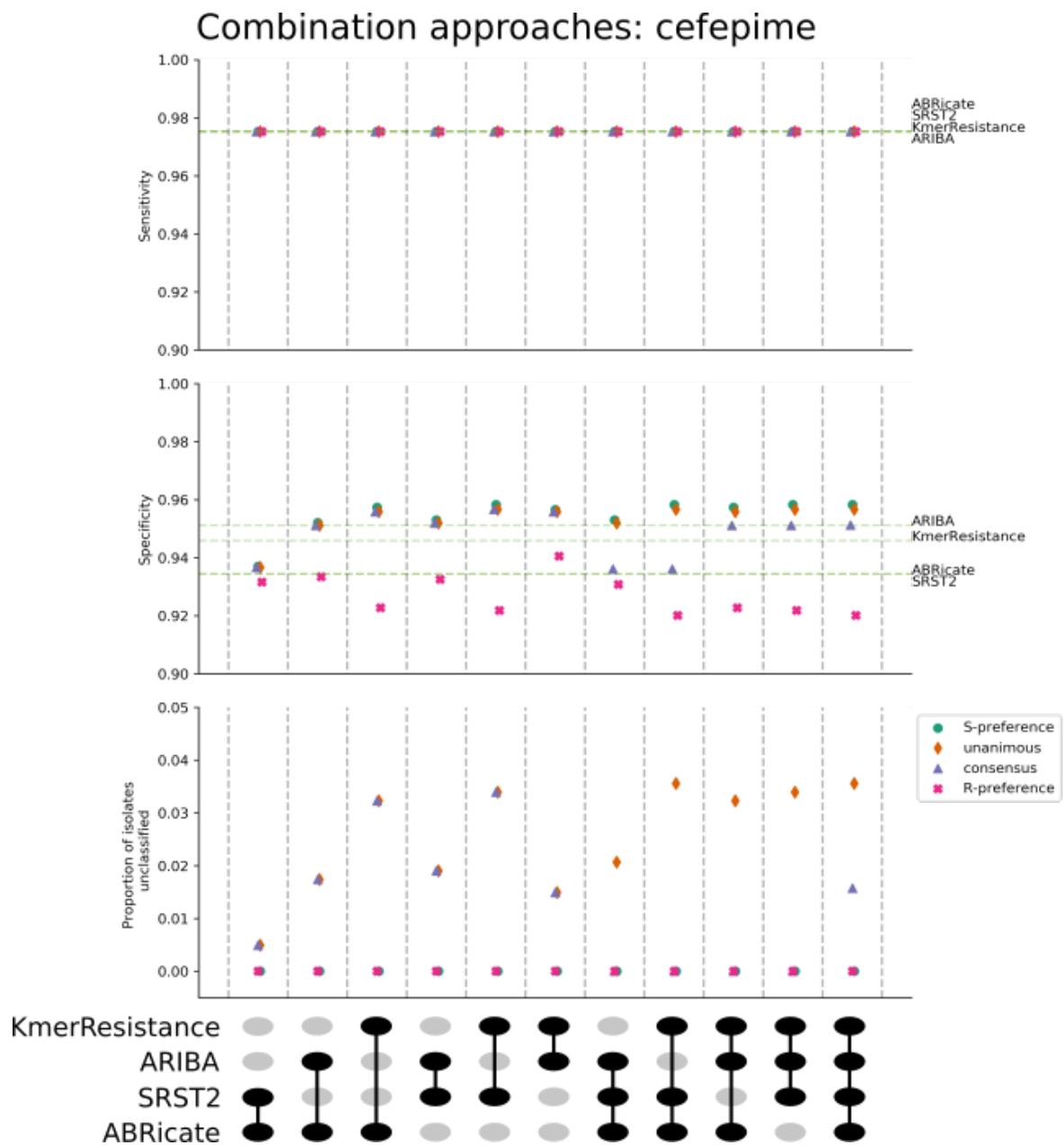
Chapter 5 Supplementary Figure 6: Combination approaches to WGS-based resistance prediction for cefuroxime



