Using genomics to understand the global spread of Escherichia coli and antimicrobial resistance

van der Putten, B.

Publication date
2022

Document Version
Final published version

Using genomics to understand the global spread of *Escherichia coli* and antimicrobial resistance Boas van der Putten

Using genomics to understand the global spread of *Escherichia coli* and antimicrobial resistance

Boas van der Putten

Using genomics to understand the global spread of *Escherichia coli* and antimicrobial resistance

Boas van der Putten

Colofon

ISBN	978-94-6421-718-6
Cover design	Marjan Taminiau
Lay-out design	Douwe Oppewal (www.oppewal.nl)
Printing	Ipskamp printing (https://proefschriften.net/)

Printing of this thesis was kindly supported by the Amsterdam UMC.

Printing of this thesis was financially supported by the Netherlands Society of Medical Microbiology (NVMM) and the Royal Netherlands Society for Microbiology (KNVM).

Funding statement
The research of this thesis was funded by grants from Netherlands Organization for Health, Research and Development (ZonMw; 50-51700-98-120), EU-H2020 programme (COMPARE, 643476), EU-Horizon2020 grant 727966 (PIGSs), the framework of the JPIAMR - Joint Programming Initiative on Antimicrobial Resistance – through the 3rd joint call, thanks to the generous funding by the Netherlands Organisation for Health Research and Development (ZonMw, grant number 547001012), the Federal Ministry of Education and Research (BMBF/DLR grant numbers 01KI1703A, 01KI1703C and 01KI1703B), the State Research Agency (AEI) of the Ministry of Science, Innovation and Universities (MINECO, grant number PCIN-2016-096), and the Medical Research Council (MRC, grant number MR/R002762/1).

Boas C.L. van der Putten was funded through an internal grant of the Amsterdam UMC ("flexibele OiO beurs")

Using genomics to understand the global spread of *Escherichia coli* and antimicrobial resistance

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op vrijdag 24 juni 2022, te 11.00 uur

door Boas Carolus Leopold van der Putten
geboren te HEUSDEN

Promotiecommissie

Promotor:	prof. dr. C. Schultsz	AMC-UvA
Copromotores:	dr. A. van der Ende	AMC-UvA
	dr. D.R. Mende	AMC-UvA
Overige leden:	prof. dr. B.H. ter Kuile	Universiteit van Amsterdam
	prof. dr. N.M. van Sorge	AMC-UvA
	prof. dr. C.A. Russell	AMC-UvA
	prof. dr. H.F.L. Wertheim	Radboud Universiteit Nijmegen
	dr. E. Franz	RIVM
	dr. A.E. Mather	Quadram Institute

Faculteit der Geneeskunde

Table of contents

Chapter 1

General introduction

Escherichia coli resistant to third-generation cephalosporins

Antimicrobial resistance (AMR) is rising across the globe, limiting the number of antimicrobials as effective therapeutic options. The global burden of AMR has proven very difficult to estimate[1] although some studies have currently estimated tens of thousands of attributable deaths in Europe[2] or possibly hundreds of thousands of attributable deaths globally per year[3]. These figures are expected to rise in the future, especially for low and middle income countries. AMR is a multifactorial problem, since factors driving an increase in resistance are diverse and many bacteria can become resistant to antimicrobials. However, some pathogens pose a more immediate threat to global health than others. Therefore, the World Health Organization has defined priority pathogens based on "bug-drug" combinations which require most attention[4]. Listed among the ones with the highest priority level are Enterobacterales resistant to carbapenems or third generation cephalosporins. In this thesis, I will focus on *Escherichia coli*, member of the Enterobacterales order, resistant to third generation cephalosporins.

Genetics of third-generation cephalosporin resistance in E. coli

Resistance to third generation cephalosporins in *E. coli* is mediated by various mechanisms. The most common mechanism is through the production of extended-spectrum β-lactamase (ESBL) enzymes which are able to inactivate third-generation cephalosporins[5].

β-lactamase enzymes can be classified in several ways on the basis of their molecular function, for example molecular class (A, B, C, D) and Bush-Jacoby group (e.g. 1, 2b, 2be, 2df)[6]. Additionally, β-lactamases can be classified into gene families, such as TEM, SHV, CTX-M or OXA[6]. Certain β-lactamases display an "extended spectrum" of activity. The term "extended broad-spectrum β-lactamases" originally referred to TEM and SHV β-lactamases which were able to hydrolyse oxyimino-cephalosporins and were inhibited by clavulanic acid[7]. Over the years, the term has shifted to "extended-spectrum β-lactamases" and has started to include a wider range of β-lactamases[7]. Currently, β-lactamases in Bush-Jacoby groups 2be, 2ber, 2de and 2e are typically considered ESBL enzymes[6]. Additionally, some extended-spectrum AmpC β-lactamases confer resistance to third-generation cephalosporins but are not commonly regarded as ESBL enzymes (e.g. bla_{CMY-2}, Bush-Jacoby group 1[8]). In *E. coli*, the most prevalent ESBL genes belong to the CTX-M gene family and to Bush-Jacoby group 2be[9].

ESBL genes, encoding ESBL enzymes, can be present on mobile genetic elements such as plasmids or integrons or can be in close vicinity of insertion sequences. These genetic elements allow efficient horizontal transfer of ESBL genes and thus of AMR. It is

suggested that CTX-M enzymes have entered the *E. coli* population through horizontal gene transfer from *Kluyvera* spp., possibly mobilised by ISEcp1, an insertion sequence commonly identified upstream of CTX-M genes[10]. Plasmids harbouring ESBL or other β-lactamase genes typically incur a small fitness cost on the bacterial host. However, bacterial hosts can accumulate mutations that offset the fitness cost of plasmid carriage (compensatory mutations)[11]. Additionally, some plasmids might incur a fitness cost on some bacterial hosts but not on others[12]. If plasmids are able to spread quickly enough and reach enough bacterial hosts capable of plasmid maintenance, these plasmids continue to exist despite potential fitness costs[12,13].

Bacterial hosts that have acquired ESBL genes gain an evolutionary advantage in environments where cephalosporins are encountered, thus promoting their further dissemination. Isolates of several *E. coli* lineages capable of efficient human colonisation harbour a limited set of ESBL genes. For particular combinations of lineage and ESBL gene, this can be a stable association over many years, facilitating the spread of ESBL genes. Isolates of sequence type (ST) 131 harbouring $bla_{\text{CTX-M-15}}$ which was first described in 2008, are forming the most notable example[14]. Currently, ST131 isolates are the most prevalent cause of extraintestinal *E. coli* infections[15] (e.g. urinary tract infections, meningitis or septicaemia). Reasons why ST131 has become globally prevalent might be its resistance to fluoroquinolones and third generation cephalosporins, adapted metabolism, acquisition of virulence factors or a combination of these and other factors[16]. The acquisition of the $bla_{\text{CTX-M-15}}$ gene offered the ST131 lineage resistance to third generation cephalosporins but at the same time offered $bla_{\text{CTX-M-15}}$ an opportunity to spread globally. Some other globally expanded lineages show a similar association with ESBL genes, such as two ST38 sublineages with the ESBL genes $bla_{\text{CTX-M-14}}$ or $bla_{\text{CTX-M-27}}$ (Chapter 4). However, it should be noted that resistance to third generation cephalosporins is not a prerequisite for global expansion of *E. coli* lineages (e.g. ST69 and ST73)[17].

One Health as an avenue to combat AMR

An obvious research avenue to combat the increasing prevalence of resistant *E. coli* is the development of novel antimicrobial therapies. This includes the development of new antimicrobials but also novel applications of existing antimicrobials. Additionally, there is increasing attention for the use of vaccines to reduce AMR. Although these strategies have produced potential useful novel therapies in recent years[18–20], additional strategies are required to curb the increasing spread of AMR pathogens[4].

Limiting the spread of ESBL genes could be an additional strategy to reduce the prevalence of AMR. This requires better understanding how ESBL genes transmit through and across populations. These populations can be defined based on geography (e.g. the spread of resistant bacteria between countries), based on host species (e.g. the spread of

resistant bacteria between humans and other animals) or on other criteria. Understanding the spread of resistant bacteria between populations can be best achieved using the conceptual framework of One Health[21]. One Health implies that the health of the human population is connected to the environment and populations of other species.

The One Health framework is highly applicable to questions in medical microbiology, as many if not most current human infectious diseases originated from animals[22]. This framework can also be applied to understand the increasing prevalence of ESBL genes. The World Organisation for Animal Health (OIE) estimated that approximately 1100 tonnes of cephalosporins were administered in animals in 2017[23]. Of these, approximately 700 tonnes were third or fourth generation cephalosporins administered in Asia and Oceania. Cephalosporin usage in animals drives the selection of (novel) resistance genes which could potentially spread to the human population. Although resistant *E. coli* from animal hosts seem to colonise humans infrequently[24,25], the human-animal interface remains a relevant pathway of AMR spread. An important example is *mcr*-1, a gene conferring resistance to the last resort antimicrobial colistin. Presence of *mcr*-1 was shown in three *E. coli* samples isolated in China in the 1980s, coinciding with the first use of colistin in food-producing animals in China[26]. The *mcr*-1 gene is now widespread in the human population, contributing to the increasing rate of colistin resistance of *E. coli*. Once bacteria with such resistance genes have reached the human population, a combination of sustained antibiotic pressure with potentially low fitness costs for the bacterial host harbouring them can result in widespread dissemination of resistance. The One Health concept can also be applied to other definitions of populations, such as populations separated by geography. Examples of this are the international emergence of *E. coli* O157:H7[27] or the findings presented in Chapter 4 of this thesis.

Bioinformatics

To study the complex epidemiology of *E. coli* in populations, genetic analyses have been developed. A landmark study introduced the multilocus sequence typing (MLST) method[28]. Originally, MLST involves sequencing (parts of) selected housekeeping genes using ddNTP Sanger sequencing. A unique number is subsequently assigned to each unique allele and these designations are stored in a central database. The allele numbers combined result in a genetic barcode, referred to as a sequence type (ST). For *E. coli* for example, the Warwick MLST scheme sequences parts of seven genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* and *recA*). As an example, the sequence type of an ST95 *E. coli* isolate is expressed as *adk*(37) *fumC*(38) *gyrB*(19) *icd*(37) *mdh*(17) *purA*(11) *recA*(26). Even though MLST was first published more than 20 years ago, the ST designations still hold value for many bacterial pathogens including *E. coli*. As molecular epidemiology progressed

using whole-genome sequencing (WGS) and sophisticated analysis methods, it turned out that STs capture the identities of naturally occurring lineages very reliably and that ST names facilitate easy communication about bacterial lineages.

Due to the advent of cost-efficient WGS, microbiologists have increasingly turned to whole-genome analyses[29]. WGS offers a number of advantages over earlier genetic typing methodologies. Where classic *E. coli* MLST investigates (parts of) seven genes, WGS provides information on the whole *E. coli* genome typically consisting of ~5000 genes and other genetic elements. The entire genome can be analysed from WGS data, enabling a higher resolution in analyses of relatedness, but also enabling typing of all antimicrobial resistance or virulence genes harboured by the strain. These advantages of WGS were clearly demonstrated in a seminal 2010 study, when WGS was applied to study the transmission and resistance mutations of methicillin-resistant *Staphylococcus aureus* (MRSA) in a hospital setting[30]. Currently, large collections of bacterial isolates are routinely whole-genome sequenced, reflected by >180,000 whole-genome sequenced *E. coli* isolates in the database Enterobase[31]. The increased accessibility of bacterial WGS has enabled research not previously possible at large scale. A limited number of examples relevant to *E. coli* would include research on adaptation to avian hosts[32], large-scale population structure[31], ST131 evolution[33,34] and the spread of resistance genes[35].

A typical WGS analysis of a bacterial isolate sequenced on an Illumina platform involves a number of steps (also covered by Schürch *et al*.[36]). The very first step is often sequence read quality control. This step comprises the trimming of low-quality ends of sequencing reads, discarding complete sequence reads if these are of bad quality, trimming adapter sequences and/or correction of sequencing errors. This ensures that subsequent analyses use good quality data. If too many issues are encountered at this quality control step, the isolate usually needs to be resequenced. The genome can be reconstructed from the remaining good quality sequence reads using *de novo* assembly[37]. This analysis yields a draft genome assembly consisting of contiguous genomic segments that could be reconstructed with high certainty ("contigs"). Genome assembly of only Illumina data almost always results in a fragmented draft assembly, as *de novo* assembly cannot resolve all repeats in a typical bacterial genome. To complete bacterial assemblies, one typically needs long read sequencing from e.g. Oxford Nanopore Technologies or Pacific Biosciences as sequence reads from these technologies span many repetitive regions, often allowing a complete resolution of an *E. coli* genome[38].

To infer meaning from assembled genomic data, the genomes need to be annotated which can be done in several ways[39]. A general annotation strategy attempts to identify all protein-coding genes by scanning for open reading frames. The predicted genes that were identified are then compared to large databases of proteins or protein domains with

known functions. Specific annotation is also possible, e.g. for antimicrobial resistance genes, virulence genes or genes from other dedicated databases.

For a set of bacterial isolates, the core genome, accessory genome and pangenome can be defined[40]. The core genome consists of the genes present in nearly all (e.g. >99%) genomes in the set of isolates. The accessory genome consists of all genes not belonging to the core genome. Antimicrobial resistance genes are typically a part of the accessory genome. Finally, the pangenome consists of all genes identified in any genome in the set of isolates, and thus forms the sum of the core and accessory genome.

To identify relatedness between bacterial isolates, comparative analyses are needed. As the core genome is present in nearly all isolates, these genes can be aligned to assess relatedness between isolates[36]. This alignment can be achieved in various ways, roughly divisible into two strategies: reference-based read mapping (heavily used in Chapter 4) and core gene alignment. Reference-based read mapping involves selecting a reference genome, on which the sequence reads can be mapped. Multiple isolates, all mapped to the same reference genome, can be compared based on the regions shared by all isolates and the reference. The other strategy, core gene alignment, typically involves the identification of all protein-coding genes in a set of isolates (comprising the pangenome) and performing an all-versus-all comparison between these genes. Some genes will be present in all isolates (core protein-coding genes), and these can subsequently be aligned and compared to assess similarity between isolates. For both strategies, differences are commonly expressed in single nucleotide polymorphisms (SNPs) which signify single nucleotide differences between isolates. Note that the first approach is heavily influenced by the choice of reference genome, and the second approach typically only considers protein-coding genes.

The results obtained with both reference-based read mapping and core gene alignment are dependent on the set of isolates. For another set of isolates, the defined core genome will almost always be different, hindering comparisons between multiple sets of isolates. Stable typing schemes can also be extracted from WGS data. Multilocus sequence types can be extracted from draft genomes and are still ubiquitous in epidemiological analyses. More comprehensive typing schemes such as core genome or whole genome MLST (cgMLST and wgMLST, respectively) offer an increased resolution compared to classic MLST and provide stable types[36]. Similar to classic MLST, cgMLST and wgMLST identify alleles of defined genes and translate these to a genomic "barcode". The difference between cgMLST and wgMLST is the group of genes considered in the comparison. As indicated by its name cgMLST only considers core genes, meaning 2513 genes for *E. coli* in the Enterobase cgMLST scheme[31]. WgMLST on the other hand considers all known genes for a species. In the Enterobase wgMLST scheme, this comprises >25,000 genes with defined alleles.

Finally, methods have been developed to associate phenotypes with bacterial genotypes, typically referred to as genome-wide association studies (GWASs)[41]. The basic principle behind these methods is to assess whether a given phenotype (e.g. resistance to third-generation cephalosporins) is more common for bacteria with a particular genotype. Genotypes can be defined in a number of ways: the presence of genes, the presence of SNPs, or more abstracted representations of genetic elements such as k-mers or unitigs. There are a number of challenges for these kinds of analyses. Usually the number of genotypes examined is far larger than the number of isolates included for which the phenotype is known. Due to this phenomenon, a GWAS requires very thorough multiple testing correction. Secondly, many bacteria display a high degree of linkage disequilibrium. This means that the presences of bacterial gene variants are associated with each other, for example due to shared ancestry. In practice this means that if gene A and gene B co-occur often due to shared ancestry, and only gene A causes the phenotype under investigation, both genes A and B will appear associated to the phenotype under investigation. Due to this phenomenon, a bacterial GWAS has to correct for population structure. Multiple methods can be used for this correction, including linear mixed models[42] and elastic net models[43]. A recent benchmarking study found that elastic net models, which have only recently been introduced in bacterial GWAS, perform well although there remains ample room for improvement[44].

Combining all methods summarised above, an enormous amount of information can be extracted from WGS data. AMR phenotypes can be predicted from inferred genes, transmission events can be reconstructed from genomic similarity and the genetic basis of phenotypes can be determined using GWAS. However, this thesis also addresses several conditions which are necessary for successful WGS analysis. For example, to accurately predict AMR phenotypes from WGS data, databases with curated genotype-phenotype relationships are needed (addressed in Chapter 2). Additionally, public sequence databases need to be complete and their contents accurately described (addressed in Chapters 3 and 8). Finally, it needs to be established through benchmarking which bioinformatics software works best on particular data and for which analysis (addressed in Chapter 6) and bioinformatics software should be tested thoroughly to ensure accuracy and reproducibility of analyses (addressed in Chapter 7).

Chapter outline

For this thesis, I have investigated which *E. coli* types can colonise humans efficiently, and which genetic elements drive this adaptation. These efficiently colonising *E. coli* types are important to consider in pathogen surveillance, as these might cause disease or play a key role in the transmission of AMR genes between populations. Unravelling

the genetics of these widely spread *E. coli* types helps in understanding their biology and to assess their potential risks to public health. Additionally, I have investigated and developed methodologies and datasets enabling impactful WGS analysis. These include phylogenetic methodologies, phylogenetic classification of *Escherichia* species, software testing methodologies, reference genome datasets and information on AMR genotype-phenotype relationships.

In the first two chapters after the introduction, I present studies introducing AMR and taxonomy of the *Escherichia* genus. Chapter 2 comprises a systematic review of ciprofloxacin resistance mechanisms in *E. coli*. Chapter 3 presents the description of a novel *Escherichia* species, *Escherichia ruysiae*.

Chapters 4 and 5 set out the main findings presented in this thesis. Chapter 4 investigates a cohort of international travellers which have returned with newly acquired ESBL-Ec. While many travellers lose the ESBL-Ec within a month, some travellers harbour the ESBL-Ec for more than a year after returning from travel. In this chapter, I investigate which ESBL-Ec lineages are associated with this long-term carriage. In Chapter 5 I investigate which genetic elements allow *E. coli* to colonise humans. To this end, a large and diverse collection of *E. coli* isolates, isolated from five different host species, is established. I compare the genomes of *E. coli* from humans with *E. coli* from other host species and present experimental characterisation for genes contributing to human colonisation.

Finally, I present research on development of methodologies and datasets, supporting other chapters. These final chapters comprise a benchmarking study of phylogenetic methods (Chapter 6), a commentary on software testing to ensure reliability and reproducibility of bioinformatic analyses (Chapter 7) and the generation of complete genomes through the combination of short and long read data (Chapter 8). Chapter 9 applies concepts from Chapter 2 (systematic review) and Chapter 5 (genetics of host adaptation) to the zoonotic swine pathogen *Streptococcus suis*. Finally, Chapter 10 contextualises the prior chapters and discusses avenues for future research.

References

1. Limmathurotsakul D, Dunachie S, Fukuda K, *et al.* Improving the estimation of the global burden of antimicrobial resistant infections. *Lancet Infect Dis*. 2019;19(11):e392-e398. doi:10.1016/S1473-3099(19)30276-2
2. Cassini A, Högberg LD, Plachouras D, *et al.* Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis*. 2019;19(1):56-66. doi:10.1016/S1473-3099(18)30605-4
3. O'Neill J. Antimicrobial resistance. *Tackling Crisis Health Wealth Nations*. Published online 2014.
4. Beyer P, Paulin S. Priority pathogens and the antibiotic pipeline: an update. *Bull World Health Organ*. 2020;98(3):151. doi:10.2471/BLT.20.251751
5. Bush K, Fisher JF. Epidemiological Expansion, Structural Studies, and Clinical Challenges of New β-Lactamases from Gram-Negative Bacteria. *Annu Rev Microbiol*. 2011;65(1):455-478. doi:10.1146/annurev-micro-090110-102911
6. Bush K, Jacoby GA. Updated Functional Classification of β-Lactamases. *Antimicrob Agents Chemother*. 2010;54(3):969-976. doi:10.1128/AAC.01009-09
7. Livermore DM. Defining an extended-spectrum beta-lactamase. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. 2008;14 Suppl 1:3-10. doi:10.1111/j.1469-0691.2007.01857.x
8. Martin LC, Weir EK, Poppe C, Reid-Smith RJ, Boerlin P. Characterization of bla_{CMY-2} Plasmids in *Salmonella* and *Escherichia coli* Isolates from Food Animals in Canada. *Appl Environ Microbiol*. 2012;78(4):1285-1287. doi:10.1128/AEM.06498-11
9. Bevan ER, Jones AM, Hawkey PM. Global epidemiology of CTX-M β-lactamases: temporal and geographical shifts in genotype. *J Antimicrob Chemother*. 2017;72(8):2145-2155. doi:10.1093/jac/dkx146
10. Humeniuk C, Arlet G, Gautier V, Grimont P, Labia R, Philippon A. Beta-lactamases of *Kluyvera ascorbata*, probable progenitors of some plasmid-encoded CTX-M types. *Antimicrob Agents Chemother*. 2002;46(9):3045-3049. doi:10.1128/AAC.46.9.3045-3049.2002
11. San Millan A, MacLean RC. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiol Spectr*. 2017;5(5):5.5.02. doi:10.1128/microbiolspec.MTBP-0016-2017
12. Alonso-del Valle A, León-Sampedro R, Rodríguez-Beltrán J, *et al.* Variability of plasmid fitness effects contributes to plasmid persistence in bacterial communities. *Nat Commun*. 2021;12(1):2653. doi:10.1038/s41467-021-22849-y
13. Ranjan A, Scholz J, Semmler T, *et al.* ESBL-plasmid carriage in *E. coli* enhances *in vitro* bacterial competition fitness and serum resistance in some strains of pandemic sequence types without overall fitness cost. *Gut Pathog*. 2018;10(1):24. doi:10.1186/s13099-018-0243-z
14. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, *et al*. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J Antimicrob Chemother*. 2008;61(2):273-281. doi:10.1093/jac/dkm464
15. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *Clin Microbiol Rev*. 2019;32(3). doi:10.1128/CMR.00135-18
16. McNally A, Kallonen T, Connor C, *et al.* Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection. *mBio*. 2019;10(2). doi:10.1128/mBio.00644-19

17. Kallonen T, Brodrick HJ, Harris SR, *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;27:1437-1449. doi:10.1101/gr.216606.116

18. Jansen KU, Knirsch C, Anderson AS. The role of vaccines in preventing bacterial antimicrobial resistance. *Nat Med*. 2018;24(1):10-19. doi:10.1038/nm.4465

19. Imai Y, Meyer KJ, Iinishi A, *et al.* A new antibiotic selectively kills Gram-negative pathogens. *Nature*. 2019;576(7787):459-464. doi:10.1038/s41586-019-1791-1

20. Kaki R, Elligsen M, Walker S, Simor A, Palmay L, Daneman N. Impact of antimicrobial stewardship in critical care: a systematic review. *J Antimicrob Chemother.* 2011;66(6):1223-1230. doi:10.1093/jac/dkr137

21. Puyvelde SV, Deborggraeve S, Jacobs J. Why the antibiotic resistance crisis requires a One Health approach. *Lancet Infect Dis*. 2018;18(2):132-134. doi:10.1016/S1473-3099(17)30704-1

22. Jones KE, Patel NG, Levy MA, *et al.* Global trends in emerging infectious diseases. *Nature*. 2008;451(7181):990-993. doi:10.1038/nature06536

23. OIE. OIE Annual Report on Antimicrobial Agents Intended for Use in Animals, Fifth Report. Published online 2021. https://www.oie.int/app/uploads/2021/05/a-fifth-annual-report-amr.pdf

24. Ludden C, Raven KE, Jamrozy D, *et al.* One Health Genomic Surveillance of *Escherichia coli* Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *mBio*. 2019;10(1):e02693-18. doi:10.1128/mBio.02693-18

25. Nguyen VT, Jamrozy D, Matamoros S, *et al.* Limited contribution of non-intensive chicken farming to ESBL-producing *Escherichia coli* colonization in humans in Vietnam: an epidemiological and genomic analysis. *J Antimicrob Chemother.* 2019;74(3):561-570. doi:10.1093/jac/dky506

26. Shen Z, Wang Y, Shen Y, Shen J, Wu C. Early emergence of *mcr*-1 in *Escherichia coli* from food-producing animals. *Lancet Infect Dis*. 2016;16(3):293. doi:10.1016/S1473-3099(16)00061-X

27. Franz E, Rotariu O, Lopes BS, *et al.* Phylogeographic Analysis Reveals Multiple International transmission Events Have Driven the Global Emergence of *Escherichia coli* O157:H7. *Clin Infect Dis*. 2019;69(3):428-437. doi:10.1093/cid/ciy919

28. Maiden MC, Bygraves JA, Feil E, *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*. 1998;95(6):3140-3145. doi:10.1073/pnas.95.6.3140

29. MacLean D, Jones JDG, Studholme DJ. Application of "next-generation" sequencing technologies to microbial genetics. *Nat Rev Microbiol*. 2009;7(4):96-97. doi:10.1038/nrmicro2088

30. Harris SR, Feil EJ, Holden MTG, *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*. 2010;327(5964):469-474. doi:10.1126/science.1182395

31. Zhou Z, Alikhan NF, Mohamed K, *et al.* The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res*. 2020;30(1):138-152. doi:10.1101/gr.251678.119

32. Mageiros L, Méric G, Bayliss SC, *et al.* Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. *Nat Commun*. 2021;12(1):765. doi:10.1038/s41467-021-20988-w

33. Stoesser N, Sheppard AE, Pankhurst L, *et al.* Evolutionary History of the Global Emergence of the *Escherichia coli* Epidemic Clone ST131. *mBio*. 7(2):e02162-15. doi:10.1128/mBio.02162-15

34. Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, *et al.* Sequential Acquisition of Virulence and Fluoroquinolone Resistance Has Shaped the Evolution of *Escherichia coli* ST131. *mBio*. 7(2):e00347-16. doi:10.1128/mBio.00347-16

35. Sheppard AE, Stoesser N, Wilson DJ, *et al*. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene bla_{KPC}. *Antimicrob Agents Chemother*. 2016;60(6):3767-3778. doi:10.1128/AAC.00464-16

36. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin Microbiol Infect*. 2018;24(4):350-354. doi:10.1016/j.cmi.2017.12.016

37. Earl D, Bradnam K, John JS, *et al*. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Res*. 2011;21(12):2224-2241. doi:10.1101/gr.126599.111

38. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genomics*. 2017;3(10):e000132. doi:10.1099/mgen.0.000132

39. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. *Brief Bioinform*. 2013;14(1):1-12. doi:10.1093/bib/bbs007

40. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev*. 2005;15(6):589-594. doi:10.1016/j.gde.2005.09.006

41. Falush D. Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol*. 2016;1(5):1-2. doi:10.1038/nmicrobiol.2016.59

42. Earle SG, Wu CH, Charlesworth J, *et al*. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*. 2016;1(5):1-8. doi:10.1038/nmicrobiol.2016.41

43. Lees JA, Mai TT, Galardini M, *et al*. Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *mBio*. 2020;11(4):e01344-20. doi:10.1128/mBio.01344-20

44. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genomics*. 2020;6(3):e000337. doi:10.1099/mgen.0.000337

Part I

Taxonomy and Antimicrobial Resistance of the *Escherichia* Genus

Chapter 2

Quantifying the contribution of four resistance mechanisms to ciprofloxacin minimum inhibitory concentration in *Escherichia coli*: a systematic review

Boas C.L. van der Putten, Daniel Remondini, Giovanni Pasquini, Victoria A. Janes, Sébastien Matamoros, Constance Schultsz

Journal of Antimicrobial Chemotherapy, Volume 74, Issue 2, February 2019, Pages 298–310, https://doi.org/10.1093/jac/dky417

Synopsis

Introduction

Reviews assessing the genetic basis of ciprofloxacin resistance in *Escherichia coli* have mostly been qualitative. However, to predict resistance phenotypes based on genotypic characteristics, it is essential to quantify the contribution of genotypic determinants to resistance. We performed a systematic review to assess the relative contribution of known genomic resistance determinants to the MIC of ciprofloxacin in *E. coli*.

Methods

PubMed and Web of Science were searched for English language studies that assessed ciprofloxacin MIC and presence or introduction of genetic determinants of ciprofloxacin resistance in *E. coli*. We included experimental and observational studies without time restrictions. Medians and ranges of MIC fold changes were calculated for individual resistance determinants and combinations thereof.

Results

We included 66 studies, describing 604 *E. coli* isolates that carried at least one genetic ciprofloxacin resistance determinant. Mutations in *gyrA* and *parC*, genes encoding targets of ciprofloxacin, contribute to the largest fold changes in ciprofloxacin resistance in *E. coli* compared to the wild type. Efflux, physical blocking or enzymatic modification, confer smaller increases in ciprofloxacin MIC than mutations in *gyrA* and *parC*. However, the presence of these other resistance mechanisms in addition to target alteration mutations further increases ciprofloxacin MIC, thus resulting in ciprofloxacin MIC fold increases ranging from 250 to 4000.

Conclusion

This quantitative review of genomic determinants of ciprofloxacin resistance in *E. coli* demonstrates the complexity of resistance phenotype prediction from genomic data and serves as a reference point for studies aiming to predict ciprofloxacin MIC from *E. coli* genomes.

Introduction

Escherichia coli is a Gram-negative bacterium able to adopt a commensal or pathogenic lifestyle in humans and animals[1]. Adding to the danger of pathogenic *E. coli* is the rise of antimicrobial resistance. *Escherichia coli* has acquired resistance to some of our most important antimicrobials, including aminopenicillins, cephalosporins, aminoglycosides, carbapenems and fluoroquinolones[2].