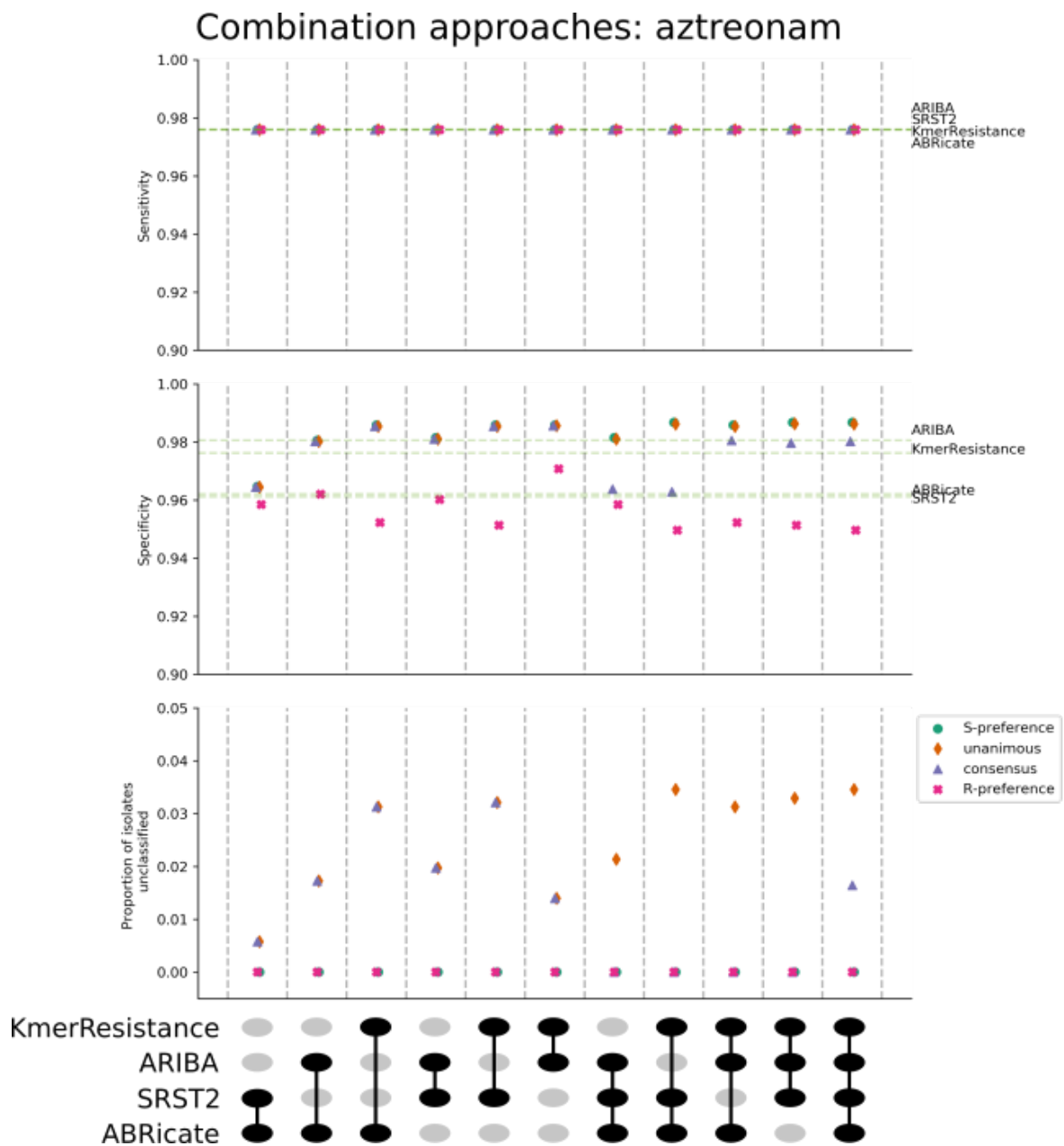
Chapter 5 Supplementary Figure 7: Combination approaches to WGS-based resistance prediction for ceftriaxone