Ciprofloxacin is an antimicrobial of the fluoroquinolone class, commonly prescribed for a wide variety of infections including infections caused by *E. coli*[3]. As is the case for other fluoroquinolones, the substrate of ciprofloxacin is the complex formed by the DNA of the bacterium and either the DNA gyrase enzyme or the topoisomerase IV enzyme[4–6]. DNA gyrase creates single-stranded breaks in the DNA to negatively supercoil the DNA during replication or transcription[7]. If ciprofloxacin binds DNA gyrase in complex with DNA, the single stranded DNA breaks cannot be religated and thus accumulate, leading to double stranded DNA breaks[8]. A similar mechanism is hypothesized for topoisomerase IV[9].

The mechanisms of ciprofloxacin resistance in *E. coli* have been investigated intensively in the past 30 years. Mutations in genes coding for DNA gyrase and topoisomerase IV contribute to ciprofloxacin resistance in *E. coli*[10,11]. In addition, efflux pumps may decrease drug accumulation whilst peptides and enzymes may block drug targets or may modify the drug, respectively (Figure 1). Numerous reviews have covered the topic of ciprofloxacin resistance in *E. coli*, but these reviews have been overwhelmingly qualitative in nature[12–19].

With the rapidly increasing availability of next generation sequencing technologies, research aimed at the prediction of a resistance phenotype from genomic data is increasing. However, these efforts typically correlate genotypic data to a categorical measure of resistance, while a quantitative resistance phenotype prediction is of clinical relevance. Therefore, we carried out a systematic review, summarizing observational and experimental studies that assessed genetic ciprofloxacin resistance determinants and the ciprofloxacin MIC conferred by these determinants in *E. coli*, to elucidate how the presence of genomic resistance determinants, either alone or in combination, affects ciprofloxacin MIC in *E. coli*. In addition, we performed an *E. coli* protein network analysis to detect potential additional determinants of ciprofloxacin resistance on the basis of the findings of the systematic review.

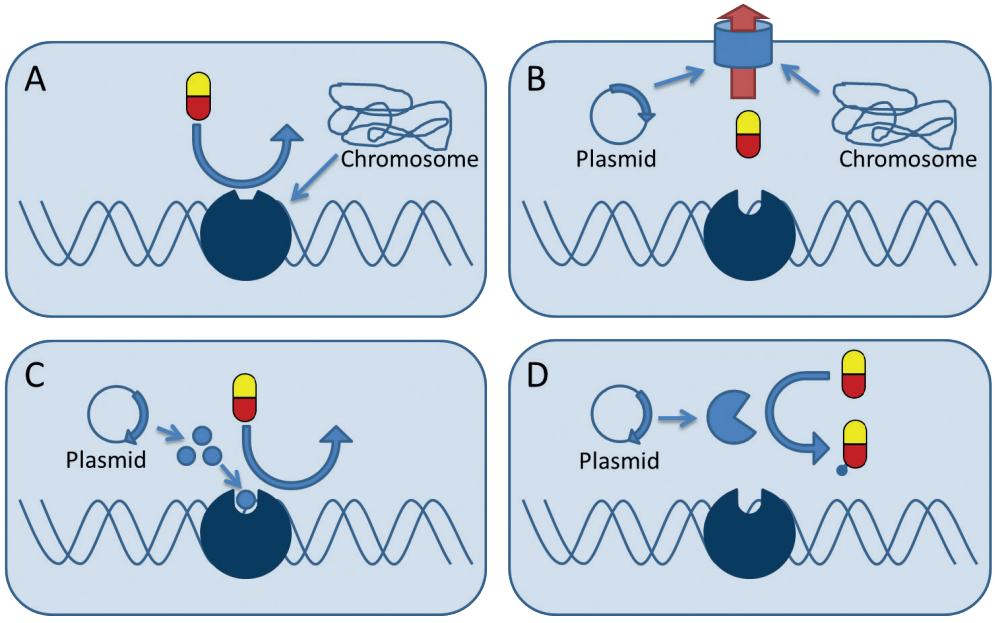


Figure 1. Schematic representation of four mechanisms of ciprofloxacin resistance in *E. coli*. A) Target alteration. B) Decreased ciprofloxacin accumulation. C) Physical blocking of ciprofloxacin target. D) Enzymatic modification of ciprofloxacin.

Methods

Systematic search

The PRISMA 2009 checklist was used as a guide for this systematic review[20]. PubMed and Web of Science were searched using a defined set of keywords, selecting original research articles in English language reporting on susceptibility test results of *Escherichia coli* isolates measured as MIC due to genetic modifications identified in clinical, carriage or environmental isolates (observational) or introduced in *E. coli* strains *in vitro* (experimental) (Supplementary methods). No time limits were applied. In addition to the defined search strategy, forward and backward citation searches of reviews and included articles was carried out. The final search was conducted on July 5th, 2018.

Inclusion and exclusion criteria for experimental and observational studies

Articles were not considered eligible for inclusion if they failed to mention any keyword (listed in the supplementary methods) describing ciprofloxacin resistance determinants in title or abstract. Eligible articles were screened by title, abstract and/or full text for inclusion based on the following inclusion and exclusion criteria (Figure 2). Studies could be included as experimental or as observational studies. For inclusion as an experimental study, the study needed to report a ciprofloxacin MIC before and after the introduction of

a genetic modification in a single *Escherichia coli* strain. Studies were eligible to be included as observational studies if the ciprofloxacin MIC of at least one *Escherichia coli* isolate was reported, together with the observed genetic determinants of ciprofloxacin resistance. *In vitro* evolution studies where *E. coli* were exposed to ciprofloxacin resulting in decreased susceptibility to ciprofloxacin, were considered observational studies, since mutations are not actively introduced in these studies. Observational studies were excluded if they failed to test for the presence of all of the following resistance determinants: mutations in Ser83 and Asp87 of *gyrA*, mutations in Ser80 and Glu84 of *parC*, mutations in *acrR* and *marR*, presence of *oqxAB*, *qepA*, *qnrA*, *qnrB*, *qnrS* and *aac(6')Ib-cr*. If studies failed to indicate unambiguously which resistance determinants were tested, the study was excluded.

Definitions

For this systematic review, the conventional definition of MIC was used, meaning the lowest concentration of ciprofloxacin that inhibits the visible growth of a bacterial culture during overnight incubation[21]. Clinical breakpoints (≤0.25 mg/L susceptible; 0.5 mg/L intermediately resistant, ≥1 mg/L resistant) and epidemiological cutoffs (0.064 mg/L) were used as defined by EUCAST[22,23].

A genomic resistance determinant was defined as a mutation in a gene or the presence of a plasmid-mediated gene that decreases ciprofloxacin susceptibility. Since currently four mechanisms of ciprofloxacin resistance in *E. coli* are known, an isolate can possess multiple resistance determinants encoding for multiple resistance mechanisms. In addition, a single resistance mechanism can be encoded by multiple resistance determinants.

Genetic modifications were defined as an experimentally acquired mutation, insertion or deletion of a nucleotide or a sequence of nucleotides in the chromosome. The introduction of plasmid-mediated genes was also considered a genetic modification. Dominance tests as described by Heisig *et al*. were considered experimental evidence[24]. In short, a dominance test relies on increasing the susceptibility of a bacterium to an antimicrobial, by introducing a plasmid containing the wild type gene that codes for the antimicrobial's target. In the studies included in this report, the MICs of bacteria with mutations in *gyrA* or *parC* were lowered by introducing a plasmid containing wild type *gyrA* or wild type *parC*.

Data extraction and analysis

The management of the literature search was performed using Pubreminer (http://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi).

All data on genetic modifications were extracted from the articles or supplementary material, together with MIC data. For experimental data, the MICs of the isolates before

and after a targeted genetic modification were extracted to calculate a fold change of ciprofloxacin MIC for each of the *E. coli* isolates.

We calculated how frequently resistance determinants were tested in the experimental data. This frequency is expressed as the number of isolates in which the genetic modification was introduced, divided by the total number of isolates included from experimental studies. The frequency can be used to estimate the strength of evidence per resistance determinant (Table S1). Furthermore, the sample sources, country of origin and isolation date of included *E. coli* isolates were extracted from the observational studies. The MIC fold change data plot and the correlation matrix were generated using the ggplot2 package RStudio version 1.1.383, running R version 3.4.2. Pearson correlation coefficients were calculated using the stats package and prepared for plotting using the reshape2 package.

Network construction

To investigate interactions between resistance determinants and to search for potential resistance determinants, a protein-protein interaction network was constructed. The *Escherichia coli* K-12 MG1655 interactome was extracted from the STRING-v10 database[25]. String-v10 aims to be more complete in terms of coverage of proteins for each organism in comparison to the other meta-interactomes available[26,27]. The functional association is the basic interaction unit of String in order to link proteins with a functional relation that are likely to contribute to a common biological purpose. Each interaction is derived from multiple sources, and we identify three groups of interactions (Table S2): PI interactions (where at least one physical protein interaction has been tested, imported from primary databases), FP interactions (determined by at least one functional prediction of an algorithm employed by String, genomic information, pathway knowledge, orthology relations) and TM interactions (supported only by automated text-mining of MedLine abstracts and full-text articles). Based on the sources, for each interaction in String a score is calculated, ranging from 0 to 1. In our analysis, only interactions with a score higher than 0.7 were retained (defined as high quality interactions by String), resulting in 3,890 nodes and 32,854 edges (with only 0.06% of the links supported only by TM interactions). Genes resulted by the systematic search were mapped to the EcoGene-3.0 database to obtain *E. coli* K-12 MG1655 identifiers (bnumber)[28], that were subsequently mapped to the MG1655 interactome.

Results

Systematic search

The systematic search yielded 5055 PubMed entries and 5873 Web of Science entries. After removal of duplicates, 1718 unique articles were screened on content by title, abstract and, if necessary, full text. This approach identified 50 articles that were included

as experimental studies. Additionally, 10 experimental studies were identified through backward/forward searches in citations of included articles and known reviews. Three articles fulfilled inclusion criteria for observational studies, of which two articles were also included as experimental studies because they provided experimental data as well (Figure 2).

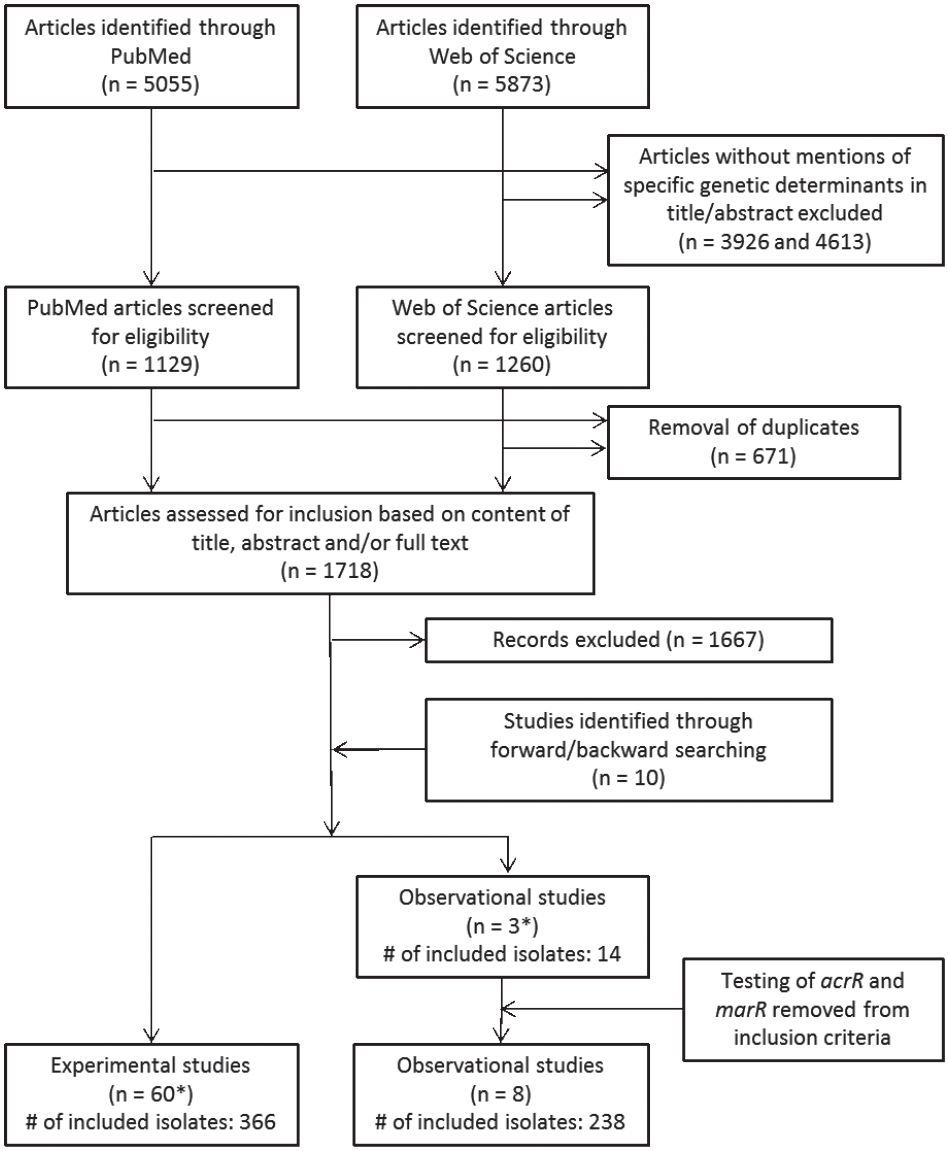


Figure 2. Flow chart adapted from the PRISMA guidelines (Moher 2009), showing the process of including articles starting from a systematic search of PubMed and Web of Science. *2 Studies contributed experimental and observational data, and were thus included for both types of articles.

The number of *E. coli* isolates which were confirmed to harbour at least one resistance determinant and for which MICs were reported, amounted to a total of 366 isolates from experimental studies (Table S1) and 238 isolates from observational studies (Table S3). A total of 43 different genomic determinants were described in the collected experimental data, of which 21 were shown to have an effect on ciprofloxacin MIC (Table 1).

Table 1. Ciprofloxacin resistance mechanisms in *Escherichia* coli and genes involved in these mechanisms. Note that in this overview, only genes are displayed that were shown to have any effect on ciprofloxacin susceptibility when mutations are present (chromosomal genes) or if the resistance gene is present (plasmid-encoded genes).

Resistance mechanism	Chromosomal genes involved in ciprofloxacin resistance	Plasmid-encoded genes involved in ciprofloxacin resistance
Target alteration	*gyrA*[12], *gyrB*[29], *parC*[11]	-
Decreased ciprofloxacin accumulation	*marR*[30], *acrRAB*[31], *tolC*[31], *soxS*[32], *rpoB*[33]	*qepA*[34], *oqxAB*[35]
Physical blocking of ciprofloxacin target	-	*qnrA*[36], *qnrB*[37], *qnrC*[38], *qnrD*[39], *qnrE*[40], *qnrS*[41]
Enzymatic modification of ciprofloxacin	-	*aac(6')-Ib-cr*[42] *crpP*[43]

Experimental studies focused primarily on mutations in Ser83 (28% of included isolates) and Asp87 (18%) of *gyrA*, S80 (15%) of *parC* and mutations in *marR* (20%). Of all plasmid-mediated resistance genes, *qnrA* (17%), *qnrS* (12%) and *aac(6')Ib-cr* (13%) were described most often. The other resistance determinants were tested in less than 10% of the experimentally modified isolates.

Target alteration mutations in *gyrA*, *gyrB*, *parC* and *parE*

Mutations in *gyrA* were the first ciprofloxacin resistance determinants to be discovered[12]. Mutations in *parC*, *gyrB* and *parE* were later also proven or implied to decrease ciprofloxacin susceptibility[11,29,44]. *gyrA* and *parC* mutations that reduce ciprofloxacin susceptibility cluster in regions termed the quinolone resistance-determining regions (QRDRs). Generally, the QRDR of *gyrA* ranges from amino acid Ala67 to Gln106[45], and the QRDR of parC from Ala64 to Gln103[11]. *gyrA* and *parC* mutations accumulate stepwise in *E. coli* when exposed to ciprofloxacin, increasing ciprofloxacin MIC concurrently[11,46–48]. The most common initial mutation is Ser83Leu in *gyrA*[46–48]. In the collected experimental data, this mutation confers a median fold increase in MIC of 24 compared to the wild type (range: 4-133x fold increase) [11,49–55]. This mutation is most often followed by Ser80Ile in *parC*[11,46,48] and finally by Asp87Asn or Asp87Gly in *gyrA*[46–48]. As mutations in *gyrA* and *parC* accumulate, ciprofloxacin MIC increases steeply. The ciprofloxacin MIC fold increase for a mutant of Ser83Leu (*gyrA*) and Ser80Ile (*parC*) is 62.5[51]. A similar double mutant of Ser83Leu (*gyrA*) and Ser80Arg (*parC*) showed a ciprofloxacin MIC fold increase of 125[53]. For a triple mutant of Ser83Leu,

Asp87Asn (*gyrA*) and Ser80Ile (*parC*) the median ciprofloxacin MIC fold increase is 2000[11,51,54]. A quadruple mutant of Ser83Leu, Asp87Asn (*gyrA*) and Ser80Ile, Glu84Lys (*parC*) has been tested, but this mutant did not show a higher ciprofloxacin MIC than triple mutants within the same study[11]. In addition, Gly81Asp and Asp82Gly mutations in *gyrA* have been tested. These mutations caused low to no decrease in ciprofloxacin susceptibility (MIC fold changes: 2.6x and 1x, respectively, Table 2) [49,56].

Only one *gyrB* mutation (Asp426Asn) was shown to slightly increase ciprofloxacin resistance (Table 2) [29]. We did not find studies that showed a decreased ciprofloxacin susceptibility due to mutations in *parE*. However, a Leu445His mutation in *parE* of *E. coli* caused a 2x fold increase in the MIC of norfloxacin, another fluoroquinolone[44].

Efflux pump genes (*acrAB*, *tolC*) and their transcriptional regulators (*marR*, *acrR* and *soxS*)

As with many other antimicrobials, bacterial efflux pumps also play a role in resistance against ciprofloxacin. Deletion of *acrAB* or *tolC* confers a clear increase in the ciprofloxacin susceptibility of *E. coli* (4-8 fold decrease in MIC)[30,31,57]. Deletions of 14 other genes or operons coding for efflux pumps in *E. coli* did not affect the ciprofloxacin MIC[31]. The deletion of transcriptional repressors of expression of efflux pumps like *marR* and *acrR* has been shown to affect ciprofloxacin MIC. The only study in our collected experimental data to investigate deletion of *acrR* showed that the MIC tripled after the repressor was deleted[51]. Nine studies investigated the effects of *marR* deletion or mutation, which reported a median fold increase in ciprofloxacin MIC of 4 (range 1.5-218x fold increase) [30,51,52,54,58–60]. A recent study by Pietsch *et al*. detected mutations in *rpoB* in an *in vitro* evolution experiment[33]. These mutations arose after accumulation of other mutations, and were shown to increase the ciprofloxacin MIC of a wild type *E. coli* by 1.5-3 fold change (Table 2). The mutations in *rpoB* were shown to increase ciprofloxacin MIC by upregulating the expression of *mdtK* (also known as *ydhE*).

Two experimental studies reported mutations in efflux pump operons, influencing ciprofloxacin MIC. The first mutation was Ala12Ser in *soxS*, leading to higher expression of *acrB*, in turn leading to a 4-fold increase in ciprofloxacin MIC[32]. The second mutation was a Gly288Asp mutation in *acrB* itself, conferring a 16.7 fold increase in ciprofloxacin MIC (Table 2) [61]. This *acrB* mutation however increased susceptibility to other antimicrobials.

Plasmid-encoded efflux pump genes *oqxAB* and *qepA*

In addition to chromosomally-encoded efflux pumps, the presence of plasmid-encoded efflux pump genes *oqxAB* and *qepA* has been shown to increase ciprofloxacin MIC in *E. coli*[34,35]. *oqxAB* confers a median fold increase in MIC of 7.5 (range 2-16x fold increase)[35,62–64], while *qepA* confers a median fold increase of 4.5 (range 2-31x fold increase, Table 2)[34,52,65–68].

qnr genes

qnrA was the first plasmid-mediated quinolone resistance (PMQR) determinant to be discovered[36]. Qnr proteins are pentapeptide repeat proteins that decrease binding of fluoroquinolones to DNA gyrase by binding the DNA:DNA gyrase complex[69]. Since 2002, many more *qnr* alleles have been discovered. Currently seven families of *qnr* genes are recognized: *qnrA*, *qnrB*, *qnrC*, *qnrD*, *qnrE*, *qnrS* and *qnrVC*[70]. In the collected experimental data, all *qnr* families have been tested for their influence on ciprofloxacin MIC of *E. coli*, except for *qnrVC*. *qnr* genes confer ciprofloxacin MIC fold increases between 4 and 125. The median ciprofloxacin MIC fold increase differed per *qnr* allele (Table 2).

aac(6')Ib-cr and *crpP*

A plasmid mediated mutant *aac(6')Ib* gene that decreased fluoroquinolone susceptibility in *E. coli* was discovered in 2006[42]. Until then, *aac(6')Ib* genes were only known to decrease *E. coli* susceptibility to aminoglycosides. A double mutation in the acetyltransferase-encoding gene enabled the resulting protein to acetylate both aminoglycosides and some fluoroquinolones, including ciprofloxacin. This novel variant, *aac(6')Ib-cr*, was shown to confer a median fold increase in ciprofloxacin MIC of 6.9 (range: 1-62.5x fold increase, Table 2)[52,71–76].

The most recently discovered ciprofloxacin resistance determinant in *E. coli* is *crpP*, a plasmid-mediated gene coding for a protein with the putative ability to phosphorylate certain fluoroquinolones such as ciprofloxacin[43]. *crpP* was first detected in a clinical isolate of *Pseudomonas aeruginosa*, but was shown to confer a 7.5 fold-change increase in ciprofloxacin MIC when conjugated to *E. coli* J53.

Table 2. Medians and ranges of ciprofloxacin MIC fold changes stratified by resistance determinants. Only data from isolates harbouring resistance determinants from a single mechanism are shown.

Resistance determinant	Median ciprofloxacin MIC fold change (range)	# of isolates	References
Gly81Asp (*gyrA*)	2.6 (1-4.2)	2	49,56
Asp82Gly (*gyrA*)	1	1	49
Ser83Trp (*gyrA*)	6.3	1	10
Ser83Leu (*gyrA*)	23.8 (4-133.3)	9	11,49–51,53–55
Asp87Asn (*gyrA*)	15.6 (7.5-15.6)	3	51,54,55
Gly81Asp, Asp82Gly (*gyrA*)	2	1	49
Ser83Leu, Asp87Asn (*gyrA*)	23.8 (15-23.8)	3	51,54,59
Ser83Leu, Asp87Gly (*gyrA*)	4266.7	1	77
Asp426Asn (*gyrB*)	8	1	29
Ser80Ile (*parC*)	1	1	51
Ser83Trp (*gyrA*), Gly78Asp (*parC*)	33.3	1	11

Ser83Leu (*gyrA*), Ser80Ile (*parC*)	62.55	1	51
Ser83Leu (*gyrA*), Ser80Arg (*parC*)	125	1	53
Asp87Asn (*gyrA*), Ser80Ile (*parC*)	23.8	1	51
Ser83Leu, Asp87Asn (*gyrA*), Ser80Ile (*parC*)	2000 (1066.7-2000)	3	11,51,54
Ser83Leu, Asp87Gly (*gyrA*), Ser80Ile (*parC*)	1024 (256-8533.3)	3	11
Ser83Leu, Asp87Asn (*gyrA*), Ser80Arg (*parC*)	2258.3 (250-4266.7)	2	11,59
Ser83Leu, D87Y (*gyrA*), Ser80Ile (*parC*)	256	1	11
Ser83Leu, Asp87Asn (*gyrA*), Glu84Lys (*parC*)	533.3	1	11
Ser83Leu, Asp87Gly (*gyrA*), Glu84Lys (*parC*)	4266.7	1	11
Ser83Leu, Asp87Asn (*gyrA*), Ser80Ile, Glu84Gly (*parC*)	1600 (1066.7-2133.3)	2	11
acrB: Gly228Asp	16.7	1	61
Δ*acrAB*	0.1 (0-0.3)	10	30,31,57
Δ*tolC*	0.3	1	31
marR (various mutations)	3.5 (1.5-4)	14	60
Δ*marR*	3.8 (2-218)	5	30,51,54,58,59
acrR (various mutations)	4 (2-16)	6	78
Δ*acrR*	2.9	1	51
soxS: Ala12Ser	4	1	32
rpoB (various mutations)	3 (1.5-3)	3	33
oqxAB	7.5 (2-16)	17	35,62–64
qepA	8.3 (1.9-64)	13	34,52,65–68,79
qepA, Δ*marR*	15	1	67
qnrA (unspecified allele)	31.3 (20.8-31.7)	12	80
qnrA1	31 (4-66.7)	37	39,50,52,53,81–89
qnrA3	31.3	1	81
qnrB1	12.5 (4-62.5)	8	52,53,85,87
qnrB2	15.6 (11.8-31.3)	4	81,90
qnrB4	15.6 (15.6-15.6)	3	91
qnrB5	15.6 (15.6-15.6)	2	72
qnrB6	15.6	1	72
qnrB19	11.9	1	82
qnrC1	31.3 (15-62.5)	3	59,38,85
qnrD1	15 (7.5-62.5)	3	59,39,85
qnrE1	62.5	1	40
qnrS (unspecified allele)	12.3 (2-83.3)	6	74,76
qnrS1	33.3 (4-125)	24	39,50,52,53,63,79,81, 82,85,87,90,92–94
qnrS2	15	1	95
aac(6')Ib-cr	6.9 (1-62.5)	28	52,42,71,73–76,79,94,96
crpP	7.5	1	43

Effect of multiple modifications on MIC

The fold change in MIC of each included experimental isolate was plotted, stratified for the resistance mechanism present (Figure 3). Target alteration resulted in the largest range of MIC fold changes which were on average higher than the fold changes observed as a result of the three other mechanisms. Whilst the presence of determinants representing different ciprofloxacin resistance mechanisms may result in a moderate fold change in MIC, the accumulation of multiple resistance determinants encoding multiple mechanisms of resistance is likely to increase the ciprofloxacin MIC significantly.

Data on plasmid-mediated resistance genes comes from either resistance genes cloned into typical lab vectors (such as pUC18), or the genes can be tested in their "native" plasmid. Generally, lab plasmids have higher copy numbers which could bias results when resistance genes cloned into lab plasmids are compared with resistance genes in their native plasmids. To assess this possible difference, we extracted information about the plasmids used in all experimental isolates, in which the effect one single plasmid-mediated resistance gene was tested. For three mechanisms (efflux, physical blocking and enzymatic modification), we compared the ciprofloxacin MIC fold change conferred by the resistance genes compared between native (n = 97) and cloned (n = 70) plasmids (Figure S1). This analysis indicated resistance genes cloned into lab plasmids did not confer higher MIC fold changes than resistance genes in native plasmids, when stratified per mechanism.

Comparison of experimental and observational data

We compared the findings from the experimental data with susceptibility test results and associated presence of mutations reported for isolates in observational studies. Because studies were excluded if isolates were not tested for the presence of all known resistance encoding determinants, only studies could be included that were published after *oqxAB* was linked to increased ciprofloxacin MIC in 2007[35]. The description of *crpP* was only recently published and was therefore not used as an inclusion criterion. Only three observational studies reported on the presence of all currently known resistance determinants[33,97,98]. Since mutations in both *acrR* and *marR* genes were shown to result in no to low fold changes in ciprofloxacin MIC, we added five observational studies that fulfilled all inclusion and exclusion criteria except testing for the presence of mutations in *acrR* and *marR* genes, in a secondary analysis. Thus, eight observational studies published between 2012 and 2018 were included, contributing data on a total of 238 strains (Table S3). The studies reported data on 1 to 92 isolates, with a median of 13.5 isolates per study. Ciprofloxacin MICs of included isolates ranged from 0.015 to 1024 mg/L with a median MIC of 1 mg/L.

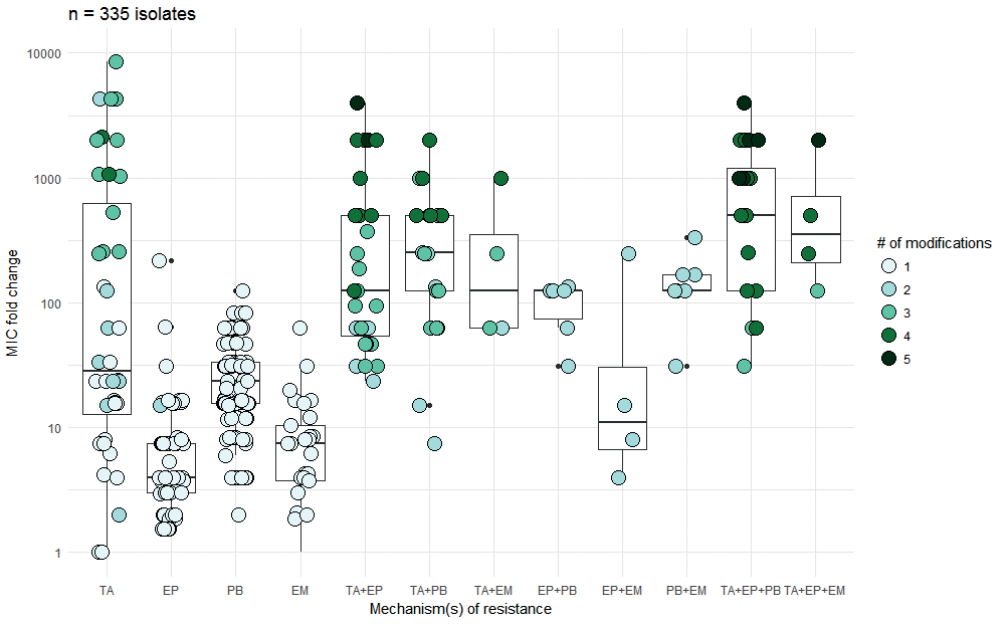


Figure 3. Median fold change (interquartile range) in ciprofloxacin MIC for each resistance mechanism or combination of resistance mechanisms experimentally tested in 335 isolates. Fold changes were calculated by dividing the MIC after modification by the MIC before modification for each isolate. Data points represent single *E. coli* isolates. Darker fill of data points indicates the presence of multiple resistance mutations or resistance genes in the isolate. Isolates that showed a decreased ciprofloxacin MIC after modification (such as deletion of *acrAB* or *tolC*) are not shown but are listed in table S1 (n = 31) [30,31,57]. TA = target alteration (mutations in *gyrA*, *gyrB* or *parC*), EP = efflux pump (mutations in *acrB*, *marR*, *acrR*, *rpoB* or presence of *qepA* or *oqxAB*), PB = physical blocking (presence of *qnrA*, *qnrB*, *qnrC*, *qnrD*, *qnrE* or *qnrS*), EM = enzymatic modification (presence of *aac(6')Ib-cr* or *crpP*).

We analysed MIC distributions for combinations of resistance determinants that were reported at least five times in the experimental and observational data. These combinations of resistance determinants included the mutation Ser83Leu in *gyrA*, presence of *qnrS1* and presence of *aac(6')Ib-cr*. Although for most combinations of resistance determinants small numbers of isolates were reported, results of experimental and observational data appear comparable with the exception for the reported MICs for *E. coli* strains solely harbouring *aac(6')Ib-cr* (Table 3).

Table 3. Median ciprofloxacin MICs for three resistance determinants that were reported at least five times in both experimental and observational data. The EUCAST epidemiological cut-off for ciprofloxacin resistance in *E. coli* is 0.064 mg/L.

Resistance determinant(s)	Median and range of ciprofloxacin MIC in experimental data (mg/L)	Number of isolates in experimental data	Median and range of ciprofloxacin MIC in observational data (mg/L)	Number of isolates in observational data
Ser83Leu (*gyrA*)	0.25 (0.06-0.38)	5	0.25 (0.125-64)	34
qnrS1	0.25 (0.032-1)	16	0.2 (0.1-4)	19
aac(6')Ib-cr	0.06 (0.004-0.5)	22	0.25 (0.25-0.5)	5

We also examined if certain combinations of resistance mechanisms were more prevalent than others in the observational data. Calculating Pearson correlation coefficients between commonly observed resistance determinants showed that *gyrA* (Ser83, Asp87) and *parC* (Ser80) mutations were positively correlated with each other. Additionally, these three mutations were shown to inversely correlate with the presence of *qnrB* and *qnrS* genes in our observational data. This inverse correlation was not observed with other frequently reported plasmid-mediated resistance determinants such as *aac(6')Ib-cr* (Figure 4).

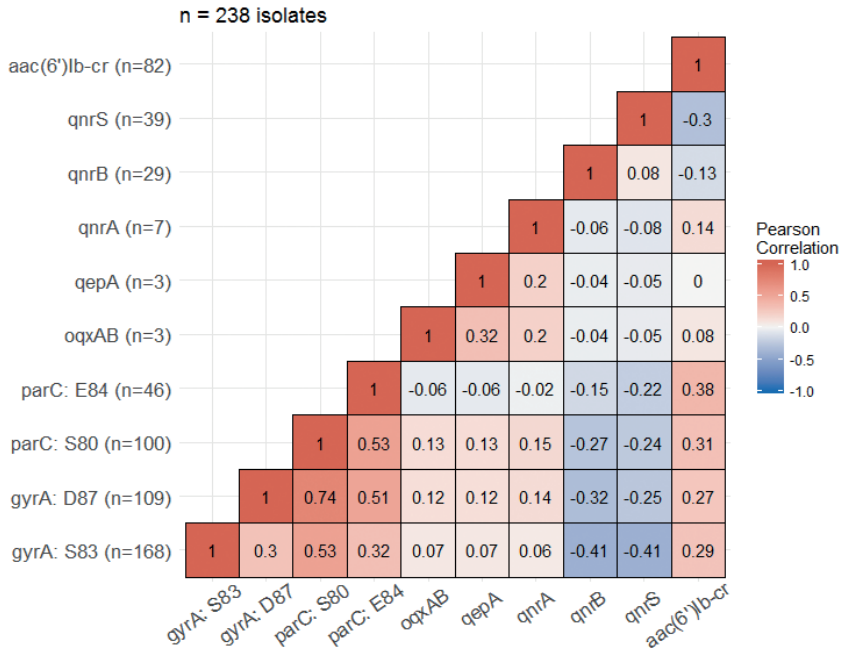


Figure 4. Matrix displaying Pearson correlation coefficients calculated between resistance determinants in a pairwise manner. All 238 strains used for this analysis were screened for all displayed resistance determinants. The reported frequencies of resistance determinants in our dataset are displayed on the y-axis. Full data is provided in table S3.

Network visualization

In order to get a global picture of the mutation landscape associated with ciprofloxacin resistance, we mapped the selected chromosomal genes onto a Protein-Protein Interaction (PPI) network. The selected genes were evaluated in a wide range of *E. coli* strains, and we mapped them to the String-v10 database referring to the *E. coli* K-12 MG1655 model organism, since it showed the highest number of matching edges and nodes among the strains available in String database. We noted that plasmid-associated genes like *oqxAB* and the *qnr* gene family were not described by interactomes in general, since interactomes mostly describe the core genome. Moreover, some genes (such as *yohG*) could not be mapped because they are not present in *E. coli* K-12 MG1655.

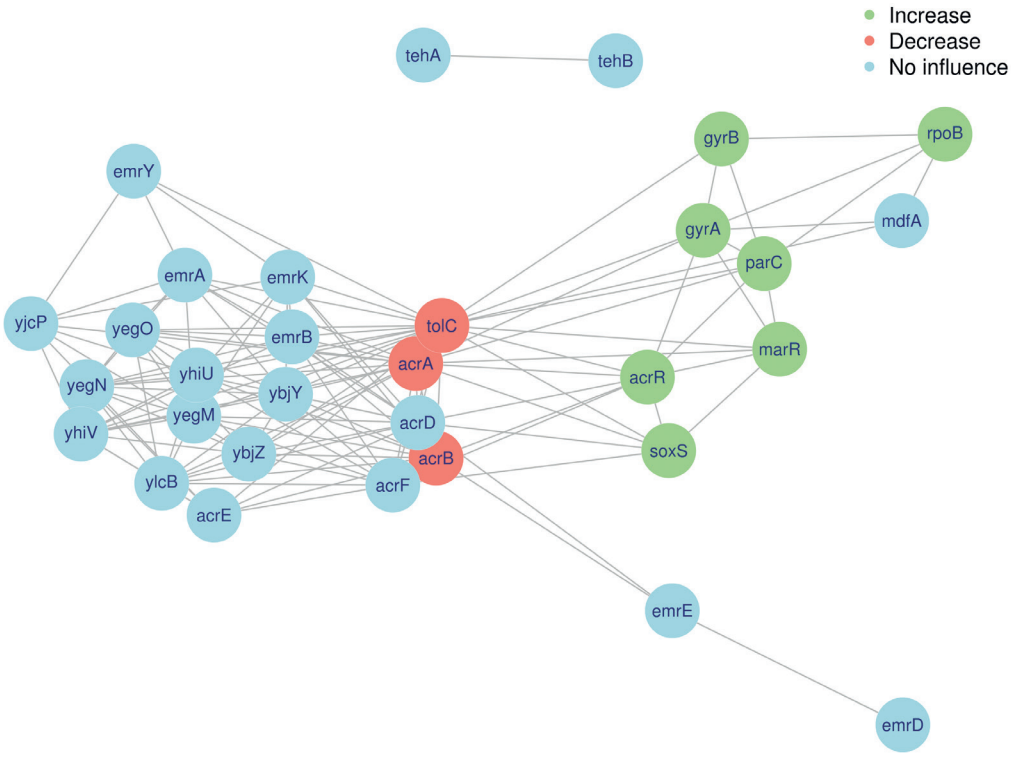


Figure 5. Network of *E. coli* ciprofloxacin resistance-associated chromosomal genes. 31 genes that were examined for their influence on ciprofloxacin and were present in the *E. coli* K-12 MG1655 genome were mapped to the String-v10 PPI database. Genes were coloured green if a mutation conferring increased ciprofloxacin resistance was observed; genes were coloured red when a mutation decreased ciprofloxacin resistance; genes were coloured blue when a mutation showed no effect on ciprofloxacin resistance. The network is displayed by R package iGraph employing the force-directed layout algorithm by Fruchterman and Reingold. The list of edges with corresponding data categories (PI, FP or TM) is available as supplementary table 3.

Of the 43 selected genes, 31 (72%) mapped to the PPI network, resulting in a fully connected sub-module. The network highlighted the close relationship between gene connectivity and ciprofloxacin resistance effects: the chosen visualization algorithm showed that genes with similar effects tightly grouped in the interactome (Figure 5). Particularly, the genes that had an increasing effect on ciprofloxacin resistance when mutated seemed to cluster, even if the genes belonged to different resistance mechanisms. As expected, close relationships between particular sets of genes were revealed. Transcriptional regulators such as *marR*, *acrR* and *soxS* were shown to interact with efflux pump genes such as *acrA*, *acrB*, *acrD*, *acrF* and *tolC*. Also, the physical interactions between *gyrA*, *gyrB* and *parC* were depicted in the network.

Discussion

This report provides a comprehensive and systematic analysis of 66 papers linking genotype of *E. coli* to a quantitative ciprofloxacin resistance phenotype, spanning the years 1989-2018 and amounting to a total of 604 isolates. Ciprofloxacin MIC in *E. coli* is largely affected by target mutations in specific residues in *gyrA* (Ser83 and Asp87) and *parC* (Ser80), conferring median MIC fold increases ranging from 24 for single Ser83Leu (*gyrA*) mutants to 1533 for triple Ser83Leu, Asp87Asn/Gly (*gyrA*) Ser80Ile/Arg (*parC*) mutants. However, accumulation of multiple resistance determinants, including those representing other resistance mechanisms, can increase ciprofloxacin MIC even further, up to MIC fold increases of 4000.

Beside the MIC fold changes that are conferred by resistance determinants, it is important to consider how these genetic resistance determinants are acquired. Various pathways leading to mutagenesis are known in *E. coli*. First and foremost, spontaneous mutagenesis leads to the accumulation of mutations at a rate of about 2.5×10^{-3} per genome per replication, although more recent studies observed mutation rates as low as 1×10^{-3} per genome per replication[99,100]. Additionally, a subset of *E. coli* isolates display a higher mutation rate because of *mutS* or *mutL* mutations[100]. Finally, specific mechanisms exist that induce mutagenesis. One such mechanism is the SOS response which is induced after DNA damage inflicted by exogenous substances, including quinolones such as ciprofloxacin[101]. Two proteins that are central in the SOS response are LexA and RecA. In the absence of DNA damage, LexA dimers are bound to a SOS box (promoter region of SOS genes) and inhibit expression of SOS genes. If DNA damage is induced, for example through the presence of ciprofloxacin, RecA will bind ssDNA that is a result of the DNA damage. The activated RecA in turn mediates the self-cleavage of LexA, derepressing the SOS box, finally leading to expression of SOS genes and thus the SOS response. This SOS response induces mutations, among others, through DNA damage repair performed by error-prone DNA polymerases[102].

Currently, four ways are known in which the SOS response affects ciprofloxacin resistance in *E. coli*. First, the SOS response induces a higher mutation rate, making it more likely that ciprofloxacin resistance mutations will arise within a fixed population[103]. Additionally, if the SOS response is knocked out in *E. coli*, ciprofloxacin MIC decreases. Clinically resistant *E. coli* that had *recA* knocked out showed MIC fold decreases of 4-8[103]. Furthermore, the SOS response has been shown to induce expression of some *qnr* gene families, for example *qnrB* and *qnrD*[104,105]. Finally, the SOS response has been shown to promote horizontal gene transfer of an integrative conjugative element (ICE) SXT in *Vibrio cholerae* and recombinant *E. coli* harbouring genes encoding for resistance to chloramphenicol, sulphonamides, streptomycin and trimethoprim in the presence of ciprofloxacin[106]. After mutagenesis through mechanisms such as the SOS response, the fitness of the mutant indicates how likely the bacterium is to survive. In absence of ciprofloxacin, *gyrA* mutations and *parC* mutations have been shown to confer limited fitness costs compared to other resistance determinants[48,51,59,67,75]. Additionally, mutations in *gyrA* and *parC* show positive epistasis, as the MIC fold change of the triple Ser83Leu, Asp87Asn (*gyrA*) and Ser80Ile (*parC*) mutant is higher (2000x fold increase) than would be expected based on the MIC fold changes conferred by the individual mutations (24x, 16x and 1x fold increases, respectively) [51,107]. This epistatic effect thus raises ciprofloxacin MIC very efficiently. This, in combination with the low fitness costs in absence of ciprofloxacin might explain why ciprofloxacin resistance mutations in *gyrA* and *parC* are the most common ciprofloxacin resistance determinants observed in *E. coli*.

Notably, other combinations of resistance determinants also show positive epistatic effects, although the observed effects are weaker. A similar positive epistatic effect was observed for chromosomal *gyrA*/*parC* mutations together with plasmid-mediated resistance determinants *qepA*[67] and *aac(6')Ib-cr*[52,75]. However, experimental studies of combinations of *gyrA* and *parC* mutations with *qnr* genes showed discordant results. One study reported a negative epistatic effect on ciprofloxacin MIC of target alteration mutations with all *qnr* genes tested (*qnrA*, *qnrB*, *qnrC*, *qnrD*, *qnrS*)[59], and another study observed a similar effect of target alteration mutations with *qnrB*, but the opposite effect for target alteration mutations with *qnrS* in terms of conferred MIC[52].

The complex relation between *gyrA*/*parC* mutations and *qnr* genes is further illustrated by our findings from the observational data. We observed a clear negative correlation between presence of *gyrA* or *parC* mutations and presence of *qnrB* and *qnrS* genes. This finding is in line with an earlier study that reported an *E. coli* population fixating *gyrA*/*parC* mutations at a reduced rate when the *E. coli* population harboured a *qnr* gene as opposed to when the *E. coli* strain did not harbour a *qnr* gene[81]. However, no additional fitness costs are usually reported for *E. coli* harbouring both *gyrA*/*parC* mutations and *qnr* genes[59]. One possible explanation was suggested by the study of Garoff *et al*., who

reported an enhanced fitness cost conferred by *qnr* genes when Lon protease was absent from an *E. coli* genome[108]. This finding shows that the fitness cost conferred by an antimicrobial resistance gene to an *E. coli* strain can be influenced by genes that do not directly play a role in antimicrobial resistance.

By mapping the selected genes onto a known *E. coli* interactome, we found a clear association between their role in ciprofloxacin resistance and their position in the network, with a significant proximity of genes that produce a similar response in terms of resistance (i.e. increase or decrease). This global picture highlights the presence of common biological functions (mostly associated with the efflux pumps and their regulation), and it suggests that system biology approaches in the future will likely be helpful to identify new targets or specific pathways related to ciprofloxacin resistance or antimicrobial resistance in general. As an example, the position in the network of *acrD* and *acrF* genes, which were not identified as resistance-associated genes in the experiments reported so far, and their biological function as efflux pump protein complexes, suggest that their role in resistance should be more deeply investigated.

Despite its comprehensiveness our study has certain limitations. First, gene expression data are not included in this review because our study aims at prediction of MIC on the basis of a DNA sequence. It has been shown that increased expression of efflux pumps such as *acrAB* or transcriptional regulators of efflux pumps such as *marA* is significantly correlated with increased fluoroquinolone MIC in *E. coli*[109,110]. Secondly, complex combinations of resistance determinants such as combinations of *gyrA*/*parC* mutations with plasmid-mediated resistance determinants have been reported sparsely in the experimental data. Therefore, the comparison of experimental and observational data for these combinations of resistance determinants is impossible using this dataset. Finally, only currently known ciprofloxacin resistance determinants could be included in this report. The very recent discovery of *crpP* suggests that more resistance determinants or resistance mechanisms are still waiting to be discovered[43]. Additionally, complex mutation patterns influencing ciprofloxacin resistance through unknown pathways may exist, but current research methods do not usually detect these kinds of effects.

One possible solution for the issues described above would be the use of advanced machine learning algorithms to predict ciprofloxacin resistance. These algorithms should be able to associate large quantities of sequence data with phenotypic metadata in an unbiased manner. One such attempt has been made for ciprofloxacin resistance already[111]. It was reported that Ser83Phe, Ser83Thr (*gyrA*), Ser80Arg (*parC*) and presence of any *qnr* gene were the most important resistance determinants according to the algorithm used. However, this study used categorical (susceptible or resistant) and not quantitative phenotype data, and included various Enterobacteriaceae species and the

results can thus not be directly compared with the data presented here for *E. coli* alone. This is exemplified by the fact that neither Ser83Phe nor Ser83Thr (*gyrA*) were reported in our observational data. For future studies, the data collected for this review could serve as a benchmark, as this review presents a comprehensive set of quantitative data on the contribution of various resistance determinants to ciprofloxacin MIC in *E. coli*.

Acknowledgments

We wish to thank the COMPARE consortium for support and helpful discussions.

Funding

This work was supported by the COMPARE Consortium, which has received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement No. 643476), and through internal funding.

Transparency

None to declare.

References

1. Tenaillon O, Skurnik D, Picard B *et al.* The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 2010; 8: 207–17.
2. ECDC. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). 2014. https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/antimicrobial-resistance-europe-2014.pdf.
3. Johns Hopkins Medicine 2016. Antibiotic Guidelines 2015-2016. https://www.medbox.org/antibiotic-guidelines-2015-2016/download.pdf.
4. LeBel M. Ciprofloxacin: chemistry, mechanism of action, resistance, antimicrobial spectrum, pharmacokinetics, clinical trials, and adverse reactions. *Pharmacother J Hum Pharmacol Drug Ther* 1988; 8: 3–30.
5. Khodursky A, Zechiedrich E, Cozzarelli N. Topoisomerase IV is a target of quinolones in *Escherichia coli*. *Proc Natl Acad Sci USA* 1995; 92: 11801–5.
6. Drlica K. Mechanism of fluoroquinolone action. *Curr Opin Microbiol* 1999; 2: 504–8.
7. Cozzarelli NR. DNA gyrase and the supercoiling of DNA. *Science* 1980; 207: 953–60.
8. Kampranis SC, Maxwell A. The DNA gyrase-quinolone complex. *J Biol Chem* 1998; 273: 22615–26.
9. Anderson VE, Gootz TD, Osheroff N. Topoisomerase IV catalysis and the mechanism of quinolone action. *J Biol Chem* 1998; 273: 17879–85.
10. Cullen ME, Wyke AW, Kuroda R *et al.* Cloning and characterization of a DNA gyrase A gene from *Escherichia coli* that confers clinical resistance to 4-quinolones. *Antimicrob Agents Chemother* 1989; 33: 886–94.
11. Heisig P. Genetic evidence for a role of *parC* mutations in development of high-level fluoroquinolone resistance in *Escherichia coli*. 1996; 40: 879–85.
12. Hooper DC, Wolfson JS, Ng EY *et al.* Mechanisms of action of and resistance to ciprofloxacin. *Am J Med* 1987; 82: 12–20.
13. Hooper DC. Mechanisms of action and resistance of older and newer fluoroquinolones. *Clin Infect Dis* 2000; 31: S24–8.
14. Hooper DC. Emerging mechanisms of fluoroquinolone resistance. *Emerging Infect Dis* 2001; 7: 337–41.
15. Hooper DC. Mechanisms of fluoroquinolone resistance. *Drug Resist Updat* 1999; 2: 38–55.
16. Webber M, Piddock LJ. Quinolone resistance in *Escherichia coli*. *Vet Res* 2001; 32: 275–84.
17. Strahilevitz J, Jacoby GA, Hooper DC *et al.* Plasmid-mediated quinolone resistance: a multifaceted threat. *Clin Microbiol Rev* 2009; 22: 664–89.
18. Hawkey PM. Mechanisms of quinolone action and microbial response. *J Antimicrob Chemother* 2003; 51: 29–35.
19. Hopkins KL, Davies RH, Threlfall EJ. Mechanisms of quinolone resistance in *Escherichia coli* and *Salmonella*: recent developments. *Int J Antimicrob Agents* 2005; 25: 358–73.
20. Moher D, Liberati A, Tetzlaff J *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009; 6: e1000097.
21. Andrews JM, Andrews JM. Determination of minimum inhibitory concentrations. *J Antimicrob Chemother* 2001; 48: 5–16.

22. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint Tables for Interpretation of MICs and Zone Diameters. Version 8.1. 2018. http://www.eucast.org/clinical_breakpoints/.

23. The European Committee on Antimicrobial Susceptibility Testing. MIC and Zone Diameter Distributions and ECOFFs. 2018. http://www.eucast.org/mic_distributions_and_ecoffs/.

24. Heisig P, Wiedemann B. Use of a broad-host-range *gyrA* plasmid for genetic characterization of fluoroquinolone-resistant gram-negative bacteria. *Antimicrob Agents Chemother* 1991; 35: 2031–6.

25. Szklarczyk D, Franceschini A, Wyder S *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015; 43: D447–52.

26. Hu P, Janga SC, Babu M *et al*. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol 2009*; 7: e96.

27. Ogris C, Guala D, Kaduk M *et al*. FunCoup 4: new species, data, and visualization. *Nucleic Acids Res* 2018; 46: D601–7.

28. Zhou J, Rudd KE. EcoGene 3.0. *Nucleic Acids Res* 2013; 41: 613–24.

29. Yoshida H, Bogaki M, Nakamura M *et al*. Quinolone resistance-determining region in the DNA gyrase *gyrB* gene of *Escherichia coli*. *Antimicrob Agents Chemother* 1991; 35: 1647–50.

30. Linde HJ, Notka F, Metz M *et al*. In vivo increase in resistance to ciprofloxacin in Es*cherichia coli* associated with deletion of the C-terminal part of *marR*. *Antimicrob Agents Chemother* 2000; 44: 1865–8.

31. Sulavik MC, Houseweart C, Cramer C *et al*. Antibiotic susceptibility profiles of *Escherichia coli* strains lacking multidrug efflux pump genes. *Antimicrob Agents Chemother* 2001; 45: 1126–36.

32. Aly SA, Boothe DM, Suh SJ. A novel alanine to serine substitution mutation in SoxS induces overexpression of efflux pumps and contributes to multidrug resistance in clinical *Escherichia coli* isolates. *J Antimicrob Chemother* 2015; 70: 2228–33.

33. Pietsch F, Bergman JM, Brandis G *et al*. Ciprofloxacin selects for RNA polymerase mutations with pleiotropic antibiotic resistance effects. *J Antimicrob Chemother* 2017; 72: 75–84.

34. Yamane K, Wachino JI, Suzuki S *et al*. New plasmid-mediated fluoroquinolone efflux pump, QepA, found in an *Escherichia coli* clinical isolate. *Antimicrob Agents Chemother* 2007; 51: 3354–60.

35. Hansen LH, Jensen LB, Sørensen HI *et al*. Substrate specificity of the OqxAB multidrug resistance pump in *Escherichia coli* and selected enteric bacteria. *J Antimicrob Chemother* 2007; 60: 145–7.

36. Tran JH, Jacoby GA. Mechanism of plasmid-mediated quinolone resistance. *Proc Natl Acad Sci USA* 2002; 99: 5638–42.

37. Jacoby GA, Walsh KE, Mills DM *et al*. *qnrB*, another plasmid-mediated gene for quinolone resistance. *Antimicrob Agents Chemother* 2006; 50: 1178–82.

38. Wang M, Guo Q, Xu X *et al*. New plasmid-mediated quinolone resistance gene, *qnrC*, found in a clinical isolate of Proteus mirabilis. *Antimicrob Agents Chemother* 2009; 53: 1892–7.

39. Cavaco LM, Hasman H, Xia S *et al*. *qnrD*, a novel gene conferring transferable quinolone resistance in *Salmonella enterica* serovar Kentucky and Bovismorbificans strains of human origin. *Antimicrob Agents Chemother* 2009; 53: 603–8.

40. Albornoz E, Tijet N, De Belder D *et al*. *qnrE1*, a member of a new family of plasmid-located quinolone resistance genes, originated from the chromosome of *Enterobacter* species. *Antimicrob Agents Chemother* 2017; 61: pii: e02555-16.

41. Hata M, Suzuki M, Matsumoto M *et al*. Cloning of a novel gene for quinolone resistance from a transferable plasmid in *Shigella flexneri* 2b. *Antimicrob Agents Chemother* 2005; 49: 801–3.

42. Robicsek A, Strahilevitz J, Jacoby GA *et al*. Fluoroquinolone-modifying enzyme: a new adaptation of a common aminoglycoside acetyltransferase. *Nat Med* 2006; 12: 83–8.

43. Chávez-Jacobo V, Hernández-Ramírez K, Romo-Rodríguez P *et al*. CrpP is a novel ciprofloxacin-modifying enzyme encoded by the *Pseudomonas aeruginosa* pUM505 plasmid. *Antimicrob Agents Chemother* 2018; 62: 1–11.

44. Breines DM, Ouabdesselam S, Ng EY *et al*. Quinolone resistance locus *nfxD* of *Escherichia coli* is a mutant allele of the *parE* gene encoding a subunit of topoisomerase IV. *Antimicrob Agents Chemother* 1997; 41: 175–9.

45. Weigel LM, Steward CD, Tenover FC. *gyrA* mutations associated with fluoroquinolone resistance in eight species of Enterobacteriaceae. *Antimicrob Agents Chemother* 1998; 42: 2661–7.

46. Liu X, Lazzaroni C, Aly SA *et al*. *In vitro* selection of resistance to pradofloxacin and ciprofloxacin in canine uropathogenic *Escherichia coli* isolates. *Vet Microbiol* 2014; 174: 514–22.

47. Heisig P, Tschorny R. Characterization of fluoroquinolone-resistant mutants of *Escherichia coli* selected *in vitro*. *Antimicrob Agents Chemother* 1994; 38: 1284–91.

48. Huseby DL, Pietsch F, Brandis G *et al*. Mutation supply and relative fitness shape the genotypes of ciprofloxacin-resistant *Escherichia coli*. *Mol Biol Evol* 2017; 34: 1029–39.

49. Truong QC, Van Nguyen JC, Shlaes D *et al*. A novel, double mutation in DNA gyrase A of *Escherichia coli* conferring resistance to quinolone antibiotics. *Antimicrob Agents Chemother* 1997; 41: 85–90.

50. Allou N, Cambau E, Massias L *et al*. Impact of low-level resistance to fluoroquinolones due to *qnrA1* and *qnrS1* genes or a *gyrA* mutation on ciprofloxacin bactericidal activity in a murine model of *Escherichia coli* urinary tract infection. *Antimicrob Agents Chemother* 2009; 53: 4292–7.

51. Marcusson LL, Frimodt-Møller N, Hughes D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog* 2009; 5: e1000541.

52. Emrich NC, Heisig A, Stubbings W *et al*. Antibacterial activity of finafloxacin under different pH conditions against isogenic strains of *Escherichia coli* expressing combinations of defined mechanisms of fluoroquinolone resistance. *J Antimicrob Chemother* 2010; 65: 2530–3.

53. Briales A, Rodríguez-Martínez JM, Velasco C *et al*. *In vitro* effect of *qnrA1*, *qnrB1*, and *qnrS1* genes on fluoroquinolone activity against isogenic *Escherichia coli* isolates with mutations in *gyrA* and *parC*. *Antimicrob Agents Chemother* 2011; 55: 1266–9.

54. Khan DD, Lagerbäck P, Cao S *et al*. A mechanism-based pharmacokinetic/pharmacodynamic model allows prediction of antibiotic killing from MIC values for WT and mutants. *J Antimicrob Chemother* 2015; 70: 3051–60.

55. Webber MA, Buckner MMC, Redgrave LS *et al*. Quinolone-resistant gyrase mutants demonstrate decreased susceptibility to triclosan. *J Antimicrob Chemother* 2017; 72: 2755–63.

56. ambau E, Bordon F, Collatz E. Novel *gyrA* point mutation in a strain of *Escherichia coli* resistant to fluoroquinolones but not to nalidixic acid. *Antimicrob Agents Chemother* 1993; 37: 1247–52.

57. Oethinger M, Kern WV, Jellen-Ritter AS *et al*. Ineffectiveness of topoisomerase mutations in mediating clinically significant fluoroquinolone resistance in *Escherichia coli* in the absence of the AcrAB efflux pump. *Antimicrob Agents Chemother* 2000; 44: 10–13.

58. Yaron S, White DG, Matthews KR. Characterization of an *Escherichia coli* O157:H7 *marR* mutant. *Int J Food Microbiol* 2003; 85: 281–91.

59. Machuca J, Briales A, Labrador G *et al*. Interplay between plasmid-mediated and chromosomal-mediated fluoroquinolone resistance and bacterial fitness in *Escherichia coli*. *J Antimicrob Chemother* 2014; 69: 3203–15.

60. Praski Alzrigat L, Huseby DL, Brandis G *et al*. Fitness cost constrains the spectrum of *marR* mutations in ciprofloxacin-resistant *Escherichia coli*. *J Antimicrob Chemother* 2017; 72: 3016–24.

61. Blair JMA, Bavro VN, Ricci V *et al*. AcrB drug-binding pocket substitution confers clinically relevant resistance and altered substrate specificity. *Proc Natl Acad Sci USA* 2015; 112: 3511–16.

62. Zhao J, Chen Z, Chen S *et al*. Prevalence and dissemination of *oqxAB* in *Escherichia coli* isolates from animals, farmworkers, and the environment. *Antimicrob Agents Chemother* 2010; 54: 4219–24.

63. Sato T, Yokota SI, Uchida I *et al*. Fluoroquinolone resistance mechanisms in an *Escherichia coli* isolate, HUE1, without quinolone resistance-determining region mutations. *Front Microbiol* 2013; 4: 125.

64. Wang J, Guo Z-W, Zhi C-P *et al*. Impact of plasmid-borne *oqxAB* on the development of fluoroquinolone resistance and bacterial fitness in *Escherichia coli*. *J Antimicrob Chemother* 2017; 72: 1293–302.

65. Yamane K, Wachino JI, Suzuki S *et al*. Plasmid-mediated *qepA* gene among *Escherichia coli* clinical isolates from Japan. *Antimicrob Agents Chemother* 2008; 52: 1564–6.