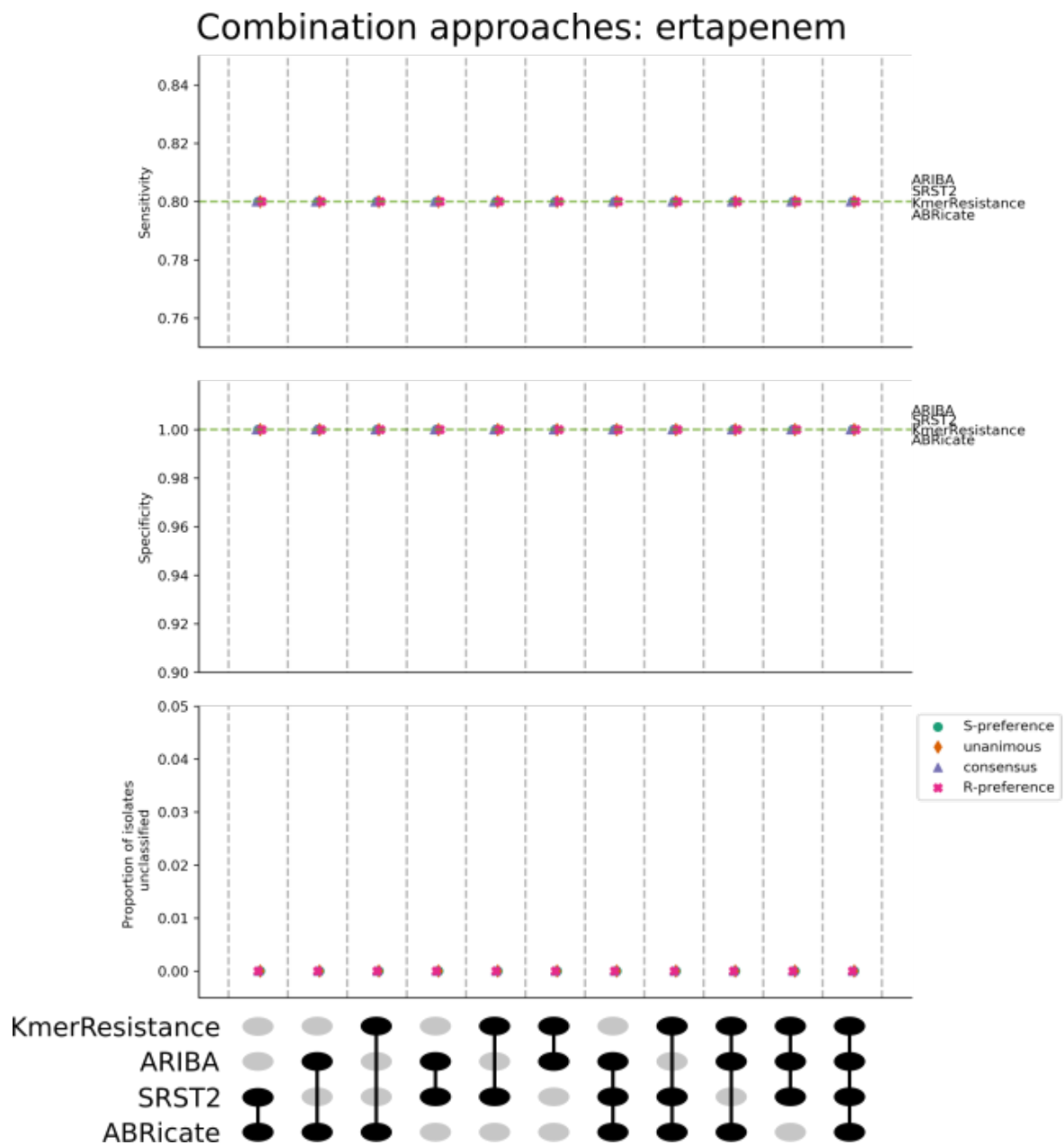
Chapter 5 Supplementary Figure 8: Combination approaches to WGS-based resistance prediction for ceftazidime



Chapter 5 Supplementary Figure 9: Combination approaches to WGS-based resistance prediction for cefepime



Chapter 5 Supplementary Figure 10: Combination approaches to WGS-based resistance prediction for aztreonam



Chapter 5 Supplementary Figure 11: Combinations approaches to WGS-based resistance prediction for ertapenem