66. Périchon B, Courvalin P, Galimand M. Transferable resistance to aminoglycosides by methylation of G1405 in 16S rRNA and to hydrophilic fluoroquinolones by QepA-mediated efflux in *Escherichia coli*. *Antimicrob Agents Chemother* 2007; 51: 2464–9.

67. Machuca J, Briales A, Díaz-de-Alba P *et al*. Effect of the efflux pump QepA2 combined with chromosomally mediated mechanisms on quinolone resistance and bacterial fitness in *Escherichia coli*. *J Antimicrob Chemother* 2015; 70: 2524–7.

68. Manageiro V, Félix D, Jones-Dias D *et al*. Genetic background and expression of the new *qepA4* gene variant recovered in clinical TEM-1- and CMY-2-producing *Escherichia coli*. *Front Microbiol* 2017; 8: 1899.

69. Tran JH, Jacoby GA, Hooper DC. Interaction of the plasmid-encoded quinolone resistance protein QnrA with *Escherichia coli* topoisomerase IV interaction of the plasmid-encoded quinolone resistance protein QnrA with *Escherichia coli* topoisomerase IV. *Antimicrob Agents Chemother* 2005; 49: 4–7.

70. Jacoby GA. *qnr* Numbering and Sequences. http://www.lahey.org/qnrstudies/.

71. Chowdhury G, Pazhani GP, Nair GB *et al*. Transferable plasmid-mediated quinolone resistance in association with extended-spectrum β-lactamases and fluoroquinolone-acetylating aminoglycoside-6′-N-acetyltransferase in clinical isolates of *Vibrio fluvialis*. *Int J Antimicrob Agents* 2011; 38: 169–73.

72. Silva-Sánchez J, Cruz-Trujillo E, Barrios H *et al*. Characterization of plasmid-mediated quinolone resistance (PMQR) genes in extended-spectrum β-lactamase-producing Enterobacteriaceae pediatric clinical isolates in Mexico. *PLoS One* 2013; 8: e77968.

73. Shaheen BW, Nayak R, Foley SL *et al*. Chromosomal and plasmid-mediated fluoroquinolone resistance mechanisms among broad-spectrum-cephalosporin-resistant *Escherichia coli* isolates recovered from companion animals in the USA. *J Antimicrob Chemother* 2013; 68: 1019–24.

74. Varela AR, Macedo GN, Nunes OC *et al*. Genetic characterization of fluoroquinolone resistant *Escherichia coli* from urban streams and municipal and hospital effluents. *FEMS Microbiol Ecol* 2015; 91: pii: fiv015.

75. Machuca J, Ortiz M, Recacha E *et al*. Impact of AAC(6′)-Ib-cr in combination with chromosomal-mediated mechanisms on clinical quinolone resistance in *Escherichia coli*. *J Antimicrob Chemother* 2016; 71: 3066–71.

76. Yanat B, Machuca J, Díaz-De-Alba P *et al*. Characterization of plasmid-mediated quinolone resistance determinants in high-level quinolone-resistant Enterobacteriaceae isolates from the community: first report of *qnrD* gene in Algeria. *Microb Drug Resist* 2017; 23: 90–7.

77. Heisig P, Schedletzky H, Falkenstein-Paul H. Mutations in the *gyrA* gene of a highly fluoroquinolone-resistant clinical isolate of *Escherichia coli*. *Antimicrob Agents Chemother* 1993; 37: 696–701.

78. Webber MA, Talukder A, Piddock LJV. Contribution of mutation at amino acid 45 of AcrR to *acrB* expression and ciprofloxacin resistance in clinical and veterinary *Escherichia coli* isolates. *Antimicrob Agents Chemother* 2005; 49: 4390–2.

79. Silva-Sanchez J, Barrios H, Reyna-Flores F *et al*. Prevalence and characterization of plasmid-mediated quinolone resistance genes in extended-spectrum β-lactamase-producing Enterobacteriaceae isolates in Mexico. *Microb Drug Resist* 2011; 17: 497–505.

80. Goto K, Kawamura K, Arakawa Y. Contribution of QnrA, a plasmid-mediated quinolone resistance peptide, to survival of *Escherichia coli* exposed to a lethal ciprofloxacin concentration. *Jpn J Infect Dis* 2015; 68: 196–202.

81. Cesaro A, Bettoni RRD, Lascols C *et al*. Low selection of topoisomerase mutants from strains of *Escherichia coli* harbouring plasmid-borne *qnr* genes. *J Antimicrob Chemother* 2008; 61: 1007–15.

82. Jakobsen L, Cattoir V, Jensen KS *et al*. Impact of low-level fluoroquinolone resistance genes *qnrA1*, *qnrB19* and *qnrS1* on ciprofloxacin treatment of isogenic *Escherichia coli* strains in a murine urinary tract infection model. *J Antimicrob Chemother* 2012; 67: 2438–44.

83. Martínez-Martínez L, Pascual A, Jacoby GA *et al*. Quinolone resistance from a transferable plasmid. *Lancet* 1998; 351: 797–9.

84. Martínez-Martínez L, Pascual A, García I *et al*. Interaction of plasmid and host quinolone resistance. *J Antimicrob Chemother* 2003; 51: 1037–9.

85. Martín-Gutiérrez G, Rodríguez-Martínez JM, Pascual A *et al*. Plasmidic *qnr* genes confer clinical resistance to ciprofloxacin under urinary tract physiological conditions. *Antimicrob Agents Chemother* 2017; 61: pii: e02615-16.

86. Rodríguez-Martínez JM, Velasco C, Pascual A *et al*. Correlation of quinolone resistance levels and differences in basal and quinolone-induced expression from three *qnrA*-containing plasmids. *Clin Microbiol Infect* 2006; 12: 440–5.

87. Rodríguez-Martínez JM, Velasco C, Briales A *et al*. Qnr-like pentapeptide repeat proteins in Gram-positive bacteria. *J Antimicrob Chemother* 2008; 61: 1240–3.

88. Wang M, Tran J, Jacoby G. Plasmid-mediated quinolone resistance in clinical isolates of *Escherichia coli* from Shanghai, China. *Antimicrob Agents Chemother* 2003; 47: 2242–8.

89. Xu X, Wu S, Ye X *et al*. Prevalence and expression of the plasmid-mediated quinolone resistance determinant *qnrA1*. *Antimicrob Agents Chemother* 2007; 51: 4105–10.

90. Jones-Dias D, Manageiro V, Francisco AP *et al*. Assessing the molecular basis of transferable quinolone resistance in *Escherichia coli* and *Salmonella* spp. from food-producing animals and food products. *Vet Microbiol* 2013; 167: 523–31.

91. Shin JH, Jung H, Lee J *et al*. High rates of plasmid-mediated quinolone resistance QnrB variants among ciprofloxacin-resistant *Escherichia coli* and *Klebsiella pneumoniae* from urinary tract infections in Korea. *Microb Drug Resist* 2008; 14: 221–6.

92. Cerquetti M, García-Fernández A, Giufrè M *et al*. First report of plasmid-mediated quinolone resistance determinant *qnrS1* in an *Escherichia coli* strain of animal origin in Italy. *Antimicrob Agents Chemother* 2009; 53: 3112–14.

93. Okumura R, Liao CH, Gavin M *et al*. Quinolone induction of *qnrVS1* in *Vibrio splendidus* and plasmid-carried *qnrS1* in *Escherichia coli*, a mechanism independent of the SOS system. *Antimicrob Agents Chemother* 2011; 55: 5942–5.

94. Xue G, Li J, Feng Y *et al.* High prevalence of plasmid-mediated quinolone resistance determinants in *Escherichia coli* and *Klebsiella pneumoniae* isolates from pediatric patients in China. *Microb Drug Resist* 2017; 23: 107–14.

95. Bönemann G, Stiens M, Pühler A *et al.* Mobilizable IncQ-related plasmid carrying a new quinolone resistance gene, *qnrS2*, isolated from the bacterial community of a wastewater treatment plant. *Antimicrob Agents Chemother* 2006; 50: 3075–80.

96. Ruiz E, Sáenz Y, Zarazaga M *et al. qnr*, *aac*(6′)-*Ib-cr* and *qepA* genes in *Escherichia coli* and *Klebsiella* spp.: genetic environments and plasmid and chromosomal location. *J Antimicrob Chemother* 2012; 67: 886–97.

97. Sekyere JO, Amoako DG. Genomic and phenotypic characterisation of fluoroquinolone resistance mechanisms in Enterobacteriaceae in Durban, South Africa. *PLoS One* 2017; 12: 1–14.

98. Vinué L, Hooper DC, Jacoby GA. Chromosomal mutations that accompany *qnr* in clinical isolates of *Escherichia coli*. *Int J Antimicrob Agents* 2018; 51: 479–83.

99. Drake JW, Charlesworth B, Charlesworth D *et al.* Rates of spontaneous mutation. *Genetics* 1998; 148: 1667–86.

100. Acharya S, Foster PL, Brooks P *et al.* The coordinated functions of the *E. coli* MutS and MutL proteins in mismatch repair. *Mol Cell* 2003; 12: 233–46.

101. Walker G. Mutagenesis and inducible responses to deoxyribonucleic acid damage in *Escherichia coli*. *Microbiol Rev* 1984; 48: 60–93.

102. Erill I, Campoy S, Barbé J. Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS Microbiol Rev* 2007; 31: 637–56.

103. Recacha E, Machuca J, Díaz de Alba P *et al.* Quinolone resistance reversion by targeting the SOS response. *MBio* 2017; 8: e00971–17.

104. Da Re S, Garnier F, Guérin E *et al.* The SOS response promotes *qnrB* quinolone-resistance determinant expression. *EMBO Rep* 2009; 10: 929–33.

105. Wang M, Jacoby GA, Mills DM *et al.* SOS regulation of *qnrB* expression. *Antimicrob Agents Chemother* 2009; 53: 821–3.

106. Beaber JW, Hochhut B, Waldor MK. SOS response promotes horizontal dissemination of antibiotic resistance genes. *Nature* 2004; 427: 72–4.

107. Hughes D, Andersson DI. Evolutionary trajectories to antibiotic resistance. *Annu Rev Microbio*l 2017; 71: 579–96.

108. Garoff L, Yadav K, Hughes D. Increased expression of Qnr is sufficient to confer clinical resistance to ciprofloxacin in *Escherichia coli*. *J Antimicrob Chemother* 2017; 73: 348–52.

109. Shigemura K, Tanaka K, Yamamichi F *et al.* Does mutation in *gyrA* and/or *parC* or efflux pump expression play the main role in fluoroquinolone resistance in *Escherichia coli* urinary tract infections?: A statistical analysis study. *Int J Antimicrob Agents* 2012; 40: 516–20.

110. Swick MC, Morgan-Linnell SK, Carlson KM *et al.* Expression of multidrug efflux pump genes *acrAB-tolC*, *mdfA*, and *norE* in *Escherichia coli* clinical isolates as a function of fluoroquinolone and multidrug resistance. *Antimicrob Agents Chemother* 2011; 55: 921–4.

111. Pesesky MW, Hussain T, Wallace M *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.

Supplementary material

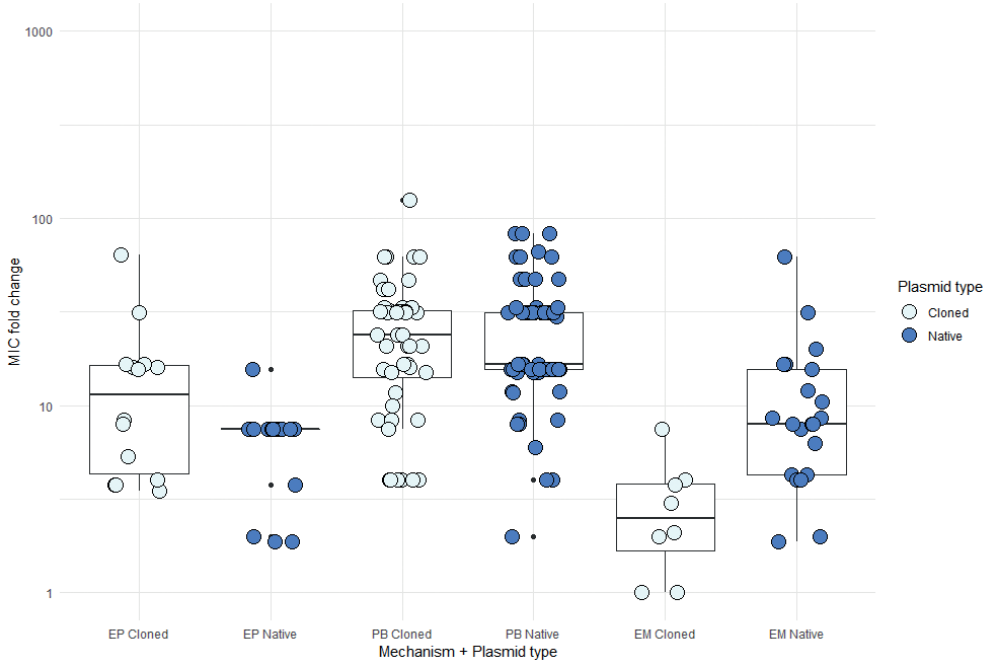


Figure S1. Ciprofloxacin MIC fold changes conferred by single plasmid-mediated resistance genes, stratified per mechanism and by plasmid type. A resistance gene is considered to be in the native plasmid if the gene has not been cloned into another vector after isolation from a clinical or environmental sample. A resistance gene is considered to be cloned when the authors have stated in the article that the gene has been cloned into another vector, or if the authors use a plasmid that has been cloned into another vector before. Mechanisms: EP = Efflux Pump (*qepA* or *oqxAB*), PB = Physical Blocking (any *qnr* gene), EM = Enzymatic Modification (*aac(6')Ib-cr* or *crpP*). This figure is also available on FigShare. Doi: 10.6084/m9.figshare.6974171.

Table S1. Full set of genetic modifications, resistance determinants present before genetic modification and MIC data extracted from 60 experimental studies, totaling 366 *E. coli* isolates. Resistance determinants are classified on mechanism, gene and if applicable, amino acid residue. For *gyrA*, *gyrB* and *parC*, specific residues are indicated and amino acids are expressed in single letter code. For *marR*, *soxS*, *acrR* and efflux pump subunits it is indicated if the gene was deleted (DEL) or mutated (MUT) in the tested isolate. Presence of *oqxAB*, *qepA*, *qnrA*, *qnrB*, *qnrC*, *qnrD*, *qnrE*, *qnrS*, *aac(6')Ib-cr* and *crpP* is indicated by 'YES', if no information on the allele was provided. Numbers indicate the alleles of these genes where possible. Information on the genetic background of genertically modified strains was extracted from the paper. No prior resistance determinants are indicated by a hyphen (-) and if genetic background of strains was not defined indicated by a question mark (?). An asterisk indicates a comment on the information is available at the end of the table. References are included at the bottom of the table.

This table (432 rows by 45 columns) is available from Figshare (https://tinyurl.com/f09a3) or with the online version of this article.

Table S2. List of edges used to visualize the PPI network. Combined_score indicates the score assigned to the interaction in the String v10 database. The label indicates the type of interaction. PI=Physical Interaction, FP=Functional Prediction, TM=Text Mining. The edge is assigned to the highest level type of interaction, where PI > FP > TM. Thus, if interactions are found between two proteins that can be assigned to Physical Interaction and Functional Prediction, the interaction will be labeled as "PI" in this table.

protein1	protein2	combined_score	label
acrB	*acrA*	999	PI
acrB	*acrR*	922	FP
acrB	*emrE*	731	FP
acrB	*cusC*	928	FP
acrB	*macA*	855	FP
acrB	*macB*	714	FP
acrB	*mdtA*	916	FP
acrB	*emrK*	727	FP
acrB	*emrB*	710	FP
acrB	*tolC*	998	PI
acrB	*acrE*	945	FP
acrB	*mdtE*	965	FP
acrA	*acrR*	962	FP
acrA	*cusC*	969	FP
acrA	*macA*	837	FP
acrA	*macB*	813	FP
acrA	*marR*	746	TM
acrA	*mdtB*	875	FP
acrA	*mdtC*	867	FP
acrA	*gyrA*	823	TM
acrA	*acrD*	999	PI
acrA	*emrA*	750	FP
acrA	*emrB*	748	FP
acrA	*parC*	749	TM
acrA	*tolC*	999	PI
acrA	*acrF*	981	FP
acrA	*mdtF*	966	FP
acrA	*soxS*	790	TM
acrA	*mdtP*	839	FP
acrR	*marR*	815	TM
acrR	*gyrA*	727	FP
acrR	*acrD*	848	FP
acrR	*parC*	756	FP
acrR	*tolC*	880	FP
acrR	*acrF*	832	FP
acrR	*soxS*	847	FP
emrE	*acrD*	728	FP
emrE	*emrD*	710	TM

protein1	protein2	combined_score	label
cusC	*macA*	833	FP
cusC	*macB*	851	FP
cusC	*mdtA*	731	FP
cusC	*mdtB*	715	FP
cusC	*mdtC*	715	FP
cusC	*acrD*	948	FP
cusC	*acrE*	950	FP
cusC	*acrF*	941	FP
cusC	*mdtE*	948	FP
cusC	*mdtF*	942	FP
mdfA	*gyrA*	817	FP
mdfA	*tolC*	811	FP
mdfA	*rpoB*	712	TM
macA	*macB*	999	PI
macA	*mdtB*	778	FP
macA	*mdtC*	716	FP
macA	*emrK*	706	FP
macA	*acrD*	858	FP
macA	*emrA*	727	FP
macA	*tolC*	998	PI
macA	*acrF*	808	FP
macA	*mdtF*	806	FP
macB	*mdtB*	777	FP
macB	*mdtC*	794	FP
macB	*acrD*	824	FP
macB	*tolC*	999	PI
macB	*acrF*	744	FP
macB	*mdtF*	749	FP
macB	*mdtP*	839	FP
tehA	*tehB*	983	FP
marR	*gyrA*	984	PI
marR	*parC*	713	TM
marR	*tolC*	846	FP
marR	*soxS*	816	TM
mdtA	*mdtB*	999	FP
mdtA	*mdtC*	999	FP
mdtA	*emrK*	776	FP
mdtA	*acrD*	969	FP
mdtA	*emrA*	718	FP
mdtA	*tolC*	964	FP
mdtA	*acrF*	908	FP
mdtA	*mdtF*	906	FP
mdtB	*mdtC*	999	PI
mdtB	*emrA*	766	FP

protein1	protein2	combined_score	label
mdtB	*emrB*	757	FP
mdtB	*tolC*	976	FP
mdtB	*acrE*	711	FP
mdtB	*mdtE*	850	FP
mdtC	*emrA*	776	FP
mdtC	*emrB*	753	FP
mdtC	*tolC*	976	FP
mdtC	*mdtE*	836	FP
gyrA	*parC*	810	PI
gyrA	*tolC*	861	FP
gyrA	*gyrB*	999	PI
gyrA	*rpoB*	972	FP
emrY	*emrK*	999	FP
emrY	*emrA*	938	FP
emrY	*tolC*	959	FP
emrY	*mdtP*	769	FP
emrK	*acrD*	734	FP
emrK	*emrB*	941	FP
emrK	*tolC*	964	FP
emrK	*mdtE*	783	FP
emrK	*mdtP*	774	FP
acrD	*emrA*	717	FP
acrD	*emrB*	714	FP
acrD	*tolC*	989	FP
acrD	*acrE*	932	FP
acrD	*mdtE*	964	FP
acrD	*soxS*	712	TM
emrA	*emrB*	999	FP
emrA	*tolC*	997	PI
emrA	*mdtP*	742	FP
emrB	*tolC*	983	FP
emrB	*acrF*	750	FP
emrB	*mdtF*	711	FP
parC	*tolC*	777	FP
parC	*gyrB*	999	FP
parC	*rpoB*	736	FP
tolC	*acrE*	969	FP
tolC	*acrF*	987	FP
tolC	*mdtE*	981	FP
tolC	*mdtF*	978	FP
tolC	*gyrB*	755	FP
tolC	*soxS*	817	TM
acrE	*acrF*	999	FP
acrE	*mdtF*	920	FP

protein1	protein2	combined_score	label
acrE	*mdtP*	845	FP
acrF	*mdtE*	959	FP
acrF	*soxS*	713	TM
mdtE	*mdtF*	999	FP
mdtE	*mdtP*	884	FP
mdtF	*mdtP*	743	FP
gyrB	*rpoB*	963	FP

Table S3. Full set of mutations, MIC data and isolate information extracted from 8 observational studies, totaling 238 E. coli isolates. Resistance determinants are stratified on mechanism, gene and if applicable, amino acid residue. For gyrA, gyrB, parC and parE, specific residues are indicated and amino acids are expressed in single letter code. For acrR, marR, soxR and rpoB, it is indicated which genetic changes where observed in the tested isolate. Presence of oqxAB, qepA, qnrA, qnrB, qnrS and aac(6')Ib-cr is indicated by 'YES', if no information on the allele was provided. Numbers indicate the alleles of these genes where possible. Hyphens indicate the residue was tested, but no differences were observed compared with wild type E. coli. References are included at the bottom of the table.

This table (254 rows by 34 columns) is available from Figshare (https://tinyurl.com/f09a3) or with the online version of this article.

Supplementary methods

The full search used for PubMed is:

(CIPROFLOXACIN[MESH] OR CIPROFLOXACIN[TIAB] OR FLUOROQUIN*[TIAB] OR DNA GYRASE[MESH] OR GYRA[TIAB]) AND ("ESCHERICHIA COLI/DRUG EFFECTS"[MESH] OR ESCHERICHIA COLI[TIAB] OR E. COLI[TIAB]) AND (ANTIBIOTIC RESISTANCE, MICROBIAL[MESH] OR MICROBIAL SENSITIVITY TEST[MESH] OR MIC[TIAB] OR RESIST*[TIAB]) AND (JOURNAL ARTICLE)

The last search was performed on July 5th, 2018 and yielded 5055 articles. Pubreminer was used to select genetic resistance determinants that were mentioned in the title or abstract of these 5055 articles. The selected resistance determinants were:

((GYRASE OR GYRASES) OR (GYRB OR GYRBS) OR (PARC OR PARCS) OR PARE OR (QNR OR QNRS) OR (SOX OR SOXS) OR (SOXR OR SOXRS) OR AAC OR ACRA OR ACRAB OR ACRB OR ACRR OR DELTAACRAB OR DELTAACRB OR DELTAMARR OR DELTATOLC OR GYRA OR GYRA1AB OR GYRA43 OR GYRA462 OR GYRA87 OR MARA OR MARO OR MAROR OR MARR OR MARRAB OR OQXA OR OQXAB OR OQXB OR QEPA OR QEPA1 OR QEPA2 OR QEPA4 OR QNRA OR QNRA1 OR QNRA3 OR QNRA6 OR QNRB OR QNRB1 OR QNRB10 OR QNRB19 OR

QNRB2 OR QNRB4 OR QNRB5 OR QNRB6 OR QNRB7 OR QNRB9 OR QNRC OR QNRD OR QNRS1 OR QNRS2 OR QNRVC1 OR QNRVC3 OR QNRVC4 OR QNRVS1 OR ROB OR TOLC)

Articles that failed to mention at least one of these resistance determinants in the title or abstract were expected to not cover genetic mechanisms of ciprofloxacin resistance and were thus excluded. This left 1129 articles identified from PubMed.

For Web of Science (WoS), the preliminary search was performed using the following set of keywords:

((TS=(Escherichia coli OR e. coli OR coli)) AND DOCUMENT TYPES: (Article)) AND ((TS=(ciprofloxacin) OR TI=(ciprofloxacin AND resistan*)) AND DOCUMENT TYPES: (Article))

The last search was performed on July 5th, 2018 and yielded 5873 articles. The same list of resistance determinants was used to exclude articles not mentioning genetically encoded ciprofloxacin resistance determinants. This left 1260 articles identified from WoS.

Next, the identified articles from PubMed and WoS were listed and 671 duplicates were removed. Finally, the remaining 1718 articles were screened on title, abstract and if needed, full text. This yielded inclusion of 50 experimental studies and 3 observational studies. After adapting the inclusion criteria for observational studies, 5 more observational studies were included. 10 experimental studies were identified through backward and forward searching of articles of interest. 2 articles were included as experimental and observational studies. This amounted to a total of 66 included studies.

Chapter 3

Escherichia ruysiae sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller

Boas C.L. van der Putten, Sébastien Matamoros, Daniel R. Mende, Edwin R. Scholl, COMBAT consortium, Constance Schultsz

International Journal of Systematic and Evolutionary Microbiology, Volume 71, Issue 2, 6 January 2021, https://doi.org/10.1099/ijsem.0.004609

Author notes

The Genbank accession number for the 16S rRNA gene sequence of strain OPT1704^T is LR745848. The Genbank accession number for the complete genome sequence of strain OPT1704^T is CABVLQ000000000.

Abstract

The *Escherichia* genus comprises four species and at least five lineages currently not assigned to any species, termed '*Escherichia* cryptic clades'. We isolated an *Escherichia* strain from an international traveller and resolved the complete DNA sequence of the chromosome and an IncI multi-drug resistance plasmid using Illumina and Nanopore whole-genome sequencing (WGS). Strain OPT1704^T can be differentiated from existing *Escherichia* spp. using biochemical (VITEK2) and genomic tests (average nucleotide identity [ANI] and digital DNA:DNA hybridisation [dDDH]). Phylogenetic analysis based on alignment of 16S rRNA sequences and 682 concatenated core genes showed similar results. Our analysis further revealed that strain OPT1704^T falls within *Escherichia* cryptic clade IV, and is closely related to cryptic clade III. Combining our analyses with publicly available WGS data of cryptic clades III and IV from Enterobase confirmed the close relationship between clades III and IV (>96% interclade ANI), warranting assignment of both clades to the same novel species. We propose *E. ruysiae* sp. nov. as a novel species, encompassing *Escherichia* cryptic clades III and IV (type strain OPT1704^T = NCCB 100732^T = NCTC 14359^T).

Introduction

Within the *Escherichia* genus, five species are recognized; *E. coli*[1], *E. hermannii*[2], *E. fergusonii*[3], *E. albertii*[4] and most recently, *E. marmotae*[5]. It has been proposed to reassign *E. hermannii* to *Atlantibacter hermannii*[6], but this reassignment is not yet formally approved. Several other species were previously reassigned from the *Escherichia* genus to other genera, such as *E. vulneris* (now *Pseudescherichia vulneris*[7]), *E. blattae* (now *Shimwellia blattae*[8]) and *E. adecarboxylata* (now *Leclercia adecarboxylata*[9]). All five current *Escherichia* species have been associated with the potential to cause animal and/or human disease[2,10–13]. Several *Escherichia* strains cannot be assigned to any of the five existing species[14]. Based on analysis of genomic data, these strains cluster into several groups, which were termed '*Escherichia* cryptic clades', numbered I through VI[14,15]. Recently, cryptic clade V was formally recognized as a separate species (*E. marmotae*), leaving at least five cryptic clades that have not been delineated at the species level[5]. Here we report the novel species *Escherichia ruysiae* sp. nov., isolated from faecal material of an international traveller. *Escherichia ruysiae* sp. nov. encompasses the closely related *Escherichia* cryptic clades III and IV.

Isolation and Ecology

We discovered a cryptic clade IV strain in our collection, previously identified as extended spectrum beta-lactamase (ESBL) producing *E. coli* as part of the COMBAT study, which investigated the acquisition of ESBL-producing Enterobacteriaceae (ESBL-E) during international travel[16]. This isolate, OPT1704^T, was further characterized in detail.

The strain was isolated from a human faecal sample provided immediately after an individual's return from a one-month journey to several Asian countries. No ESBL-E were detected in a faecal sample collected immediately before departure, suggesting the ESBL gene, and possibly strain OPT1704^T, were acquired during travel. The traveller reported diarrhoea during travel but no antibiotic usage. No ESBL-E were isolated in follow-up faecal samples, suggesting loss of the OPT1704^T strain or the ESBL gene within one month after return from travel.

Genome Features

The whole-genome sequence of strain OPT1704^T was determined using a combination of the Illumina HiSeq and Oxford Nanopore Technologies (ONT) sequencing platforms. Strain OPT1704^T was grown o/n in liquid LB at 37 °C. DNA for Illumina sequencing

was extracted using the Qiagen Blood and Tissue kit (cat nr. 69506, Qiagen) and the sequencing library was prepared using the Illumina Nextera XT DNA Library Preparation kit (cat nr. FC-131-1096, Illumina), both according to manufacturer's instructions. DNA for ONT sequencing was extracted using the Qiagen MagAttract HMW DNA extraction kit (cat nr. 67563, Qiagen) and the sequencing library was prepared using the native barcoding and ligation sequencing kits (cat. nr. EXP-NBD114 and SQK-LSK109, respectively, Oxford Nanopore Techonologies) according to manufacturer's instructions. The Illumina sequencing run yielded a total of 6.3×10^6 paired-end reads, with a mean read length of 151 bp. Default parameters were used in bioinformatic analyses unless noted otherwise. Illumina reads were filtered using fastp (using flag "--disable-length-filtering", version 0.19.5,[17]) and downsampled using seqtk (version 1.3-r106, https://github.com/lh3/seqtk) to provide a theoretical coverage depth of 100X with the assumption that the strain OPT1704^T has a genome size of approximately 5×10^6 bp. The ONT sequencing run yielded a total of 2.5×10^4 reads, with a mean read length of 9078 bp before filtering. ONT reads were filtered on length and on read identity using Filtlong (version 0.2.0, https://github.com/rrwick/Filtlong) with Illumina reads as a reference, leaving 1.5×10^4 reads with a mean length of 12580 bp. This provided a theoretical coverage depth of ~38X of ONT reads. The combined assembly using Unicycler (version 0.4.6[18]) of Illumina and Nanopore reads resulted in a completely assembled genome, consisting of one circular chromosome (4,651,588 bp) and one circular plasmid (116,086 bp). The GC content of the complete strain OPT1704^T genome was 50.6%.

Putative resistance and virulence genes were predicted from the complete genome using ABRicate (https://github.com/tseemann/abricate) with the CARD[19] and VFDB[20] databases. Strain OPT1704^T harbours 6 resistance genes on its IncI plasmid, associated with reduced susceptibility to fluoroquinolones (*qnrS1*), aminoglycosides (*aph(6)-Id* & *aph(3'')-Ib*), cephalosporins ($bla_{CTX\text{-}M\text{-}14}$), trimethoprim (*dfrA14*) and sulphonamides (*sul2*), corresponding with its reduced susceptibility to fluoroquinolones (norfloxacin, MIC: 2 mg/L and ciprofloxacin, MIC: 0.5 mg/L), cephalosporins (cefuroxime, MIC: >32 mg/L and cefotaxime, MIC: 4 mg/L) and trimethoprim-sulfamethoxazole (MIC: >8 mg/L), assessed using VITEK2 (BioMérieux). However, strain OPT1704^T was susceptible to tobramycin (MIC: ≤1 mg/L) and gentamicin (MIC: ≤1 mg/L) despite presence of aminoglycoside resistance genes *aph(6)-Id* and *aph(3'')-Ib*. The *aph(6)-Id* gene encodes an aminoglycoside modifying enzyme that mediates resistance against streptomycin[21]. The *aph(3'')-Ib* gene encodes an aminoglycoside modifying enzyme mediating resistance against tobramycin and gentamicin[22] but the *aph(3'')-Ib* variant identified in strain OPT1704^T possesses a Glu18Lys mutation which maps to the catalytic phosphorylase kinase

domain (assessed with InterPro[23]). This could potentially inhibit enzymatic function, explaining the observed susceptibility to gentamicin and tobramycin, based on clinical breakpoints[24]. Furthermore, several putative virulence genes were predicted from the genome sequence associated with siderophore function (*chuX*, *entS*, *fepABD*), fimbriae (*fimBCDGI*), a type II secretion system (*gspGHI*) and capsular polysaccharide biosynthesis (*kpsD*). These predicted virulence genes, when present in *Escherichia coli,* are not typically associated with a specific clinical syndrome such as diarrhoeal disease.

Physiology and Chemotaxonomy

Strain OPT1704^T formed circular, grey-white colonies on a Columbia sheep (COS) blood agar plate when incubated overnight at 37 ˚C. No haemolysis was observed. Individual cells were observed using transmission electron microscopy (TEM) and were rod-shaped and on average 0.7 by 1.9 µm in size (Fig. S1). Bacteria were fixed with McDowell fixative (4% v/v PFA and 1% v/v GA in 0.1 M phosphate buffer) with 1.5% lysine acetate (Merck) for 4 hours and postfixed with 1% osmium tetraoxide (Electron Microscopy Sciences) for 1 hour. Afterwards, the bacteria were dehydrated using an ethanol series and embedded in Epon 812 (Ladd Research). Copper grids covered with formvar were used to collect 60-70 nm sections made using a Leica EM FC6 ultramicrotome (Leica). Sections were stained with uranyl acetate (Merck) and lead citrate (Laurylab). Electron micrographs were collected using an FEI Tecnai T12 Biotwin electron microscope (FEI Company) operated at 120 kV and equipped with an EMSIS Xarosa camera. Subsequently, we tested motility using the hanging drop method, oxidase presence using an oxidase strip (cat. nr. 40560, Sigma Aldrich) and catalase presence using H_2O_2[25]. The strain was shown to be Gram-negative, non-motile, oxidase-negative and catalase-positive. The strain was capable to grow in the absence of oxygen. On COS blood plates, it showed growth in the temperature range of 20-42 °C. The strain was also able to grow in NaCl concentrations ranging from 0% to 6% w/v in lysogeny broth overnight at 37 ˚C, but not at NaCl concentrations from 7% to 10% w/v (1% steps). MALDI-TOF (Bruker) and VITEK2 (BioMérieux) systems both identified strain OPT1704^T as *E. coli* with high confidence scores (score>2 for MALDI-TOF and “Excellent identification” for VITEK2, see Supplemental information for MALDI-TOF spectrum). Comparison of the output of the VITEK2 biochemical test with published biochemical reactions of other *Escherichia* species revealed that *E. ruysiae* sp. nov. str. OPT1704^T is distinct from other *Escherichia* species based on a combination of biochemical markers (table 1) (2–5,26).

Table 1. Comparison of biochemical markers which differentiate *E. ruysiae* sp. nov. from other *Escherichia* species. + and – indicate that ≥85% of tested strains is positive or negative for that biochemical marker, respectively. Data for *E. albertii*, *E. coli*, *E. fergusonii* and *E. marmotae* summarised from literature (2–5,26).

	E. ruysiae	*E. albertii*	*E. coli*	*E. fergusonii*	*E. hermannii*	*E. marmotae*
ONPG	+	+	+	+	+	–
Lysine decarboxylase	-	+	+	+	-	+
Ornithine decarboxylase	+	+	+*	+	+	–
Fermentation of:						
Adonitol	-	–	–	+	-	–
d-Xylose	-	–	+	+	+	+
Cellobiose	-	–	–	+	+	–
d-Sorbitol	+	–	+	–	-	+

*50-85% of *E. coli* possess this biochemical property.

16S rRNA and whole-genome phylogeny

Next, we calculated 16S rRNA sequence similarities, ANI values and digital DNA:DNA hybridisation (dDDH) values between strain OPT1704^T and type strains of the four other *Escherichia* species, representative genomes of the other three *Escherichia* cryptic clades, and *S. enterica* serovar Typhimurium (table 2). Representative genomes for the *Escherichia* cryptic clades I, II, III and VI were selected from Enterobase[27], using the genomes with the highest contiguity. Clades VII and VIII in Enterobase only consisted of a single strain and were not used in further analyses. We used three separate tools to calculate average nucleotide identity (ANI) (fastANI[28], OrthoANIu[29] and ANI calculator from Enveomics[30]). Multiple ANI calculation algorithms were employed to increase confidence in the genomic species delineation, as different ANI algorithms can output different ANI values. In this study, calculated ANI values were similar across ANI calculation algorithms. We also included calculation of the digital DNA:DNA hybrisation (dDDH) values between strains using the DSMZ Genome-to-Genome Distance Calculator to calculate[31]. The output of formula 2 of the DSMZ Genome-to-Genome Distance Calculator was used, as recommended by the authors of the tool. 16S rRNA genes were extracted from whole genomes using barrnap (version 0.9, https://github.com/tseemann/barrnap) and SNPs between strains were counted using snp-dists (version 0.6, https://github.com/tseemann/snp-dists). Extracted 16S rRNA gene segments were 1538 bp long for all strains and were manually aligned and checked. The alignment is provided in the supplementary material.

Table 2. Comparison of strain OPT1704^T 16S rRNA and whole-genome sequence with type strains of *E. albertii*, *E. coli*, *E. fergusonii*, *E. marmotae*, representative genomes of *Escherichia* cryptic clades I, II, III and VI and *S. enterica* serovar Typhimurium. In bold are the values that warrant assignment of strain OPT1704^T to a novel species (<98.7% 16S rRNA sequence similarity, <95-96% ANI, <70% dDDH). ANI: average nucleotide identity, dDDH: digital DNA:DNA hybridisation.

	***E. ruysiae* sp. nov. OPT1704^T**				
	16S rRNA sequence similarity (%)	**ANI (%, fastANI)**	**ANI (%, OrthoANIu)**	**ANI (%, ANI calculator Enveomics)**	**dDDH (%)**
E. albertii NBRC 107761^T	**98.6**	**90.0**	**90.0**	**89.2**	**39.8**
E. coli ATCC 11775^T	98.7	**92.8**	**92.4**	**92.0**	**48.3**
E. fergusonii ATCC 35469^T	98.9	**89.4**	**88.2**	**89.7**	**36.7**
E. hermanni NCTC 12129^T	**98.0**	**80.2**	**77.7**	**80.1**	**21.4**
E. marmotae HT073016^T	98.9	**92.2**	**92.2**	**91.4**	**47.1**
S. enterica Typhimurium LT2^T	**97.5**	**82.1**	**80.7**	**81.8**	**24.0**
Escherichia cryptic clade I 89-3506	99.0	**92.5**	**92.1**	**91.8**	**47.8**
Escherichia cryptic clade II MOD1-EC7253	99.2	**92.0**	**91.7**	**91.0**	**45.5**
Escherichia cryptic clade III E4694	99.7	96.6	96.5	96.3	70.8
Escherichia cryptic clade VI UHCL_3L	**98.4**	**91.6**	**91.7**	**91.3**	**45.9**

Strain OPT1704^T showed 98.7-98.9% 16S rRNA sequence similarity to *E. coli* ATCC 11775^T, *E. fergusonii* ATCC 35469^T and *E. marmotae* HT073016^T, which would not warrant assignment to a novel species based on the current threshold for species delineation (less than 98.7% sequence similarity[32]). However, the threshold for species delineation on the basis of 16S rRNA sequence has changed often and thresholds of up to 99% sequence similarity have been proposed previously[33]. In contrast, ANI analysis and dDDH did support assignment of strain OPT1704^T to a novel species, together with the representative strain of *Escherichia* cryptic clade III (table 2). The analyses also suggested that strain OPT1704^T falls within the *Escherichia* genus. This novel species, encompassing both *Escherichia* cryptic clades III and IV, was assigned *E. ruysiae* sp. nov. with strain OPT1704^T as the proposed type strain.

Assigning a novel species to a particular genus is challenging and currently no clear guidelines exist. Several approaches have been proposed, such as phylogenomics[32] or

counting shared genes[34]. Our phylogenomic analyses show OPT1704T clusters closely with other *Escherichia* spp., and clusters further away from *E. hermannii* and *S. enterica*. This clustering pattern is well supported by the bootstrapping analysis, for both the alignment 16S rRNA genes and 682 concatenated core genes. Another commonly used approach is calculating the percentage of conserved proteins (POCP[34]). Strain OPT1704T had the highest POCP with organisms in the *Escherichia* genus, clearly above the 'universal' cut-off of 50% shared proteins. This cut-off seems to be inadequate for Enterobacterales: *Salmonella enterica* and *Escherichia hermannii* showed a POCP of more than 60% with all true *Escherichia* strains (table S1). As the original POCP approach was developed using only 17 genera and has not been verified for Enterobacterales, the high rate of genetic exchange in this order[35] may necessitate an alternative POCP cut-off to the previously proposed cut-off of 50%.

To gain a better understanding of the *Escherichia* genus, we produced two phylogenies, based on 16S rRNA sequence (Fig. 1) and on an alignment of 682 core genes (Fig. 2). In short, rRNA genes were predicted from whole genomes using barrnap (version 0.9, https://github.com/tseemann/barrnap) and a tree was generated using FastTree (version 2.1.10[36]). For the core gene alignment, genomes were first annotated with Prokka (version 1.14.0[37]) and a core gene alignment was produced using Roary (version 3.12.0[38]) and MAFFT (version 7.307[39]). The phylogeny was inferred using a generalised time reversible model using base frequencies from the SNP alignment and free rate heterogeneity (GTR+F+R4 model) in IQ-tree (version 1.6.6[40]), as advised by ModelFinder[41]. Phylogenies were rooted on the *Salmonella enterica* serovar Typhimurium str. LT2^T genome. Both phylogenies showed that strain OPT1704^T clusters closely with the strain MOD1-EC7259 from *Escherichia* cryptic clade III, and away from the current *Escherichia* species. Comparing the number of SNPs extracted from the core genome of the strains included in table 2 showed the same results, as the two strains with the smallest number of SNPs were strains OPT1704T (clade IV) and MOD1-EC7259 (clade III, table S2).

Chun *et al*.[32] proposed that strains with >95-96% genome-wide ANI between each other should be assigned to the same species. If cryptic clades III and clade IV would share >95-96% ANI, this would mean both clades should be assigned to the same novel species, *E. ruysiae*. To assess this for a larger number of strains than the type strains presented in table 2, we downloaded all 65 available WGS from clade III and clade IV strains from Enterobase and compared ANI between all genomes using fastANI (version 1.1[28]). This analysis revealed that within 33 clade III genomes, the median ANI is 98.6% (range: 97.7%-99.9%), while within 32 clade IV genomes, the median ANI is 98.9% (range: 98.6%-99.9%, table S3). Between clade III and clade IV genomes, the median ANI is 96.5% (range 96.1%-96.8%). This suggests clades III and IV should be assigned to the same novel species, *E. ruysiae* sp. nov. Subsequently, based on ANI analysis we selected 10 representative

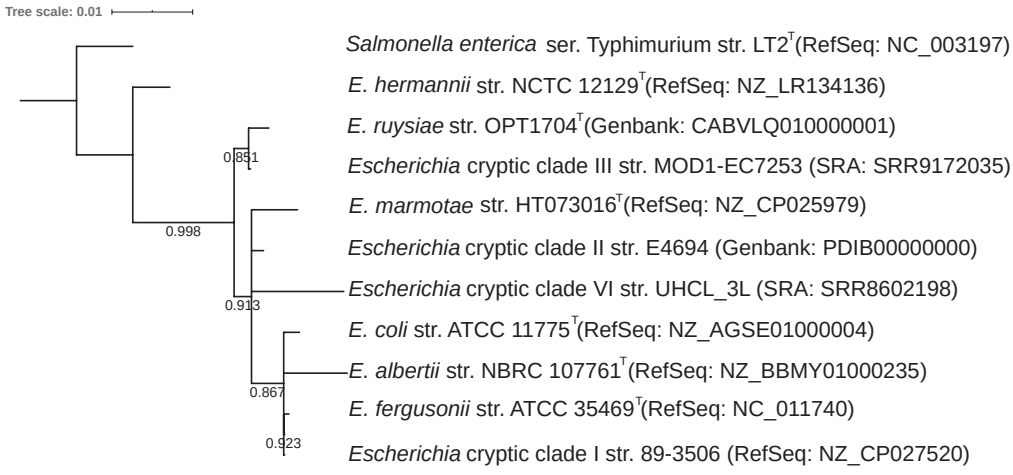


Figure 1. 16S phylogeny of *E. ruysiae* str. OPT1704^T with type strains of other *Escherichia spp.*, other *Escherichia* cryptic clades and *Salmonella enterica* serovar Typhimurium as outgroup. Numbers indicate bootstraps on a scale of 0 to 1. Phylogeny available at https://itol.embl.de/tree/14511722611226771596704407.

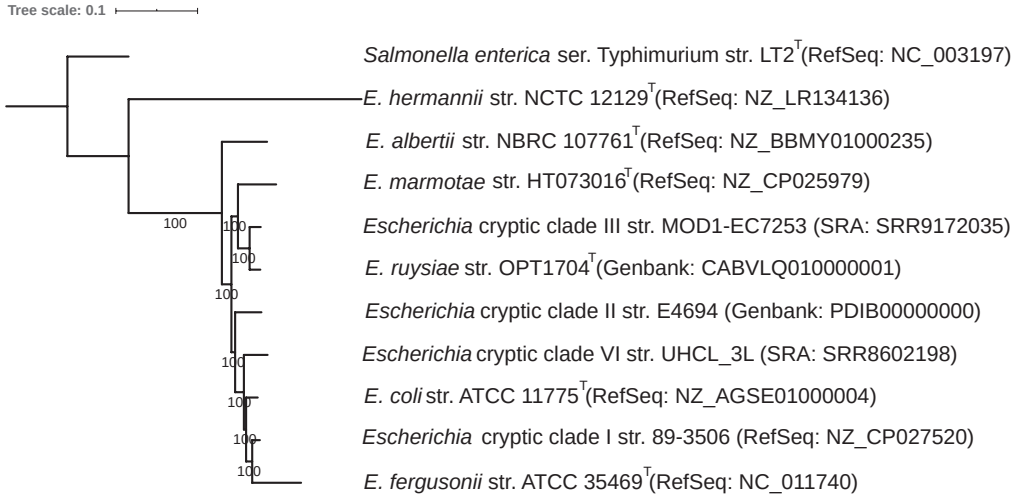


Figure 2. Phylogeny based on 682 concatenated core genes including *E. ruysiae* str. OPT1704T with type strains of other *Escherichia spp.*, other *Escherichia* cryptic clades and *Salmonella enterica* serovar Typhimurium as outgroup. Numbers indicate bootstraps on a scale of 0 to 100. Phylogeny available at https://itol.embl.de/tree/14511722611160731596708541.

clade IV genomes and 10 representative clade III genomes to be analysed using the TYGS platform[42]. The TYGS platform employs dDDH estimation and 16S and core genome phylogenetics to define species and subspecies within a given set of genomes, with an

upload limit of 20 user-provided genomes. The TYGS analysis also indicated that clade III and clade IV should be assigned to a single species, but could be delineated into two separate subspecies (table S4). Currently, no IJSEM guidelines exist for the delineation of subspecies based on genomic data. However, *E. ruysiae* could potentially be delineated further into two subspecies (representing the current clades III and IV, respectively) in the future, after a type strain for cryptic clade III has been identified. In the meantime, we propose to term clades III and IV genomic lineages of *E. ruysiae* sp. nov.

Finally, we annotated the genomes of the strains provided in table 2 using the EggNOG database[43] and extracted categories of cluster of orthologous genes (COG categories). Strain OPT1704^T did not encode a different profile of COG categories compared to other *Escherichia* type strains (table S5). Possibly, a gene ontology analysis which includes more genomes from all species might elucidate the different functional profiles, but this is out of the scope of the current study.

Based on phenotypic and genotypic data presented above, the niche for *E. ruysiae* sp. nov. cannot be exactly defined yet. Although OPT1704^T was isolated from human faeces, earlier studies have indicated that strains belonging to *E. ruysiae* sp. nov. do not adhere well to human-derived cell lines[44]. This finding is highlighted by the fact that we could not detect OPT1704^T anymore using ESBL microarray a month after we first detected it, although this might also be caused by loss of the ESBL gene. In conclusion, it seems that the human gut is not the primary niche for *E. ruysiae* sp. nov.

Description of *Escherichia ruysiae* sp. nov.

Escherichia ruysiae (ruy'si.ae N.L. gen. n. *ruysiae* named after Anna Charlotte Ruys, professor of microbiology at the University of Amsterdam from 1940 to 1969). Cells are Gram-negative, facultatively anaerobic, non-sporulating, non-motile rods with a size of approximately 1 by 2 µm. Colonies are circular, convex, grey-white and semi-transparent when grown overnight at 37 °C on COS sheep blood agar plates. The species is catalase-positive and oxidase-negative and grows at temperatures between 20 and 42 °C and NaCl concentrations between 0% and 6% w/v. In the VITEK2 GN biochemical test set it yields a positive result for Beta-Galactosidase, D-Glucose, D-Maltose, D-Mannitol, D-Mannose, D-Sorbitol, D-Trehalose, Saccharose/Sucrose, D-Tagatose, Gamma-Glutamyl-Transferase, Fermentation Glucose, Tyrosine Arylamidase, Succinate Alkalinisation, Alpha-Galactosidase, Ornithine Decarboxylase, Courmarate, Beta-Glucoronidase, 0/129 Resistance (Comp.Vibrio.) and Ellman and negative for Ala-Phe-Pro-Arylamidase, Adonitol, L-Pyrrolydonyl-Arylamidase, L-Arabitol, D-Cellobiose, H_2S Production, Beta-N-Acetyl Glucosaminidase, Glutamyl Arylamidase Pna, Beta-Glucosidase, Beta-Xylosidase,

Beta-Alanine Arylamidase Pna, L-Proline Arylamidase, Lipase, Palatinose, Urease, Citrate (Sodium), Malonate, 5-Keto-D-Gluconate, L-Lactate Alkalinisation, Alpha-Glucosidase, Beta-N-Acetyl-Galactosaminidase, Phosphatase, Glycine Arylamidase, Lysine Decarboxylase, L-Histidine Assimilation, Glu-Gly-Arg-Arylamidase, L-Malate Assimilation and L-Lactate Assimilation (table S6).

The type strain, OPT1704^T (= NCCB 100732^T = NCTC 14359^T), was isolated from faecal material of an international traveller returning from Asia.

The 16S rRNA sequence is deposited in ENA under accession LR745848. Raw Illumina and Nanopore whole-genome sequencing data, as well as the complete genome assembly are deposited under project PRJEB34275.

Author statements

Funding information

The COMBAT study was funded by Netherlands Organization for Health, Research and Development (ZonMw; 50-51700-98-120) and EU-H2020 programme (COMPARE, 643476).

Acknowledgements

The authors would like to thank Rob Weijts and Patricia Brinke for their help in phenotypic characterization of type strain OPT1704^T of *Escherichia ruysiae* sp. nov., and Arie van der Ende, Thomas Roodsant and Kees van der Ark for the helpful discussions. We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster.

The members of the COMBAT consortium, in alphabetical order: Maris S. Arcilla, Martin C.J. Bootsma, Perry J. van Genderen, Abraham Goorhuis, Martin Grobusch, Jarne M. van Hattem, Menno D. de Jong, Damian C. Melles, Nicky Molhoek, Astrid M.L. Oude Lashof, John Penders, Constance Schultsz, Ellen E. Stobberingh, Henri A. Verbrugh.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

The COMBAT study was approved by the Medical Research Ethics Committee, Maastricht University Medical Centre (METC 12-4-093). All participants provided written informed consent.

References

1. Castellani A, Chambers AJ. Manual of tropical medicine. *William Wood* 1919
2. Brenner DJ, Davis BR, Steigerwalt AG, Riddle CF, McWhorter AC *et al.* Atypical biogroups of *Escherichia coli* found in clinical specimens and description of Escherichia hermannii sp. nov. *J Clin Microbiol* 1982; 15:703–713
3. Farmer JJ, Fanning GR, Davis BR, O'Hara CM, Riddle C *et al. Escherichia fergusonii* and *Enterobacter taylorae*, two new species of Enterobacteriaceae isolated from clinical specimens. *J Clin Microbiol* 1985; 21:77–81
4. Huys G, Cnockaert M, Janda JM, Swings J. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *Int J Syst Evol Microbiol* 2003; 53:807–810
5. Liu S, Jin D, Lan R, Wang Y, Meng Q *et al. Escherichia marmotae* sp. nov., isolated from faeces of Marmota himalayana. *Int J Syst Evol Microbiol* 2015; 65:2130–2134
6. Hata H, Natori T, Mizuno T, Kanazawa I, Eldesouky I *et al.* Phylogenetics of family Enterobacteriaceae and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. *Microbiol Immunol* 2016; 60:303–311
7. Alnajar S, Gupta RS. Phylogenomics and comparative genomic studies delineate six main clades within the family Enterobacteriaceae and support the reclassification of several polyphyletic members of the family. *Infection, Genetics and Evolution* 2017; 54:108–127
8. Priest FG, Barker M. Gram-Negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *Int J Syst Evol Microbiol* 2010; 60:828–833
9. Tamura K, Sakazaki R, Kosako Y, Yoshizaki E. *Leclercia adecarboxylata* gen. nov., comb. nov., formerly known as *Escherichia adecarboxylata*. *Curr Microbiol* 1986; 13:179–184
10. Liu S, Feng J, Pu J, Xu X, Lu S *et al.* Genomic and molecular characterisation of *Escherichia marmotae* from wild rodents in Qinghai-Tibet plateau as a potential pathogen. *Sci Rep* 2019; 9:1–9
11. Ooka T, Seto K, Kawano K, Kobayashi H, Etoh Y *et al.* Clinical Significance of *Escherichia albertii*. *Emerg Infect Dis* 2012; 18:488–492
12. Russo T, Johnson JR. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes Infect* 2003; 5:449–456
13. Savini V, Catavitello C, Talia M, Manna A, Pompetti F *et al.* Multidrug-Resistant *Escherichia fergusonii*: a case of acute cystitis. *J Clin Microbiol* 2008; 46:1551–1552
14. Walk ST. The "Cryptic" *Escherichia*. *EcoSal Plus* 2015; 6:
15. Gangiredla J, Mammel MK, Barnaba TJ, Tartera C, Gebru ST *et al.* Draft genome sequences of *Escherichia albertii*, *Escherichia fergusonii*, and strains belonging to six cryptic lineages of *Escherichia* spp. *Genome Announc* 2018; 6:e00271–18
16. Arcilla MS, van Hattem JM, Haverkate MR, Bootsma MCJ, van Genderen PJJ *et al.* Import and spread of extended-spectrum ß-lactamase-producing Enterobacteriaceae by international travellers (combat study): a prospective, multicentre cohort study. *Lancet Infect Dis* 2017; 17:78–85
17. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018; 34:i884–i890
18. 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

19. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA *et al.* The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 2013; 57:3348–3357

20. Chen L. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res* 2004; 33:D325–D328

21. Sundin GW, Bender CL. Ecological and genetic analysis of copper and streptomycin resistance in *Pseudomonas syringae* pv. syringae. *Appl Environ Microbiol* 1993; 59:1018–1024

22. Ojdana D, Sieńko A, Sacha P, Majewski P, Wieczorek P *et al.* Genetic basis of enzymatic resistance of *E. coli* to aminoglycosides. *Adv Med Sci* 2018; 63:9–13

23. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* 2019; 47:D351–D360

24. The European Committee on Antimicrobial Susceptibility Testing Breakpoint tables for interpretation of MICs and zone diameters Version 9; 2019

25. Leber AL. Clinical Microbiology Procedures Handbook John Wiley & Sons; 2020

26. Abbott SL, O'Connor J, Robin T, Zimmer BL, Janda JM. Biochemical properties of a newly described *Escherichia* species, *Escherichia albertii. J Clin Microbiol* 2003; 41:4852–4854

27. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Achtman M *et al.* The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 2020; 30:138–152

28. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018; 9:5114

29. Yoon S-H, Ha S-M, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek* 2017; 110:1281–1286

30. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr* 2016; 4:e1900v1

31. Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 2013; 14:60

32. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* 2018; 68:461–466

33. Kim M, Oh H-S, Park S-C, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014; 64:346–351

34. Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C *et al.* A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* 2014; 196:2210–2215

35. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* 2020; 11:3602

36. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; 5:e9490

37. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30:2068–2069

38. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015; 31:3691–3693

39. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; 30:772–780

40. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; 32:268–274

41. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; 14:587–589

42. Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat Commun* 2019; 10:2182

43. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; 47:D309–D314

44. Vignaroli C, Di Sante L, Magi G, Luna GM, Di Cesare A *et al.* Adhesion of marine cryptic *Escherichia* isolates to human intestinal epithelial cells. *Isme J* 2015; 9:508–515

Supplementary material

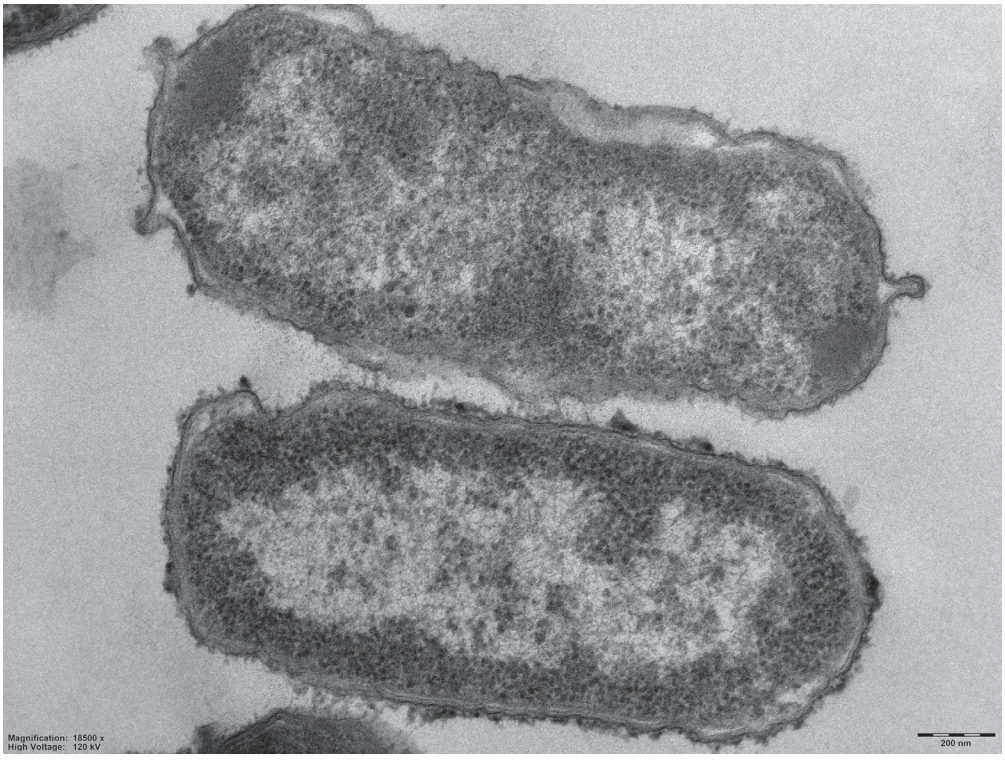


Figure S1. Transmission electron micrograph of *E. ruysiae* str. OPT1704T at a magnification of 18,500×.

Table S1. Percentage of conserved proteins (POCP) between type strains and representative strains spanning the *Escherichia* genus, including *Salmonella enterica* as outgroup. POCP was calculated according to Qin *et al.* (2014).

	E. albertii str. NBRC 107761^T (RefSeq: NZ_BBMY01000235)	*E. coli* str. ATCC11775^T (RefSeq: NZ_AGSE01000004)	*E. fergusonii* str. ATCC 35469^T (RefSeq: NC_011740)	*E. hermannii* str. NCTC 12129^T (RefSeq: NZ_LR134136)	*E. marmotae* str. HT073016^T (RefSeq: NZ_CP025979)	*E. ruysiae* str. OPT1704^T (Genbank: CABVLQ010000001)	*Escherichia* cryptic clade I str. 89-3506 (RefSeq: NZ_CP027520)	*Escherichia* cryptic clade II str. E4694 (Genbank: PDIB00000000)	*Escherichia* cryptic clade III str. MOD1-EC7253 (SRA: SRR9172035)	*Escherichia* cryptic clade VI str. UHCL_3L (SRA: SRR8602198)	*Salmonella enterica* ser. Typhimurium str. LT2^T (RefSeq: NC_003197)
E. albertii str. NBRC 107761^T (RefSeq: NZ_BBMY01000235)	100.0%	84.6%	78.4%	65.9%	81.8%	81.5%	85.1%	83.0%	81.7%	79.1%	75.8%
E. coli str. ATCC11775^T (RefSeq: NZ_AGSE01000004)	75.6%	100.0%	76.5%	64.6%	79.6%	81.1%	87.8%	81.3%	79.6%	75.6%	72.7%
E. fergusonii str. ATCC 35469^T (RefSeq: NC_011740)	74.4%	81.9%	100.0%	67.0%	80.4%	78.4%	82.2%	81.6%	79.2%	74.7%	75.6%
E. hermannii str. NCTC 12129^T (RefSeq: NZ_LR134136)	67.0%	73.8%	71.6%	100.0%	71.4%	72.3%	73.6%	72.1%	71.6%	67.2%	72.3%
E. marmotae str. HT073016^T (RefSeq: NZ_CP025979)	75.1%	83.7%	77.9%	63.3%	100.0%	82.7%	84.1%	80.5%	80.3%	77.0%	73.9%
E. ruysiae str. OPT1704^T (Genbank: CABVLQ010000001)	76.8%	86.0%	77.9%	67.0%	84.6%	100.0%	86.5%	84.9%	89.1%	76.3%	75.5%
Escherichia cryptic clade I str. 89-3506 (RefSeq: NZ_CP027520)	72.5%	83.4%	72.9%	60.8%	75.7%	78.1%	99.9%	77.9%	76.2%	72.1%	70.6%
Escherichia cryptic clade II str. E4694 (Genbank: PDIB00000000)	76.6%	85.0%	79.5%	65.1%	82.6%	83.2%	85.1%	100.0%	84.1%	75.7%	74.5%
Escherichia cryptic clade III str. MOD1-EC7253 (SRA: SRR9172035)	81.3%	89.7%	83.6%	69.8%	88.8%	93.9%	90.0%	90.5%	100.0%	80.5%	80.4%
Escherichia cryptic clade VI str. UHCL_3L (SRA: SRR8602198)	75.6%	81.2%	74.8%	62.6%	78.7%	77.3%	82.0%	78.1%	76.4%	100.0%	72.7%
Salmonella enterica ser. Typhimurium str. LT2^T (RefSeq: NC_003197)	67.7%	73.5%	71.9%	63.8%	72.3%	72.6%	74.7%	71.6%	72.2%	68.4%	100.0%

Table S2. SNP distances between type strains and representative strains spanning the *Escherichia* genus. SNPs were extracted from a 220 kbp core gene alignment produced with Roary and MAFFT. A darker cell colour indicates a higher number of SNPs separating the respective strains.

	E. albertii str. NBRC 107761^T (RefSeq: NZ_BBMY01000235)	*E. coli* str. ATCC11775^T (RefSeq: NZ_AGSE01000004)	*E. fergusonii* str. ATCC 35469^T (RefSeq: NC_011740)	*E. hermannii* str. NCTC 12129^T (RefSeq: NZ_LR134136)	*E. marmotae* str. HT073016^T (RefSeq: NZ_CP025979)	*E. ruysiae* str. OPT1704^T (Genbank: CABVLQ010000001)	*Escherichia* cryptic clade I str. 89-3506 (RefSeq: NZ_CP027520)	*Escherichia* cryptic clade II str. E4694 (Genbank: PDIB00000000)	*Escherichia* cryptic clade III str. MOD1-EC7253 (SRA: SRR9172035)	*Escherichia* cryptic clade VI str. UHCL_3L (SRA: SRR8602198)	*Salmonella enterica* ser. Typhimurium str. LT2^T (RefSeq: NC_003197)
E. albertii str. NBRC 107761^T (RefSeq: NZ_BBMY01000235)	0	12755	15901	24572	14213	13402	13000	12872	13380	13825	26287
E. coli str. ATCC11775^T (RefSeq: NZ_AGSE01000004)	12755	0	11564	24022	11921	9379	5363	8892	9659	7276	25772
E. fergusonii str. ATCC 35469^T (RefSeq: NC_011740)	15901	11564	0	24504	15649	13805	10062	13048	13991	13010	26471
E. hermannii str. NCTC 12129^T (RefSeq: NZ_LR134136)	24572	24022	24504	0	24460	24038	24035	23970	24133	24273	23490
E. marmotae str. HT073016^T (RefSeq: NZ_CP025979)	14213	11921	15649	24460	0	10759	11967	11768	10588	12944	26531
E. ruysiae str. OPT1704^T (Genbank: CABVLQ010000001)	13402	9379	13805	24038	10759	0	9503	10267	4811	10714	25829
Escherichia cryptic clade I str. 89-3506 (RefSeq: NZ_CP027520)	13000	5363	10062	24035	11967	9503	0	9351	9777	7566	25869
Escherichia cryptic clade II str. E4694 (Genbank: PDIB00000000)	12872	8892	13048	23970	11768	10267	9351	0	10249	10422	25757
Escherichia cryptic clade III str. MOD1-EC7253 (SRA: SRR9172035)	13380	9659	13991	24133	10588	4811	9777	10249	0	10862	25800
Escherichia cryptic clade VI str. UHCL_3L (SRA: SRR8602198)	13825	7276	13010	24273	12944	10714	7566	10422	10862	0	26159
Salmonella enterica ser. Typhimurium str. LT2^T (RefSeq: NC_003197)	26287	25772	26471	23490	26531	25829	25869	25757	25800	26159	0

Table S3. FastANI analysis for *Escherichia* cryptic clades III and IV (*E. ruysiae* sp. nov.)

This table (4161 rows by 8 columns) is available from Figshare (https://tinyurl.com/1ebef) or with the online version of this article.

Table S4. TYGS results for *E. ruysiae* sp. nov. strains. Ten uploaded genomes belong to clade III, while the other ten belong to clade IV. Genome names correspond to the assembly barcodes from Enterobase.

This table (21 rows by 18 columns) is available from Figshare (https://tinyurl.com/1ebef) or with the online version of this article.

Table S5. Gene ontology analysis of type strains and other representative strains spanning the *Escherichia* genus, with *Salmonella enterica* included as outgroup. Functional categories are based on the EggNOG database and ordered according to higher-level functions. J: Translation, ribosomal structure and biogenesis; A: RNA processing and modification; K: Transcription; L: Replication, recombination and repair; D: Cell cycle control, cell division, chromosome partitioning; V: Defense mechanisms; T: Signal transduction mechanisms; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; U: Intracellular trafficking, secretion, and vesicular transport; O: Posttranslational modification, protein turnover, chaperones; C: Energy production and conversion; G: Carbohydrate transport and metabolism; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; H: Coenzyme transport and metabolism; I: Lipid transport and metabolism; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; S: Function unknown

This table (13 rows by 24 columns) is available from Figshare (https://tinyurl.com/1ebef) or with the online version of this article.

Table S6. Output from the biochemical VITEK2 analysis using the GN card.

Well	Test	Mnemonic	Result
2	Ala-Phe-Pro-ARYLAMIDASE	APPA	-
3	ADONITOL	ADO	-
4	L-Pyrrolydonyl-ARYLAMIDASE	PyrA	-
5	L-ARABITOL	lARL	-
7	D-CELLOBIOSE	dCEL	-
9	BETA-GALACTOSIDASE	BGAL	+
10	H2S PRODUCTION	H2S	-
11	BETA-N-ACETYL GLUCOSAMINIDASE	BNAG	-
12	Glutamyl Arylamidase PNA	AGLTp	-
13	D-GLUCOSE	dGLU	+
14	GAMMA-GLUTAMYL-TRANSFERASE	GGT	+
15	FERMENTATION GLUCOSE	OFF	+
17	BETA-GLUCOSIDASE	BGLU	-
18	D-MALTOSE	dMAL	+
19	D-MANNITOL	dMAN	+
20	D-MANNOSE	dMNE	+
21	BETA-XYLOSIDASE	BXYL	-
22	BETA-Alanine arylamidase pNA	BAlap	-
23	L-Proline ARYLAMIDASE	ProA	-
26	LIPASE	LIP	-
27	PALATINOSE	PLE	-
29	Tyrosine ARYLAMIDASE	TyrA	+
31	UREASE	URE	-
32	D-SORBITOL	dSOR	+
33	SACCHAROSE/SUCROSE	SAC	-
34	D-TAGATOSE	dTAG	-
35	D-TREHALOSE	dTRE	+
36	CITRATE (SODIUM)	CIT	-
37	MALONATE	MNT	-
39	5-KETO-D-GLUCONATE	5KG	-
40	L-LACTATE alkalinisation	lLATk	-
41	ALPHA-GLUCOSIDASE	AGLU	-
42	SUCCINATE alkalinisation	SUCT	+
43	Beta-N-ACETYL-GALACTOSAMINIDASE	NAGA	-
44	ALPHA-GALACTOSIDASE	AGAL	+
45	PHOSPHATASE	PHOS	-
46	Glycine ARYLAMIDASE	GlyA	-

47	ORNITHINE DECARBOXYLASE	ODC	+
48	LYSINE DECARBOXYLASE	LDC	-
53	L-HISTIDINE assimilation	IHISa	-
56	COURMARATE	CMT	+
57	BETA-GLUCORONIDASE	BGUR	+
58	0/129 RESISTANCE (comp.vibrio.)	O129R	+
59	Glu-Gly-Arg-ARYLAMIDASE	GGAA	-
61	L-MALATE assimilation	IMLTa	-
62	ELLMAN	ELLM	+
64	L-LACTATE assimilation	ILATa	-

Supplementary information. MALDI-TOF spectrum for strain OPT1704^T and full length 16S rRNA gene alignment used for figure 1 and table 2.

This file is available from Figshare (https://tinyurl.com/1ebef) or with the online version of this article.

Part II

Escherichia coli adapted to
Human Intestinal Colonisation

Chapter 4

Extraintestinal pathogenic *Escherichia coli* (ExPEC) are associated with prolonged carriage of extended-spectrum ß-lactamase-producing *E. coli* acquired during travel

Boas C.L. van der Putten, Jarne M. van Hattem, John Penders, COMBAT Consortium, Daniel R. Mende, Constance Schultsz

Submitted for publication
bioRxiv, https://doi.org/10.1101/2020.09.23.309856

Abstract

Objectives Extended-spectrum ß-lactamase-producing *Escherichia coli* (ESBL-Ec) are frequently acquired during international travel, contributing to the global spread of antimicrobial resistance. Human-adapted ESBL-Ec are predicted to exhibit increased intestinal carriage duration, resulting in a higher likelihood of onward human-to-human transmission. Yet, bacterial determinants of increased carriage duration are unknown. Previous studies analysed small traveler cohorts, with short follow-up times, or did not employ high-resolution molecular typing, and were thus unable to identify bacterial traits associated with long-term carriage.

Methods In a prospective cohort study of 2001 international travelers, we analysed 160 faecal ESBL-Ec isolates from all 38 travelers who acquired ESBL-Ec during travel and subsequently carried ESBL-Ec for at least 12 months after return, by whole-genome sequencing. For 17 travelers, we confirmed the persistence of ESBL-Ec strains through single nucleotide variant typing. To identify determinants of increased carriage duration, we compared the 17 long-term carriers (≥12 months carriage) with 33 age-, sex- and destination-matched short-term carriers (<1 month carriage). Long-read sequencing was employed to investigate ESBL plasmid persistence.

Results We show that in healthy travelers with very low antibiotic usage, extraintestinal pathogenic lineages of *E. coli* (ExPEC) are significantly more likely to persist than other *E. coli* lineages. The long-term carriage of *E. coli* from ExPEC lineages is mainly driven by sequence type 131 and phylogroup D *E. coli*.

Conclusions Although ExPEC frequently cause extra-intestinal infections such as bloodstream infections, our results imply that ExPEC are also efficient intestinal colonizers, which potentially contributes to their onward transmission.

Introduction

International travel contributes significantly to the spread of extended-spectrum ß-lactamase (ESBL) gene positive *Escherichia coli* (ESBL-Ec)[1,2]. Genetically diverse ESBL-Ec are frequently acquired during international travel[3–5]. Whilst travel-acquired ESBL-Ec typically are lost during travel or within the first month after return (Figure 1A)[6], ESBL-Ec and ESBL genes have been detected for more than 12 months after return in a proportion of travelers[1,3–5,7]. ESBL genes can persist through at least two different mechanisms[8]. Bacterial strains that carry ESBL-encoding genes in their chromosome or on a stable plasmid can persist over time as part of the local microbiome (strain persistence, Figure 1B). In addition, ESBL genes can be located on mobile genetic elements (MGEs) which persist in the microbiome by their ability to transfer between different bacterial hosts (MGE persistence, Figure 1C).

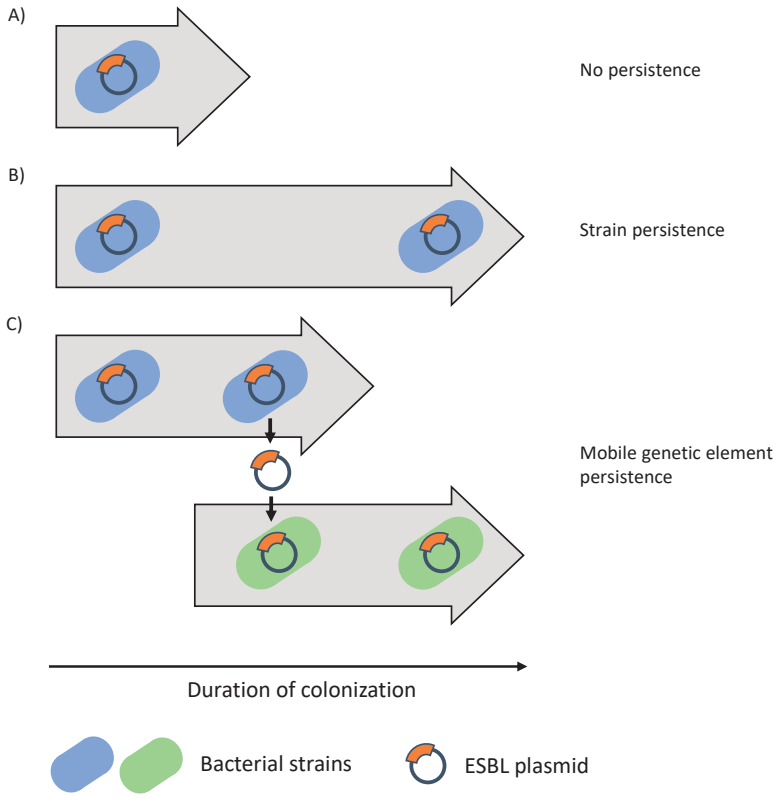


Figure 1. Schematic overview of main persistence mechanism of ESBL genes in the gut of returning travelers. Gray arrows indicate carriage duration over time. A) No persistence, where the colonizing strain is lost during the follow-up period. B) Strain persistence, where a bacterial strain harboring an ESBL gene is continuously present. C) Persistence through mobile genetic elements, where the original strain harboring the ESBL gene has been lost, but the ESBL gene (located on a mobile genetic element) has been passed on to another strain.

ESBL-Ec lineages which are capable of long-term colonization and are adapted to the human intestinal tract are likely to contribute to onward transmission of ESBL-Ec and ESBL genes. However, it is unknown which ESBL-Ec lineages are capable of persistence after return from travel, due to a lack of studies with a sufficiently large sample size and a prospective longitudinal study design. Additionally, high-resolution typing methods such as whole-genome sequencing (WGS) are needed to reliably determine whether the strains have been carried over a long period. One study employed WGS to investigate persistence in 16 travelers who acquired ESBL-Ec abroad and showed that only one traveler carried a travel-acquired ESBL-Ec strain for at least 7 months[7]. A very recent study analysed data from 11 travellers which were colonised for >3 months[9]. Due to the low sample sizes in both studies, few ESBL-Ec attributes could be identified that were significantly associated with long-term carriage. Armand-Lefèvre *et al*. reported an association between phylogroups B2/D/F with prolonged carriage duration ($p = 0.02$) [9]. Other studies either did not employ WGS or focused on short-term carriage only.

We studied a cohort of 2001 Dutch international travelers (COMBAT cohort)[1]. Six hundred and thirty-three travelers acquired ESBL gene-positive Enterobacterales during travel abroad, of whom 38 travelers (6.0%) were colonized for ≥12 months after acquisition of ESBL gene-positive Enterobacterales, all of which were *E. coli*. Here we report on host and bacterial characteristics associated with ESBL genes persistence and the mechanism through which the ESBL genes persisted. We identified an association between extraintestinal pathogenic *E. coli* and long-term carriage.

Materials and methods

Travelers and isolates selection

We included all 38 travelers from the COMBAT cohort (N=2001) who were colonized for ≥12 months with *Escherichia coli* which possessed ESBL genes belonging to a single ESBL gene group at return from travel and at all subsequent time points (1, 3, 6, and 12 months after return), and who tested ESBL-negative before travel (see reference 1 for a detailed description of sampling and microbiological methods)[1]. In short, faecal samples were inoculated in tryptic soy broth containing 50 mg/L vancomycin to select for Enterobacterales. After overnight incubation, the broth was subcultured on chromID ESBL agar plates (bioMérieux). Morphologically different colonies were isolated, with a maximum of five isolates per faecal sample. Eighty-five ESBL-Ec isolates sampled at return from travel (T0) and 75 ESBL-Ec isolates sampled 12 months after return from travel (T12) were included in the current study (Figure 2).

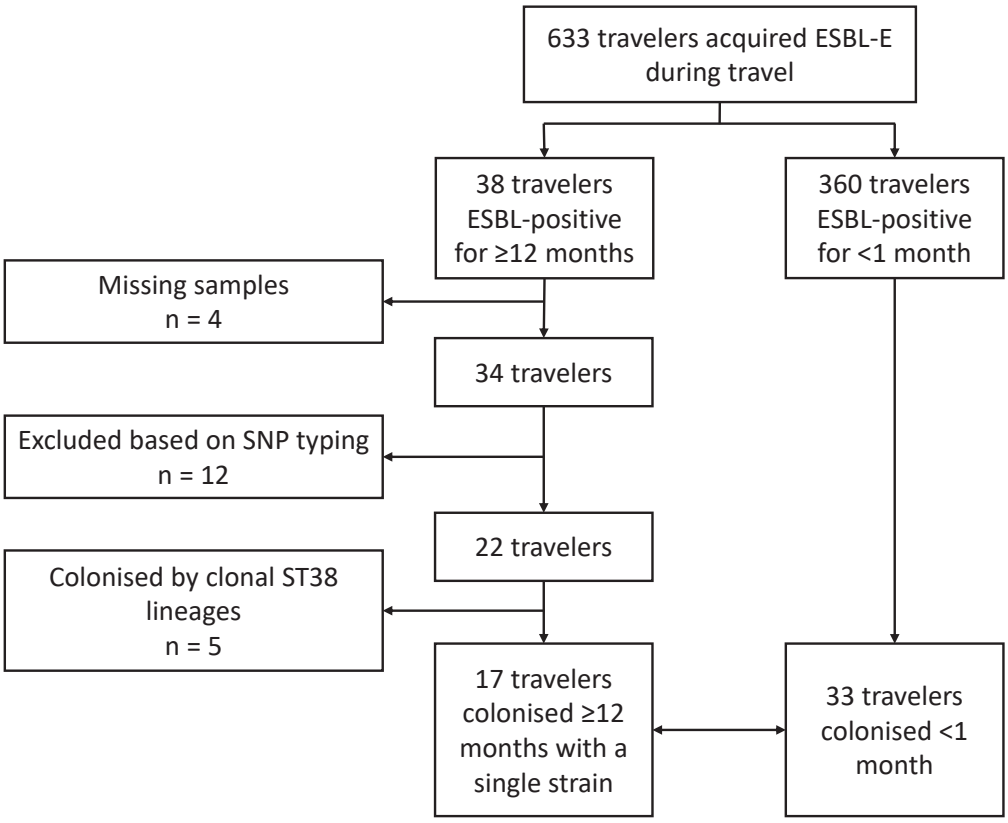


Figure 2. Study flowchart. Strain persistence was defined as isolates from a single traveler, sample twelve months apart with identical ESBL alleles, identical MLST profiles and 25 or fewer genome-wide SNPs difference. Clonal ST38 isolates were excluded. Seventeen long-term carriers (colonized ≥12 months with a single strain) were matched on sex, age and travel destination to thirty-three short-term carriers (colonized <1 month).

In silico typing

DNA extraction and library preparation were performed on all available isolates using the Qiagen Blood and Tissue DNA extraction kit (Cat No./ID: 69506) and Kapa HTP library prep kit (Kit Code KK8234), respectively. Whole-genome sequencing was performed on Illumina HiSeq 2500 by the Amsterdam UMC Core Facility Genomics. Sequencing data were analysed using a Snakemake v5.7.110 pipeline, available at https://github.com/boasvdp/COMBAT (v1.1.0 archived at https://doi.org/10.5281/zenodo.4582689). In short, Illumina sequencing data were trimmed using fastp v0.20.0[11], assembled using the Shovill wrapper v1.0.9 (https://github.com/tseemann/shovill) for SPAdes[12], and resistance genes were identified using AMRfinderplus v3.2.3[13]. *E. coli* phylogroups were predicted using EzClermont v0.4.3[14]. Multi-locus sequence typing (MLST) was performed with the mlst script (https://github.com/tseemann/mlst), using the Achtman scheme for *Escherichia coli*[15].

Analysis of strain persistence

We defined strain persistence as the presence of one or more ESBL-Ec isolates in faecal samples from a single traveler obtained at return from travel (T0), as well as 12 months thereafter (T12), with a maximum of 25 SNPs per 5 Mbp difference in the core genome[16].

To assess strain persistence, we calculated single nucleotide polymorphism differences between isolates to identify which isolates belonged to the same strain. We mapped sequencing reads using Snippy v4.4.5 (https://github.com/tseemann/snippy) on MLST-specific reference genomes selected with ReferenceSeeker v1.6.3[17] to obtain core genome alignments. Reference genomes used for all MLST sequence types, together with core genome alignment lengths, can be found in table S1. For each MLST, IQtree v1.6.12[18] was used to infer a phylogeny from the core genome alignment under the model advised by Modelfinder[19]. Recombination events in the core genome alignment were identified using ClonalFrameML v1.12[20] and masked using maskrc-svg v0.5 (https://github.com/kwongj/maskrc-svg). SNP differences were counted using snp-dists v0.7.0 (https://github.com/tseemann/snp-dists) and alignment lengths were calculated using a modified version of snp-dists v0.7.0 (https://github.com/boasvdp/snp-dists). SNP counts were scaled to 5 Mbp, to approximate the number of genome-wide SNPs[16].

To determine whether ESBL plasmids had persisted independent of bacterial host, we employed long read sequencing (Oxford Nanopore Technologies). We generated Oxford Nanopore Technologies sequencing data according to Van der Putten *et al.* (2020)[21]. In short, strains were grown overnight at 37 °C in liquid LB. DNA extraction and library preparation were performed using the Qiagen MagAttract HMW DNA Kit (Cat. No. 67563) and ONT native barcoding kit (Cat. No. EXP-NBD114), respectively. The library was subsequently sequenced on an ONT MinION flowcell. Raw read data was filtered using Filtlong v0.2.0 (https://github.com/rrwick/Filtlong) and assembled with corresponding Illumina data using Unicycler v0.4.8[22]. Quality control was implemented at several steps in the pipeline using FastQC v0.11.8[23], Quast v4.6.3[24], and MultiQC v1.6[25]. Plasmid comparison was performed using ANICalculator[26].

Comparison with short-term carriers

Seventeen long-term carriers were matched by age (range +/-7 years), sex, and travel destination (United Nations subregions) to thirty-three travelers who were colonized for less than a month using SPSS 26 (Figure 2). Illumina WGS was performed as described before on all ESBL-Ec isolated at return from travel from these short-term carriers (total 41 isolates).

Plotting and statistical analysis

Data were plotted using ggplot2 v3.1.1[27], ggthemes v2.4.0[28], and patchwork[29] in R v3.5.1. Tabular data were analysed using Pandas v0.24.2[30] in Python v3.6.7. Statistical analysis was performed using Fisher's exact test as implemented in the Python library SciPy[31].

Results

Data characteristics

Out of 2001 Dutch international travelers, we included 34 travelers with samples available, out of all 38 travelers whose faecal samples were positive for *E. coli* harbouring the same ESBL group gene at return (T0), one month after return from travel (T1) and 12 months after return from travel (T12), but were ESBL-negative before travel (Figure 2). We included a median of 1 ESBL-Ec isolate per traveler per timepoint (range: 1-5 isolates). SNP typing showed that acquired ESBL-Ec isolates were genetically diverse, with some travelers acquiring up to four distinct strains.

ESBL-Ec strain persistence

SNP analysis

Twenty-two out of 34 travelers harboured persistent strains based on SNP typing (Figure 2). Of the 12 travelers who did not harbour persistent strains, ten carried strains with differing MLST profiles between T0 and T12, and two carried strains from a single ST but with the number of core genome-wide SNP differences exceeding the threshold (25 SNPs for a 5 Mbp genome)[16].

Clonal ST38 lineages

Non-persistent, yet highly conserved strains that are widely disseminated within the human population could potentially be misidentified as persistent when using SNV analyses. Hence, we compared SNV distances between isolates obtained from unrelated travelers (i.e., belonging to different households) to identify strains shared across our study population. Conserved strains belonging to two ST38 lineages were identified in five unrelated travelers. Lineage ST38-$bla_{CTX\text{-}M\text{-}27}$ was identified in three unrelated travelers and lineage ST38-$bla_{CTX\text{-}M\text{-}14}$ in two other, unrelated travelers. All five unrelated travelers returned from different countries and continents, suggesting these two lineages have disseminated widely. Whole genomes from lineages ST38-$bla_{CTX\text{-}M\text{-}27}$ and ST38-$bla_{CTX\text{-}M\text{-}14}$ were also abundantly present in public data and a phylogenetic analysis of their core genomes confirmed their high similarity (Figure S1). Hence, we could not conclusively determine whether these highly clonal strains had persisted in the five travelers concerned, or whether these strains were re-acquired from an unknown source. Therefore,

we excluded the travelers harbouring these clonal ST38 strains from further analyses to identify lineages potentially associated with persistence. ST38 strains colonizing four additional travelers were not related to the ST38-$bla_{CTX\text{-}M\text{-}27}$, and ST38-$bla_{CTX\text{-}M\text{-}14}$ lineages and were thus not excluded from further analyses.

We identified two related travelers who harboured highly similar ST69 isolates (9 genome-wide SNPs). The most likely explanation of this observation is that the travellers acquired this strain from the same source or one from the other, and that it has persisted in both travellers since.

Plasmid analysis

For six travelers, we detected identical ESBL genes from isolates sampled at return from travel and twelve months thereafter, but we could exclude ESBL gene persistence in persistent isolates based on SNP typing (Table S1). To determine whether the plasmid carrying these ESBL genes had persisted independent of bacterial host (Figure 1B), we employed long read sequencing. For one traveler, we detected a pair of ESBL-Ec isolates belonging to different MLST types in which an almost identical plasmid (99.8% nucleotide identity) was identified at T0 and T12, suggesting ESBL gene persistence by plasmid transfer between different *E. coli* hosts.

In summary, we identified persistent ESBL-Ec strains in 17 out of 34 travelers with prolonged ESBL-Ec carriage and persistent ESBL-plasmid in one traveler (Table S1). Although we included multiple ESBL-Ec isolates for some travelers, we did not observe multiple persistent strains within any single traveler.

Comparison of ESBL-Ec from long-term and short-term carriers

For each long-term carrier (≥12 months carriage) harbouring a persistent strain, two short-term carriers (<1 month carriage) were matched by age, sex and travel destination. For one long-term carrier, only a single matching short-term carrier could be identified, resulting in a comparison of 17 isolates from 17 long-term carriers and 42 isolates from 33 matched short-term carriers, which were sequenced.

Antibiotic usage was low before, during, and after travel and similar between long-term and short-term carriers (Table 1). None of the travelers were admitted to the hospital during or after travel in either group. One single traveler returned to the same country as visited during index travel, within 12 months after return from index travel.

Table 1. Characteristics of long-term and short-term carriers. Matching was performed on sex, age and travel destination (United Nations subregions). These characteristics are depicted in italic. Data are presented as number (%) for all characteristics except age and travel duration, which is presented as median (IQR).

		Long-term carriers (n = 17)	Short-term carriers (n = 33)
Male		4 (23.5%)	8 (24.2%)
Age (years)		50.2 (41.7-59.4)	51.4 (36.5-58.1)
Continent visited during index travel	*Asia*	16 (94.1%)	31 (93.9%)
	North and South America	1 (5.9%)	2 (6.1%)
Travel duration (days)		20 (17-28)	19 (14-21)
Admitted to hospital during index travel		0 (0%)	0 (0%)
Travel to the same country as index travel, within 12 months after return from index travel		1 (5.9%)	0 (0%)
Antibiotic usage	Within 3 months before index travel	0 (0%)	0 (0%)
	During index travel	3 (17.6%)	2 (6.1%)
	Within 1 month after return from index travel	0 (0%)	0 (0%)
	Within 1 to 3 months after return from index travel	1 (5.9%)	2 (6.7%)
	Within 3 to 6 months after return from index travel	0 (0%)	2 (7.1%)
	Within 6 to 12 months after return from index travel	3 (17.6%)	2 (7.4%)
Traveler's diarrhoea		6 (35.3%)	9 (27.3%)

Following the recent definition of common extraintestinal pathogenic *E. coli* (ExPEC) lineages according to Manges *et al.* (2019)[32], persistent strains belonged significantly more often to ExPEC lineages than non-persistent strains. In 15 out of 17 long-term carriers, ExPEC were identified as the persistent strain while in only 7 out of 33 short-term carriers ExPEC were detected (odds ratio 27.86, 95% confidence interval: 5.11-151.74). This difference appears to be driven mostly by ST131 and phylogroup D strains.

Discussion

We demonstrated long-term carriage of travel-acquired ESBL-positive *Escherichia coli* in 17 travelers out of a cohort of 2001 travelers, which was driven by persistence of ESBL-Ec belonging to ExPEC lineages. In a single traveler, the persistence of ESBL-Ec seemed to be due to an ESBL plasmid which shifted between bacterial hosts after colonising the traveler's gut. Our study strengthens the finding of Armand-Lefèvre *et al.*[9] that ESBL-Ec lineage is associated with persistent carriage after acquisition during travel. Interestingly, we come to similar conclusions although we have focused on long-term carriage (≥12

months) as compared to the previous study (>3 month carriage). Earlier work could not make strong assertions about the persistence of strains, either due to the limited number of included travelers, because the typing methods employed were insufficiently discriminating between isolates, because of a limited duration of follow-up, or because the study population was not representative of community exposure[3–5,7]. In a recent study in Laos, a group of travelers attending a course at local hospitals acquired a very high diversity of ESBL-Ec immediately upon arrival and it was shown that this acquisition of resistant *E. coli* during travel is highly dynamic[6]. However, the short follow-up in this study did not allow analysis of long-term outcomes of the acquisition of ESBL-Ec.

Previous studies have shown that a limited number of *E. coli* lineages contribute to a large fraction of *E. coli*-mediated extraintestinal disease, including urinary tract infections and bloodstream infections[32]. These extraintestinal pathogenic *E. coli* are commonly referred to as ExPEC. While ExPEC display pathogenic potential, epidemiological studies often find ExPEC colonizing the human gut in the absence of symptoms[33,34]. In fact, it has been proposed that ExPEC have evolved towards particularly efficient intestinal colonization and their extraintestinal virulence is an evolutionary "byproduct"[35]. The long-term persistence of ExPEC lineages, acquired after relatively short travel duration as observed in our study (median 20 days, interquartile range: 17 – 29 days) suggests that ExPEC lineages have spread globally successfully due to their adaptation to the human intestinal tract. It should however be noted that we restricted our analysis to ESBL-producing *E. coli* and duration of persistence of susceptible ExPEC in this cohort is unknown.

Whole-genome sequencing allowed us to detect two clonal ST38 lineages which were shared between unrelated travelers. By additional analysis of publicly available WGS data, these ST38 lineages were shown to have spread globally. The extremely high degree of similarity within these ST38 lineages interfered with reliable strain identification based on core genome DNA sequence analysis. Future studies should explore the application of SNP typing using the pangenome rather than only the core genome, for example through software like Pandora[36]. These novel methods however currently display a higher error rate than the core genome analysis approach used in the present study[36]. Awareness of the circulation of ST38 or other rapidly expanding lineages that are indistinguishable by their core genome is important, for example for management of suspected outbreaks of ST38 in hospital settings.

Our current study focused solely on ESBL-producing *E. coli* and found an association between ExPEC carriage and increased carriage duration. Our approach could also be applied to non-ESBL-producing ExPEC, to determine whether acquisition of ExPEC in general, independent of their antibiotic susceptibility profile, is more likely to results in increased carriage duration after travel. The frequency of sampling can be considered

a limitation of this study. However, when comparing between two timepoints, twelve months apart, we could still identify very closely related isolates in samples from the same traveler, indicating strain persistence. Additionally, we only had isolates available for household members of four long-term carriers. Including more household members in future studies might allow us to estimate the likelihood of onward transmission following long-term carriage.

Applying genomic epidemiology to a large traveler cohort, we have shown that ESBL-positive *E. coli* acquired during travel are able to persist for more than a year. The strains that showed the longest carriage duration belonged predominantly to pathogenic ExPEC lineages. Our data imply that long-term carriage of resistant *E. coli* is governed by bacterial characteristics which are associated with lineage. This finding possibly allows a more precise risk assessment for international travelers returning with travel-acquired resistant *E. coli*.

Author statements

Funding information

The COMBAT study was funded by Netherlands Organization for Health, Research and Development (ZonMw; 50-51700-98-120) and EU-H2020 programme (COMPARE, 643476). BP was funded through an internal grant of the Amsterdam UMC ("flexibele OiO beurs").

Declaration of interests

We declare no competing interests.

Data and code availability

All Illumina and Oxford Nanopore Technologies sequencing data used in this study are currently available free of restrictions at NCBI under project accession number PRJEB40103. Metadata linking isolates to travelers, required to reproduce our analyses, are currently available free of restrictions at the GitHub repository of this project (https://www.github.com/boasvdp/COMBAT, v1.1.0 archived through Zenodo at https://doi.org/10.5281/zenodo.4582689).

All code is available free of restrictions under the MIT license at https://www.github.com/boasvdp/COMBAT (v1.1.0 archived through Zenodo at https://doi.org/10.5281/zenodo.4582689).

Ethical statement

The COMBAT study was approved by the Medical Research Ethics Committee, Maastricht

University Medical Centre (METC 12-4-093). All participants provided written informed consent.

Acknowledgements

The authors would like to thank Arie van der Ende for the helpful discussions. We thank SURFsara (https://www.surfsara.nl) for the support in using the Lisa Compute Cluster.

References

1. Arcilla MS, van Hattem JM, Haverkate MR, *et al*. Import and spread of extended-spectrum ß-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. *Lancet Infect Dis*. 2017;17(1):78–85. doi:10.1016/S1473-3099(16)30319-X

2. Ruppé E, Andremont A, Armand-Lefèvre L. Digestive tract colonization by multidrug-resistant Enterobacteriaceae in travellers: An update. *Travel Med Infect Dis*. 2018;21:28–35. doi:10.1016/j.tmaid.2017.11.007

3. Ruppé E, Armand-Lefèvre L, Estellat C, *et al*. High Rate of Acquisition but Short Duration of Carriage of Multidrug-Resistant Enterobacteriaceae After Travel to the Tropics. *Clin Infect Dis Off Publ Infect Dis Soc Am*. 2015;61(4):593–600. doi:10.1093/cid/civ333

4. Paltansing S, Vlot JA, Kraakman MEM, *et al*. Extended-spectrum ß-lactamase-producing enterobacteriaceae among travelers from the Netherlands. *Emerg Infect Dis*. 2013;19(8):1206–1213. doi:10.3201/eid.1908.130257

5. Pires J, Kuenzli E, Kasraian S, *et al*. Polyclonal Intestinal Colonization with Extended-Spectrum Cephalosporin-Resistant Enterobacteriaceae upon Traveling to India. *Front Microbiol*. 2016;7:1069. doi:10.3389/fmicb.2016.01069

6. Kantele A, Kuenzli E, Dunn SJ, *et al*. Dynamics of intestinal multidrug-resistant bacteria colonisation contracted by visitors to a high-endemic setting: a prospective, daily, real-time sampling study. *Lancet Microbe*. 2021;0(0). doi:10.1016/S2666-5247(20)30224-X

7. Bevan ER, McNally A, Thomas CM, Piddock LJV, Hawkey PM. Acquisition and Loss of CTX-M-Producing and Non-Producing *Escherichia coli* in the Fecal Microbiome of Travelers to South Asia. *mBio*. 2018;9(6). doi:10.1128/mBio.02408-18

8. Sheppard AE, Stoesser N, Wilson DJ, *et al*. Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene blaKPC. *Antimicrob Agents Chemother*. 2016;60(6):3767–3778. doi:10.1128/AAC.00464-16

9. Armand-Lefèvre L, Rondinaud E, Desvillechabrol D, *et al*. Dynamics of extended-spectrum beta-lactamase-producing Enterobacterales colonization in long-term carriers following travel abroad. *Microb Genomics*. 2021;7(7). doi:10.1099/mgen.0.000576

10. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2522. doi:10.1093/bioinformatics/bts480

11. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560

12. Bankevich A, Nurk S, Antipov D, *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012;19(5):455–477. doi:10.1089/cmb.2012.0021

13. Feldgarden M, Brover V, Haft DH, *et al*. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother*. 2019;63(11). doi:10.1128/AAC.00483-19

14. Waters NR, Abram F, Brennan F, Holmes A, Pritchard L. Easy phylotyping of *Escherichia coli* via the EzClermont web app and command-line tool. *Access Microbiol*. Published online 2020. doi:10.1099/acmi.0.000143

15. Wirth T, Falush D, Lan R, *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol*. 2006;60(5):1136–1151. doi:10.1111/j.1365-2958.2006.05172.x

16. Gorrie CL, Silva AGD, Ingle DJ, *et al*. Systematic analysis of key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria. *bioRxiv*. Published online September 25, 2020:2020.09.24.310821. doi:10.1101/2020.09.24.310821

17. Schwengers O, Hain T, Chakraborty T, Goesmann A. ReferenceSeeker: rapid determination of appropriate reference genomes. *J Open Source Softw*. 2020;5(46):1994. doi:10.21105/joss.01994

18. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*. 2015;32(1):268–274. doi:10.1093/molbev/msu300

19. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587–589. doi:10.1038/nmeth.4285

20. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput Biol*. 2015;11(2):e1004041. doi:10.1371/journal.pcbi.1004041

21. van der Putten BCL van der, Roodsant TJ, Haagmans MA, Schultsz C, Ark KCH van der. Five Complete Genome Sequences Spanning the Dutch *Streptococcus suis* Serotype 2 and Serotype 9 Populations. *Microbiol Resour Announc*. 2020;9(6). doi:10.1128/MRA.01439-19

22. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput Biol*. 2017;13(6):e1005595. doi:10.1371/journal.pcbi.1005595

23. Andrews S, others. FastQC: A Quality Control Tool for High Throughput Sequence Data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

24. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075. doi:10.1093/bioinformatics/btt086

25. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047–3048. doi:10.1093/bioinformatics/btw354

26. Varghese NJ, Mukherjee S, Ivanova N, *et al*. Microbial species delineation using whole genome sequences. *Nucleic Acids Res*. 2015;43(14):6761–6771. doi:10.1093/nar/gkv657

27. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Springer; 2016.

28. Arnold JB. ggthemes: Extra Themes, Scales and Geoms for "ggplot2." R Package Version. 2017;3(0).

29. Pedersen TL. patchwork: The Composer of ggplots. R Package Version 00. 2017;1.

30. team T pandas development. Pandas-Dev/Pandas: Pandas. Zenodo; 2020. doi:10.5281/zenodo.3509134

31. Jones E, Oliphant T, Peterson P, others. SciPy: Open Source Scientific Tools for Python.; 2001. http://www.scipy.org/

32. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *Clin Microbiol Rev*. 2019;32(3). doi:10.1128/CMR.00135-18

33. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol*. 2021;19(1):37–54. doi:10.1038/s41579-020-0416-x

34. Verschuuren TD, van Hout D, Arredondo-Alonso S, *et al*. Comparative genomics of ESBL-producing *Escherichia coli* (ESBL-Ec) reveals a similar distribution of the 10 most prevalent ESBL-Ec clones and ESBL genes among human community faecal and extra-intestinal infection isolates in the Netherlands (2014–17). *J Antimicrob Chemother*. 2021;76(4):901–908. doi:10.1093/jac/dkaa534

35. Le Gall T, Clermont O, Gouriou S, *et al*. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Mol Biol Evol*. 2007;24(11):2373–2384. doi:10.1093/molbev/msm172

36. Colquhoun RM, Hall MB, Lima L, *et al*. Nucleotide-resolution bacterial pan-genomics with reference graphs. *bioRxiv*. Published online November 24, 2020:2020.11.12.380378. doi:10.1101/2020.11.12.380378

Supplementary material

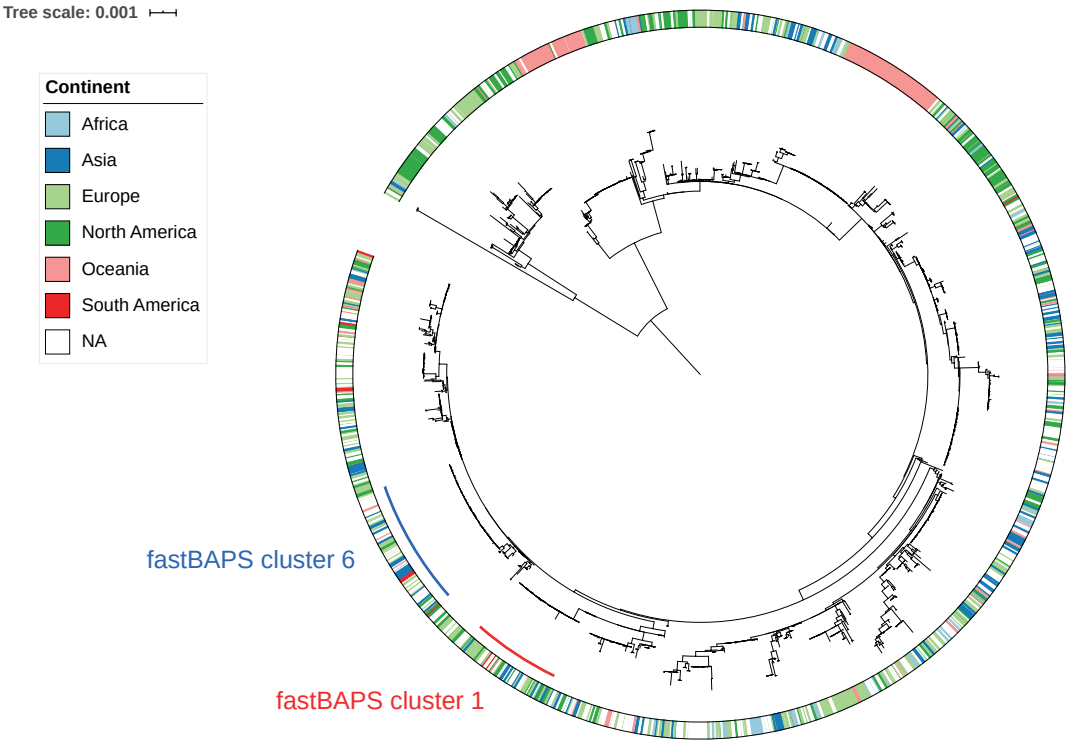


Figure S1. Core genome phylogeny of 1805 clonal complex 38 *E. coli* strains. Outer ring indicates continent on which the strain was isolated. The two clonal ST38 lineages present in the COMBAT collection are marked in red and blue. Tree and metadata available through iTOL: https://itol.embl.de/tree/21312783472680715892106 57.

Table S1. All SNP comparisons between isolates with relevant metadata (169 rows). Supplemental information: Members of the COMBAT consortium. ST = sequence type, SNPs = single nucleotide polymorphisms. Only isolates typed as the same sequence type were compared, as sequence type-specific reference genomes were used.

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4557050	ERR4557075	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	2	3434934	2.91
ERR4557051	ERR4557075	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3434481	4.37
ERR4557052	ERR4557075	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	4	3433509	5.82
ERR4557053	ERR4557075	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3435304	4.37
ERR4557050	ERR4557076	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	2	3434906	2.91
ERR4557051	ERR4557076	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3434435	4.37
ERR4557052	ERR4557076	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3433419	4.37
ERR4557053	ERR4557076	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	1	3435093	1.46
ERR4557050	ERR4557077	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	1	3435788	1.46
ERR4557051	ERR4557077	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3434867	4.37
ERR4557052	ERR4557077	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3433755	4.37
ERR4557053	ERR4557077	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3436609	4.36
ERR4557050	ERR4557078	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	1	3435192	1.46
ERR4557051	ERR4557078	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	4	3434596	5.82
ERR4557052	ERR4557078	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	3	3433567	4.37
ERR4557053	ERR4557078	trav032	trav032	10	GCA_004771235.1	within_traveler_between_timepoint	2	3435601	2.91
ERR4554399	ERR4554595	trav030	trav030	131	GCA_003856615.1	within_traveler_between_timepoint	1	5071211	0.99
ERR4554400	ERR4554595	trav030	trav030	131	GCA_003856615.1	within_traveler_between_timepoint	2	5073002	1.97
ERR4554411	ERR4554595	trav040	trav030	131	GCA_003856615.1	between_traveler	154	4829693	159.43
ERR4554537	ERR4554595	trav076	trav030	131	GCA_003856615.1	between_traveler	163	4894509	166.51
ERR4554452	ERR4554595	trav095	trav030	131	GCA_003856615.1	between_traveler	171	4872128	175.49
ERR4554399	ERR4554598	trav030	trav040	131	GCA_003856615.1	between_traveler	152	4827746	157.42

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554400	ERR4554598	trav030	trav040	131	GCA_003856615.1	between_traveler	152	4828824	157.39
ERR4554411	ERR4554598	trav040	trav040	131	GCA_003856615.1	within_traveler_between_timepoint	2	4858865	2.06
ERR4554537	ERR4554598	trav076	trav040	131	GCA_003856615.1	between_traveler	88	4845115	90.81
ERR4554452	ERR4554598	trav095	trav040	131	GCA_003856615.1	between_traveler	50	4858185	51.46
ERR4554399	ERR4555027	trav030	trav076	131	GCA_003856615.1	between_traveler	164	4868185	168.44
ERR4554400	ERR4555027	trav030	trav076	131	GCA_003856615.1	between_traveler	163	4869344	167.37
ERR4554411	ERR4555027	trav040	trav076	131	GCA_003856615.1	between_traveler	91	4827445	94.25
ERR4554537	ERR4555027	trav076	trav076	131	GCA_003856615.1	within_traveler_between_timepoint	2	4983867	2.01
ERR4554452	ERR4555027	trav095	trav076	131	GCA_003856615.1	between_traveler	106	4863330	108.98
ERR4554399	ERR4555023	trav030	trav095	131	GCA_003856615.1	between_traveler	171	4872106	175.49
ERR4554400	ERR4555023	trav030	trav095	131	GCA_003856615.1	between_traveler	172	4873062	176.48
ERR4554411	ERR4555023	trav040	trav095	131	GCA_003856615.1	between_traveler	51	4869683	52.36
ERR4554537	ERR4555023	trav076	trav095	131	GCA_003856615.1	between_traveler	107	4890556	109.39
ERR4554452	ERR4555023	trav095	trav095	131	GCA_003856615.1	within_traveler_between_timepoint	2	4913577	2.04
ERR4554414	ERR4554599	trav002	trav002	2141	GCA_904863375.1	within_traveler_between_timepoint	31	4576666	33.87
ERR4554555	ERR4557044	trav017	trav017	38	GCA_006965465.1	within_traveler_between_timepoint	3	4472047	3.35
ERR4554438	ERR4557044	trav033	trav017	38	GCA_006965465.1	between_traveler	22	4451424	24.71
ERR4554420	ERR4557044	trav044	trav017	38	GCA_006965465.1	between_traveler	254	4427084	286.87
ERR4554427	ERR4557044	trav052	trav017	38	GCA_006965465.1	between_traveler	232	4269517	271.69
ERR4554446	ERR4557044	trav064	trav017	38	GCA_006965465.1	between_traveler	244	4345723	280.74
ERR4554533	ERR4557044	trav068	trav017	38	GCA_006965465.1	between_traveler	22	4472807	24.59
ERR4554546	ERR4557044	trav079	trav017	38	GCA_006965465.1	between_traveler	307	4135626	371.17
ERR4554552	ERR4557044	trav081	trav017	38	GCA_006965465.1	between_traveler	242	4270979	283.31
ERR4554553	ERR4557044	trav081	trav017	38	GCA_006965465.1	between_traveler	241	4274610	281.90

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554435	ERR4557044	trav093	trav017	38	GCA_006965465.1	between_traveler	229	4241472	269.95
ERR4554555	ERR4554605	trav017	trav033	38	GCA_006965465.1	between_traveler	17	4451452	19.09
ERR4554438	ERR4554605	trav033	trav033	38	GCA_006965465.1	within_traveler_between_timepoint	2	4449351	2.25
ERR4554420	ERR4554605	trav044	trav033	38	GCA_006965465.1	between_traveler	275	4440742	309.63
ERR4554427	ERR4554605	trav052	trav033	38	GCA_006965465.1	between_traveler	209	4273436	244.53
ERR4554446	ERR4554605	trav064	trav033	38	GCA_006965465.1	between_traveler	218	4344858	250.87
ERR4554533	ERR4554605	trav068	trav033	38	GCA_006965465.1	between_traveler	21	4452239	23.58
ERR4554546	ERR4554605	trav079	trav033	38	GCA_006965465.1	between_traveler	269	4138173	325.02
ERR4554552	ERR4554605	trav081	trav033	38	GCA_006965465.1	between_traveler	206	4268099	241.33
ERR4554553	ERR4554605	trav081	trav033	38	GCA_006965465.1	between_traveler	206	4271105	241.16
ERR4554435	ERR4554605	trav093	trav033	38	GCA_006965465.1	between_traveler	241	4254864	283.21
ERR4554555	ERR4554603	trav017	trav044	38	GCA_006965465.1	between_traveler	263	4478085	293.65
ERR4554438	ERR4554603	trav033	trav044	38	GCA_006965465.1	between_traveler	244	4459022	273.60
ERR4554420	ERR4554603	trav044	trav044	38	GCA_006965465.1	within_traveler_between_timepoint	5	5149259	4.86
ERR4554427	ERR4554603	trav052	trav044	38	GCA_006965465.1	between_traveler	241	4494758	268.09
ERR4554446	ERR4554603	trav064	trav044	38	GCA_006965465.1	between_traveler	315	4605683	341.97
ERR4554533	ERR4554603	trav068	trav044	38	GCA_006965465.1	between_traveler	265	4478650	295.85
ERR4554546	ERR4554603	trav079	trav044	38	GCA_006965465.1	between_traveler	138	4325947	159.50
ERR4554552	ERR4554603	trav081	trav044	38	GCA_006965465.1	between_traveler	233	4483049	259.87
ERR4554553	ERR4554603	trav081	trav044	38	GCA_006965465.1	between_traveler	237	4485930	264.16
ERR4554435	ERR4554603	trav093	trav044	38	GCA_006965465.1	between_traveler	183	4802615	190.52
ERR4554555	ERR4555016	trav017	trav052	38	GCA_006965465.1	between_traveler	245	4296252	285.13
ERR4554438	ERR4555016	trav033	trav052	38	GCA_006965465.1	between_traveler	246	4292227	286.56
ERR4554420	ERR4555016	trav044	trav052	38	GCA_006965465.1	between_traveler	264	4482517	294.48
ERR4554427	ERR4555016	trav052	trav052	38	GCA_006965465.1	within_traveler_between_timepoint	5	4505101	5.55

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554446	ERR4555016	trav064	trav052	38	GCA_006965465.1	between_traveler	291	4407431	330.12
ERR4554533	ERR4555016	trav068	trav052	38	GCA_006965465.1	between_traveler	245	4297007	285.08
ERR4554546	ERR4555016	trav079	trav052	38	GCA_006965465.1	between_traveler	300	4209782	356.31
ERR4554552	ERR4555016	trav081	trav052	38	GCA_006965465.1	between_traveler	46	4484813	51.28
ERR4554553	ERR4555016	trav081	trav052	38	GCA_006965465.1	between_traveler	46	4488997	51.24
ERR4554435	ERR4555016	trav093	trav052	38	GCA_006965465.1	between_traveler	229	4357874	262.74
ERR4554555	ERR4555024	trav017	trav064	38	GCA_006965465.1	between_traveler	246	4366778	281.67
ERR4554438	ERR4555024	trav033	trav064	38	GCA_006965465.1	between_traveler	245	4358687	281.05
ERR4554420	ERR4555024	trav044	trav064	38	GCA_006965465.1	between_traveler	333	4578450	363.66
ERR4554427	ERR4555024	trav052	trav064	38	GCA_006965465.1	between_traveler	276	4400994	313.57
ERR4554446	ERR4555024	trav064	trav064	38	GCA_006965465.1	within_traveler_between_timepoint	6	4628676	6.48
ERR4554533	ERR4555024	trav068	trav064	38	GCA_006965465.1	between_traveler	247	4367513	282.77
ERR4554546	ERR4555024	trav079	trav064	38	GCA_006965465.1	between_traveler	295	4300485	342.98
ERR4554552	ERR4555024	trav081	trav064	38	GCA_006965465.1	between_traveler	281	4392685	319.85
ERR4554553	ERR4555024	trav081	trav064	38	GCA_006965465.1	between_traveler	280	4396496	318.44
ERR4554435	ERR4555024	trav093	trav064	38	GCA_006965465.1	between_traveler	298	4411789	337.73
ERR4554555	ERR4555026	trav017	trav068	38	GCA_006965465.1	between_traveler	18	4498692	20.01
ERR4554438	ERR4555026	trav033	trav068	38	GCA_006965465.1	between_traveler	25	4477126	27.92
ERR4554420	ERR4555026	trav044	trav068	38	GCA_006965465.1	between_traveler	260	4453171	291.93
ERR4554427	ERR4555026	trav052	trav068	38	GCA_006965465.1	between_traveler	230	4290343	268.04
ERR4554446	ERR4555026	trav064	trav068	38	GCA_006965465.1	between_traveler	242	4366742	277.09
ERR4554533	ERR4555026	trav068	trav068	38	GCA_006965465.1	within_traveler_between_timepoint	0	4499405	0.00
ERR4554546	ERR4555026	trav079	trav068	38	GCA_006965465.1	between_traveler	302	4156541	363.28
ERR4554552	ERR4555026	trav081	trav068	38	GCA_006965465.1	between_traveler	239	4291956	278.43
ERR4554553	ERR4555026	trav081	trav068	38	GCA_006965465.1	between_traveler	238	4295673	277.02

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554435	ERR4555026	trav093	trav068	38	GCA_006965465.1	between_traveler	234	4267685	274.15
ERR4554555	ERR4555031	trav017	trav079	38	GCA_006965465.1	between_traveler	312	4153885	375.55
ERR4554438	ERR4555031	trav033	trav079	38	GCA_006965465.1	between_traveler	315	4152022	379.33
ERR4554420	ERR4555031	trav044	trav079	38	GCA_006965465.1	between_traveler	270	4273604	315.89
ERR4554427	ERR4555031	trav052	trav079	38	GCA_006965465.1	between_traveler	299	4201059	355.86
ERR4554446	ERR4555031	trav064	trav079	38	GCA_006965465.1	between_traveler	297	4297433	345.56
ERR4554533	ERR4555031	trav068	trav079	38	GCA_006965465.1	between_traveler	314	4154406	377.91
ERR4554546	ERR4555031	trav079	trav079	38	GCA_006965465.1	within_traveler_between_timepoint	8	4340737	9.22
ERR4554552	ERR4555031	trav081	trav079	38	GCA_006965465.1	between_traveler	302	4193370	360.09
ERR4554553	ERR4555031	trav081	trav079	38	GCA_006965465.1	between_traveler	307	4197181	365.72
ERR4554435	ERR4555031	trav093	trav079	38	GCA_006965465.1	between_traveler	239	4102435	291.29
ERR4554555	ERR4557046	trav017	trav081	38	GCA_006965465.1	between_traveler	241	4294521	280.59
ERR4554438	ERR4557046	trav033	trav081	38	GCA_006965465.1	between_traveler	242	4290428	282.02
ERR4554420	ERR4557046	trav044	trav081	38	GCA_006965465.1	between_traveler	247	4458270	277.01
ERR4554427	ERR4557046	trav052	trav081	38	GCA_006965465.1	between_traveler	44	4481491	49.09
ERR4554446	ERR4557046	trav064	trav081	38	GCA_006965465.1	between_traveler	280	4395706	318.49
ERR4554533	ERR4557046	trav068	trav081	38	GCA_006965465.1	between_traveler	241	4295110	280.55
ERR4554546	ERR4557046	trav079	trav081	38	GCA_006965465.1	between_traveler	294	4198871	350.09
ERR4554552	ERR4557046	trav081	trav081	38	GCA_006965465.1	within_traveler_between_timepoint	3	4495704	3.34
ERR4554553	ERR4557046	trav081	trav081	38	GCA_006965465.1	within_traveler_between_timepoint	4	4499814	4.44
ERR4554435	ERR4557046	trav093	trav081	38	GCA_006965465.1	between_traveler	211	4346688	242.71
ERR4554555	ERR4555018	trav017	trav093	38	GCA_006965465.1	between_traveler	274	4305804	318.18
ERR4554438	ERR4555018	trav033	trav093	38	GCA_006965465.1	between_traveler	257	4287895	299.68
ERR4554420	ERR4555018	trav044	trav093	38	GCA_006965465.1	between_traveler	202	4821939	209.46
ERR4554427	ERR4555018	trav052	trav093	38	GCA_006965465.1	between_traveler	251	4390824	285.82

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554446	ERR4555018	trav064	trav093	38	GCA_006965465.1	between_traveler	327	4448827	367.51
ERR4554533	ERR4555018	trav068	trav093	38	GCA_006965465.1	between_traveler	277	4306653	321.60
ERR4554546	ERR4555018	trav079	trav093	38	GCA_006965465.1	between_traveler	267	4147096	321.91
ERR4554552	ERR4555018	trav081	trav093	38	GCA_006965465.1	between_traveler	245	4383790	279.44
ERR4554553	ERR4555018	trav081	trav093	38	GCA_006965465.1	between_traveler	243	4387295	276.94
ERR4554435	ERR4555018	trav093	trav093	38	GCA_006965465.1	within_traveler_between_timepoint	4	4835465	4.14
ERR4554431	ERR4554604	trav007	trav007	393	GCA_015136715.1	within_traveler_between_timepoint	4	4994439	4.00
ERR4554547	ERR4554604	trav010	trav007	393	GCA_015136715.1	between_traveler	147	4981334	147.55
ERR4554431	ERR4555032	trav007	trav010	393	GCA_015136715.1	between_traveler	148	4988815	148.33
ERR4554547	ERR4555032	trav010	trav010	393	GCA_015136715.1	within_traveler_between_timepoint	2	5072018	1.97
ERR4557041	ERR4555028	trav011	trav011	405	GCA_002164645.3	within_traveler_between_timepoint	5	4997262	5.00
ERR4554581	ERR4555028	trav086	trav011	405	GCA_002164645.3	between_traveler	131	4642925	141.07
ERR4554582	ERR4555028	trav086	trav011	405	GCA_002164645.3	between_traveler	131	4641393	141.12
ERR4557041	ERR4557045	trav011	trav086	405	GCA_002164645.3	between_traveler	129	4644236	138.88
ERR4554581	ERR4557045	trav086	trav086	405	GCA_002164645.3	within_traveler_between_timepoint	2	4701378	2.13
ERR4554582	ERR4557045	trav086	trav086	405	GCA_002164645.3	within_traveler_between_timepoint	2	4699250	2.13
ERR4554425	ERR4555017	trav051	trav051	449	GCA_011068385.1	within_traveler_between_timepoint	21	4905965	21.40
ERR4554442	ERR4554592	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	272	4453487	305.38
ERR4554397	ERR4554592	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	274	4451745	307.74
ERR4557048	ERR4554592	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	273	4453946	306.47
ERR4557049	ERR4554592	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	274	4454521	307.55
ERR4554442	ERR4557067	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	274	4455411	307.49
ERR4554397	ERR4557067	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	277	4454323	310.93
ERR4557048	ERR4557067	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	276	4456491	309.66
ERR4557049	ERR4557067	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	277	4457939	310.68

Isolate (T1)	Isolate (T12)	Traveler (isolate T1)	Traveler (isolate T12)	ST	Reference genome	Comparison type	SNPs	Alignment lengths	SNPs (scaled to 5Mbp)
ERR4554442	ERR4557068	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	272	4455408	305.25
ERR4554397	ERR4557068	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	277	4454313	310.93
ERR4557048	ERR4557068	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	274	4456442	307.42
ERR4557049	ERR4557068	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	275	4457846	308.44
ERR4554442	ERR4557069	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	275	4455523	308.61
ERR4554397	ERR4557069	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	280	4454416	314.29
ERR4557048	ERR4557069	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	277	4456573	310.78
ERR4557049	ERR4557069	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	278	4457943	311.80
ERR4554442	ERR4557070	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	275	4455531	308.61
ERR4554397	ERR4557070	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	280	4454409	314.30
ERR4557048	ERR4557070	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	277	4456565	310.78
ERR4557049	ERR4557070	trav026	trav026	617	GCA_003030365.1	within_traveler_between_timepoint	278	4457947	311.80
ERR4554447	ERR4555020	trav046	trav046	69	GCA_900622685.1	within_traveler_between_timepoint	19	5003966	18.98
ERR4554448	ERR4555020	trav047	trav046	69	GCA_900622685.1	between_traveler	14	5000488	14.00
ERR4554447	ERR4555021	trav046	trav046	69	GCA_900622685.1	within_traveler_between_timepoint	22	5005469	21.98
ERR4554448	ERR4555021	trav047	trav046	69	GCA_900622685.1	between_traveler	9	5001360	9.00
ERR4554447	ERR4555019	trav046	trav047	69	GCA_900622685.1	between_traveler	14	4997095	14.01
ERR4554448	ERR4555019	trav047	trav047	69	GCA_900622685.1	within_traveler_between_timepoint	11	4993532	11.01
ERR4554447	ERR4555022	trav046	trav047	69	GCA_900622685.1	between_traveler	12	5005636	11.99
ERR4554448	ERR4555022	trav047	trav047	69	GCA_900622685.1	within_traveler_between_timepoint	12	5001331	12.00
ERR4554454	ERR4555025	trav098	trav098	-	GCA_009720465.1	within_traveler_between_timepoint	20	4149494	24.10

Chapter 5

Genome-wide association reveals host-specific genomic traits in *Escherichia coli*

Sumeet K. Tiwari[$] & Boas C.L. van der Putten[$], Thilo M. Fuchs, Trung N. Vinh, Martin Bootsma, Rik Oldenkamp, Roberto La Ragione, Sebastien Matamoros, Ngo T. Hoa, Christian Berens, Joy Leng, Julio Álvarez, Marta Ferrandis-Vila, Jenny M. Ritchie, Angelika Fruth, Stefan Schwarz, Lucas Domínguez, María Ugarte-Ruiz, Astrid Bethe, Charlotte Huber, Vanessa Johanns, Ivonne Stamm, Lothar H. Wieler, Christa Ewers, Amanda Fivian-Hughes, Herbert Schmidt, Christian Menge, Torsten Semmler*, Constance Schultsz*
[$]Equal contribution
*Equal contribution

Submitted for publication
bioRxiv, https://doi.org/10.1101/2022.02.08.479532

Abstract

Escherichia coli is an opportunistic pathogen that can colonize or infect various host species. There is a significant gap in our understanding to what extent genetic lineages of *E. coli* are adapted or restricted to specific hosts. In addition, genomic determinants underlying such host specificity are unknown.

By analyzing a randomly sampled collection of 1,198 whole-genome sequenced *E. coli* isolates from four countries (Germany, UK, Spain, and Vietnam), obtained from five host species (human, pig, cattle, chicken, and wild boar) over 16 years, from both healthy and diseased hosts, we demonstrate that certain lineages of *E. coli* are frequently detected in specific hosts. We report a novel *nan* gene cluster, designated *nan-9*, putatively encoding acetylesterases and determinants of uptake and metabolism of sialic acid, to be associated with the human host as identified through genome wide association studies. *In silico* characterization predicts *nan-9* to be involved in sialic acid (Sia) metabolism. *In vitro* growth experiments with a representative Δ*nan E. coli* mutant strain, using sialic acids 5-*N*-acetyl neuraminic acid (Neu5Ac) and *N*-glycolyl neuraminic acid (Neu5Gc) as the sole carbon source, indicate an impaired growth behaviour compared to the wild-type.

In addition, we identified several additional *E. coli* genes that are potentially associated with adaptation to human, cattle and chicken hosts, but not for the pig host. Collectively, this study provides an extensive overview of genetic determinants which may mediate host specificity in *E. coli*. Our findings should inform risk analysis and epidemiological monitoring of (antimicrobial resistant) *E. coli*.

Introduction

Escherichia coli is a Gram-negative bacterium which has been isolated from various host species, including humans, cattle, chickens and pigs[1]. Because *E. coli* can colonize or infect multiple host species, this bacterium can act as a reservoir for genes encoding antimicrobial resistance (AMR)[2] that can be transmitted between different host species. The likelihood that *E. coli* and its AMR encoding genes persist in a new host after transmission depends on multiple factors[3,4]. For example, small changes in metabolic pathways may enable *E. coli* to colonize or infect a host more efficiently[1]. Several studies have suggested that highly successful *E. coli* clones, such as the sequence type 131 (ST131) clone[5,6] or clonal complex 87 (ST58 and ST155) *E. coli* facilitate the spread of AMR *E. coli* in the human population[7] whilst other studies have shown that different lineages of AMR *E. coli* vary in their ability to spread[8]. These findings both indicate that AMR genes, at least to some extent, hitchhike on bacterial strains that are specifically equipped to colonize a given host. Beyond classical virulence or adhesion factors, genetic and functional traits defining different degrees of host adaptation[3,9] and thereby indirectly impacting on the spread of AMR between host species, have not been identified thus far.

Comparative genomic analysis of bacterial populations from multiple hosts has revealed signatures of host-adaptation in bacterial genomes[10]. The emergence of large-scale bacterial genome-wide association studies (GWAS) allowed for the detection of genes or genomic variants that are associated with resistance, pathogenicity, and host adaptive traits[11–13]. Here, we have applied population-based bacterial GWAS to identify host-associated genomic determinants in a diverse panel of 1,198 *E. coli* isolates, irrespective of their AMR pattern. Isolates were recovered from five different host species, including healthy and diseased individuals from four different countries in two continents over 16 years. The *pan*-genome was analyzed for specific host association followed by a *k-mer* based bacterial GWAS approach to identify host-specific genomic determinants and their potential role in host-adaptation.

Material and Methods

Sampling strategy

A panel of 1,213 *E. coli* isolates from four countries (Germany, UK, Spain, and Vietnam), obtained from five host species (human, pig, cattle, chicken, and wild boar) during three time periods (2003-2007, 2008-2012 and 2013-2018) from both healthy and diseased hosts were selected randomly from existing strain collections and newly collected isolates. Out of 120 possible strata (defined as a unique combination of country, host, time-period, and host health status), 42 strata contained isolates. We included all isolates available per

stratum if there were less than 30 isolates and performed a random selection of up to a maximum of 30 isolates if more were available. Potentially duplicate isolates that were part of an outbreak, isolated at a single location within a short timeframe, or from a single farm or a single individual were excluded. Only one isolate per individual was included in the analyses. Isolates included per stratum are shown in Table S1.

DNA extraction and sequencing

The DNA of the *E. coli* isolates from Germany was extracted using the QIAamp DNA Mini Kit (Qiagen) following the manufacturer's instructions. The DNA concentration was evaluated fluorometrically by using Qubit™ 2.0 fluorometer (Invitrogen, USA) and the associated Qubit™ dsDNA HS Assay Kit (0.2-100ng) and Qubit™ BR Assay Kit (2-1000ng), respectively. The libraries were generated using Nextera DNA library preparation (Illumina, https://www.illumina.com). The sequencing was performed using the Illumina MiSeq and HiSeq systems, generating 2 × 250 bp and 2 × 150 bp reads, respectively.

The DNA of the *E. coli* isolates from the UK was purified using a Promega DNA Wizard® genomic purification kit and quantified using Nanodrop. Libraries were generated using Nextera XT technology (Illumina), and DNA sequencing of isolates was performed at the Animal and Plant Health Agency (APHA, Surrey, UK, https://www.gov.uk/government/-organisations/animal-and-plant-healthagency) using an Illumina MiSeq system generating 2 × 150 bp reads.

For *E. coli* isolates from Spain, DNA was extracted using the DNA blood and tissue Qiagen kit according to the manufacturer's instruction. The total amount of DNA was quantified using a Qubit fluorometer and frozen at -20°C until further analysis. Libraries were prepared using Nextera XT DNA Library preparation (Illumina), and DNA samples were sequenced using a MiSeq platform (2 × 300 cycle V3 Kit).

The DNA of the *E. coli* isolates from Vietnam was extracted using the Wizard Genomic DNA purification kit (Promega, Madison, WI, USA) following the manufacturer's instructions. The concentration of the DNA was measured fluorometrically by using picogreen (Invitrogen). The sequencing was performed using an Illumina HiSeq 4000 system, which generates 2 × 150 bp reads.

Quality control

Adapter sequences were removed from raw reads using flexbar v3.0.3[14,15] with trimming mode (-ae) ANY. Low-quality bases within raw reads (Phred score value <20) were trimmed using a sliding window approach (-q WIN). FastQC v0.11.7[16] and MultiQC v1.6[17] were used for quality control before and after processing steps.

Genome assembly and annotation

Adapter-trimmed reads were assembled using SPAdes v3.13.1[18] using read correction. Scaffolds smaller than 500bp were discarded. QUAST v5.0.0[19] was used to assess assembly quality using default parameters. Draft assemblies were excluded if the N50 was below an aribrary value of 30 kbp or consisted of more than 900 contigs. Draft genomes were annotated using prokka v1.13[20] with a genus-specific blast for *Escherichia*. Phylogroups were predicted using ClermonTyper v1.4.1[21], and sequence types (STs) of the isolates were identified *in silico* using the Achtman seven gene MLST scheme using mlst (https://github.com/tseemann/mlst).

Pan-genome and phylogenetic analysis

Roary v3.12.0[22] was used to define the *pan*-genome of the population, using paralog splitting. The core genes were aligned using prank[23] on default parameters. The core gene alignment was used to construct the phylogenetic tree using RaxML 8.2.4[24] with 100 bootstraps under a General Time Reversible (GTR) substitution model with the Gamma model of rate heterogeneity and Lewis ascertainment bias correction[25]. The core gene phylogeny was corrected for recombination using ClonalFrameML[26] using default parameters. Phylogenetic Clusters (or BAPS clusters) within the dataset were defined using hierBAPS[27,28] based on the core gene alignment. The accessory gene clustering was performed using package Rtsne v0.15[29,30] with 5000 iterations and perplexity 15 in R v3.6.1. iTOL[31] and Microreact[32] were used to visualize the population structure in the context of available metadata. The function chisq.test from the MASS library[33] (v7.3-51.1) was used in R[34] (v3.5.2) to perform X^2-tests of independence between phylogenetic clusters and host species. Tests were carried out on the full dataset (14 phylogenetic clusters vs. five hosts and nine phylogroups vs. five host species).

Genome-wide association study (GWAS)

We excluded the wild boar *E. coli* isolates from the GWAS analysis, because of their low number (n=29). GWAS was performed to screen *k-mers* for associations with their host (pig, human, chicken, and cattle). Assemblies were shredded into *k-mers* of 9-100 bases using FSM-lite (https://github.com/nvalimak/fsm-lite). The association between *k-mers* and host phenotype was carried out using Fast-LMM linear mixed model implemented in pyseer[35] using a pairwise similarity matrix derived from the phylogenetic tree as population correction. A GWAS analysis was carried out for each host (pig, human, chicken, and cattle). To reduce false-positive associations, isolates from the host of interest were compared with an equal number of isolates from each of the other hosts, designated control isolates. This analysis was repeated 100 times per host of interest by selecting the control strains from other hosts per iteration[36]. The selection of control isolates was random and with replacement except for stratification by phylogenetic clusters to minimize phylogenetic bias. The statistical significance threshold was

estimated based on the number of unique *k-mers* patterns for each run[35]. *K-mers*, which were significantly associated with 90% of the runs per host, were retained and mapped to reference genomes (Table S2) using a fastmap algorithm in bwa[35,37] . An arbitrary cut-off of a minimum of 10 *k-mers* mapped per gene was chosen for further analysis to reduce false-positives. *In silico* characterization and gene ontology (GO) assignment was performed using Blast2GO[38], and Clusters of Orthologous Groups (COGs) were assigned using CD-search[39,40].

Prevalence of a human-associated nan gene cluster

All available *E. coli* genome assemblies in NCBI RefSeq were downloaded on Nov 29th, 2019, using NCBI-genome-download (https://github.com/kblin/ncbi-genome-download). Using a custom ABRicate (https://github.com/tseemann/abricate) database, consisting of the nine genes of the novel human-associated *nan* gene cluster, all downloaded genomes (n=17,994) were scanned. STs for all the genomes were assigned as described above.

Construction of mutants and phenotypic experiments

Mutants Δ*nan-9* (Amp^R) and Δ*nanRATEK* of extra-intestinal pathogenic *E. coli* (ExPEC) strain IMT12185 (ST131; RKI 20-00501; Amp^R) were constructed using the Datsenko-Wanner method[41]. The genomic DNA of the wild-type and the mutant strains was isolated using a QIAamp DNA Mini Kit (QIAGEN). Libraries were prepared using the Nextera XT DNA Library preparation kit (Illumina), and MinION one-dimensional (1D) libraries were constructed using the SQK-RBK004 kit (Nanopore technologies, Oxford, UK) and loaded according to the manufacturer's instructions onto an R9.4 flow cell. MinIon sequencing data were collected for 48 h and the paired-end Illumina sequencing was performed using MiSeq. Hybrid assembly using Illumina and MinION reads was performed using unicycler v0.4.8[42] with default parameters to complete both strains' genomes. The absence of the desired genes was confirmed based on the assembly followed by annotation using prokka v1.13[20].

Carbon utilization and chemical sensitivity of the deletion mutants and their parental strain were tested using a Biolog Phenotypic Array system, using the PM1 MicroPlate and the Gen III MicroPlate according to the manufacturer's instructions.

Growth curve analysis

E. coli strains were grown at 37°C aerobically in lysogeny broth (LB) (10 g/l tryptone, 5 g/l yeast extract, 5 g/l NaCl, pH 7.5) or in minimal medium (MM). MM is M9 mineral medium (33.7 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.55 mM NaCl, 9.35 mM NH_4Cl) supplemented with 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$. As carbon and energy source, either 27.8 mM [0.5% w/v] glucose, 6.47 mM [0.2% w/v] 5-N-acetyl neuraminic acid (Neu5Ac), or 6.15 mM [0.1% w/v] N-glycolylneuraminic acid (Neu5Gc) (all purchased from Sigma-Aldrich, Taufkirchen, Germany) was added. If appropriate, the following antibiotics were used: ampicillin

sodium salt (150 mg/ml) or kanamycin (50 µg/ml). For solid media, 1.5% agar (w/v) was added. For all growth experiments, bacterial strains were grown in LB medium overnight at 37°C, washed twice in PBS and then adjusted to an optical density at 600 nm (OD_{600}) of 0.005 in the desired liquid growth medium, or streaked on agar plates. Growth curves were obtained from bacterial cultures incubated at 37°C with gentle agitation in 96-well microtitre plates containing 200 µl medium. The OD_{600} was measured by an automatic reader (Epoch2T; BioTek, Bad Friedrichshall, Germany) at appropriate time intervals as indicated.

Results

Data collection

After WGS quality control, 14 isolates were excluded because of poor quality sequences. One additional isolate was excluded since this isolate was identified as *Escherichia marmotae* (formerly cryptic clade V)[43,44], a species commonly mistaken for *E. coli*. Our final collection comprised 1,198 *E. coli* whole-genome sequences with metadata (Table S1), which also contained 8 cryptic clade I isolates, which were included as *E. coli* based on the recommended species cut-off of 95-96% average nucleotide identity[43]. Our collection consisted of 22.1% (n=265) cattle, 28.1% (n=337) chicken, 27.3% (n=327) human, 20.3% (n=240) pigs and 2.4% (n=29) wild boar isolates (Fig. S1A). Fifty-one percent (n=612), 19.4% (n=233), 14.5% (n=174) and 14.9% (n=179) of these isolates were from Germany, Spain, the UK, and Vietnam, respectively (Fig. S1A). Chicken isolates were from all four countries, human isolates from Germany, the UK and Vietnam, pig isolates from Germany, Spain and Vietnam, cattle isolates from Germany and Spain and only Spain provided wild boar isolates. In total, 35.5% (n=426) of the isolates were from hosts with reported disease, whereas 62.0% (n=743) were from hosts without reported disease, while host health status was unknown for the wild boar isolates (2.4%, n=29). Of the 1198 isolates analyzed, 1140 were grouped into 358 different STs, and 58 could not be assigned to any known ST. The population structure of the collection closely resembles that of the ECOR collection[45], indicating that it represents most of the known diversity of *E. coli sensu stricto* (Fig. S2).

Pan-genome analysis

The *pan*-genome of the 1,198 *E. coli* isolates consisted of 77,130 genes, of which 1,956 genes belonged to the core genome (i.e., present in at least 99% of the isolates). The population structure of the collection based on core genome single-nucleotide polymorphisms (SNPs) was defined using Bayesian analysis of population structure (BAPS), which assigns isolates to discrete clusters. Most of the isolates were assigned to phylogroups B1 (n=366, 30.55%), A (n=313, 26.12%) and B2 (n=213, 17.77%). The remaining isolates were distributed

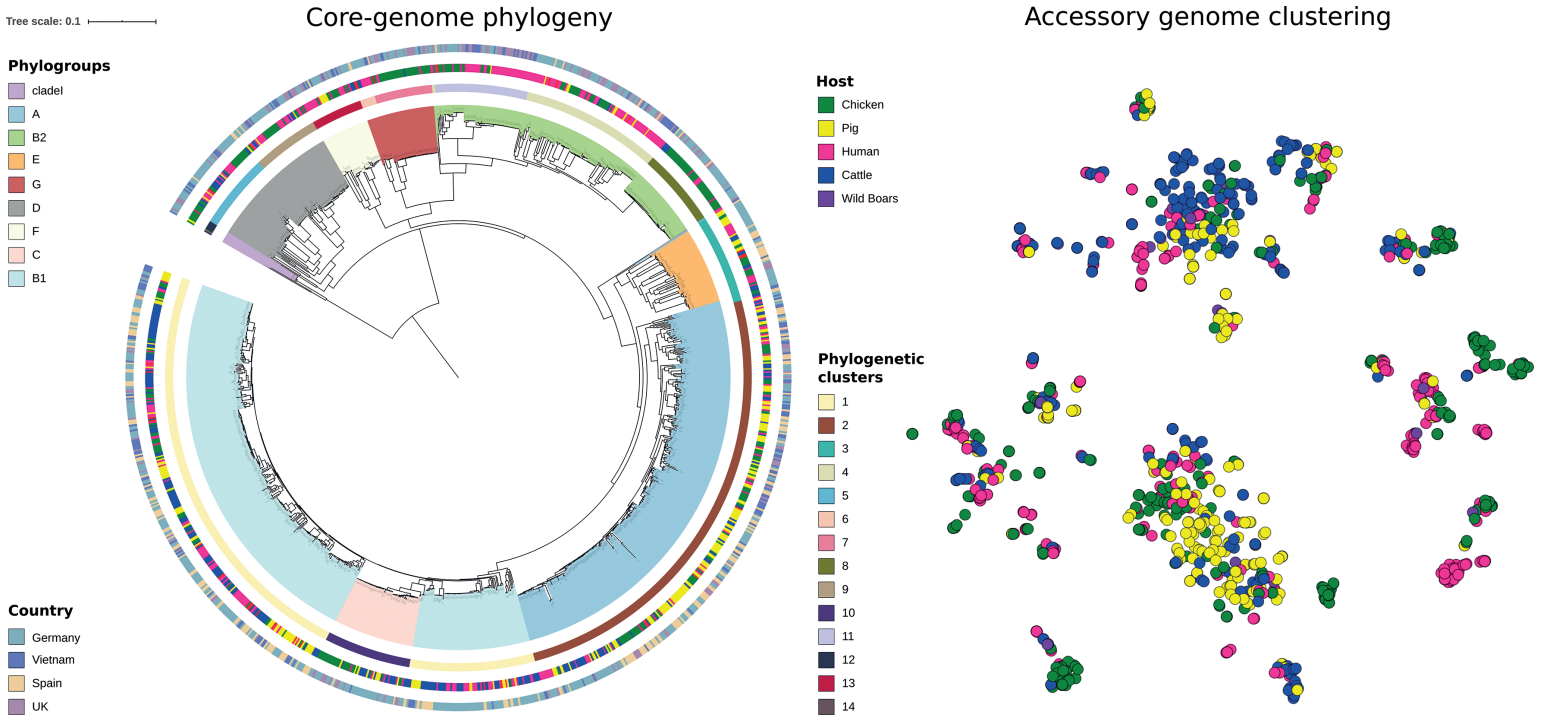


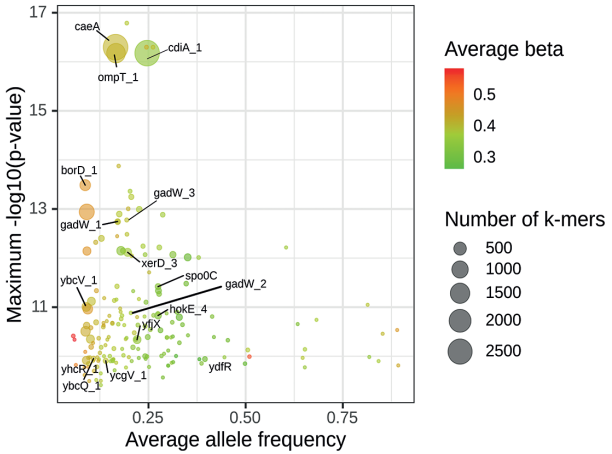
Fig 1. Distribution of 1,198 isolates with host species by core-genome phylogeny (left) and clustering based on accessory gene content (right). Clades on the phylogeny represent phylogroups, inner-ring represents phylogenetic clusters, middle-ring represents host-species, and outer ring indicates the geographical region.

among phylogroups D (n=97, 8.09%), E (n=55, 4.59%), G (n=49, 4.09%), F (n=35, 2.92%), C (n=60, 5.0%), and clade I (n=8, 0.6%). A comparison of phylogenetic clusters, phylogroups, country, host, and a maximum likelihood (ML) tree based on 110,920 core-genome SNPs is shown in Fig 1. The χ^2-tests for independence revealed a positive correlation between host status and phylogenetic clusters (at $p < 2.26e^{-16}$, df=52) and between phylogroups and hosts ($p<2.2e^{-16}$, df=32). This indicates that specific phylogenetic clusters (Fig. S1 B&C) and phylogroups, such as B1 (cattle), A (pig), B2 (human and chicken), and G (chicken) were enriched within different hosts in our collection (Fig. S1D).

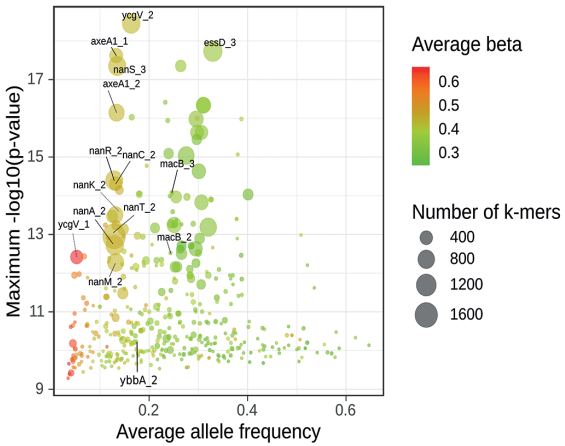
Clustering of isolates based both on core gene alignment and on accessory gene profile appeared to be correlated with phylogroups. The interactive visualization of data is also available on Microreact (https://microreact.org/project/ouDOdcFxc). A minimum spanning tree was built on the allelic profiles of 358 (n=1,140 isolates) known STs and 58 isolates belonging to unknown STs using GrapeTree[46] along with the host distribution (Fig. S3). Several sequence types, of which at least ten isolates were available, appeared to be linked with certain host species. ST33 (n= 10/10, 10 human isolates out of all 10 isolates), ST73 (n=11/17), ST131 (n=37/42) and ST1193 (n=12/12) were associated with a human host. ST131 was also found in chickens (n=4/42) and pigs (n=1/42) in this collection. ST23 (n=18/22), ST95 (n=25/31), ST115 (n=11/11), ST117 (n=30/33), ST140 (n=19/20) and ST752 (n=29/30) were associated with the chicken host.

GWAS

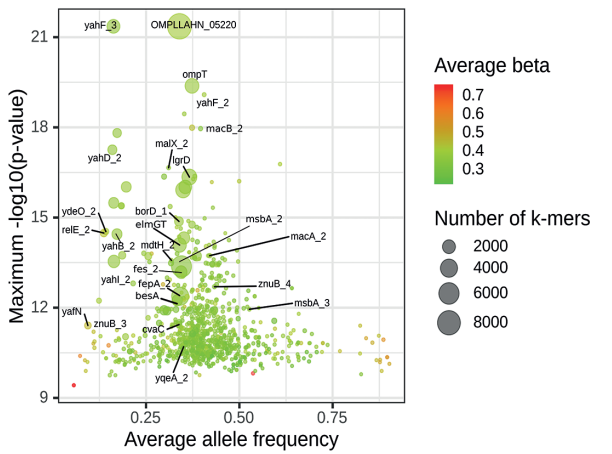
The genome-wide association analysis was performed on 1,169 *E. coli* isolates from cattle, chickens, humans, and pigs. The 29 wild boar isolates were excluded because of their small group size. Genome-wide association analysis revealed the positive association (ß>0) of 27,854, 16,164, and 69,307 *k-mers* with *E. coli* isolates from humans, cattle, and chickens at a likelihood ratio test *p-value* less than 1.87×10^{-9}, 2.16×10^{-9}, and 1.9×10^{-9} respectively (reported as "lrt-pvalue"). There were no *k-mers* significantly associated with the pig host. The significant *k-mers* accounted for 426, 179, and 915 bacterial genes associated with isolation from human, cattle, and chicken hosts, respectively (Fig 2 and Table S3). An arbitrary cut-off of at least 10 *k-mers* mapped per gene was chosen to select genes for *in silico* functional characterization as well as COG assignment using Blast2GO[38] (Table S4) and CD-search[39,40] (Fig. S4).



A) Cattle host

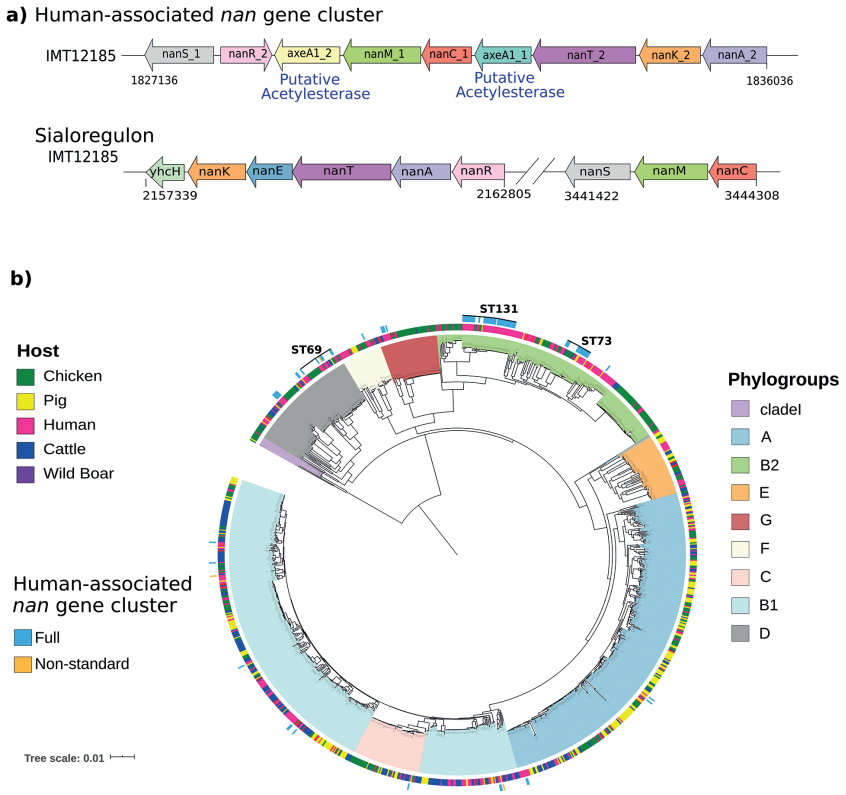


B) Human host



C) Chicken host

Figure 2. Plots representing the *E. coli* genes or gene variants associated with the a) Cattle host, b) Human host, and c) Chicken host. The bubble size represents the number of k-mers mapped to a specific gene, and the color gradient represents the effect size (ß).



C)

Our collection			RefSeq	
STs	**# Genomes**	**%**	**# Genomes**	**%**
131	35/42	83	1073/1256	85
73	14/17	82	253/372	68
69	6/31	19	155/302	51
38	3/6	50	57/196	29
10	4/109	4	66/1183	6
Other	24/993	3	512/14685	3
Total	**86/1198**	**7**	**2116/17994**	**12**

Figure 3. a) Genetic architecture of the human-associated *nan*-9 gene cluster and the sialoregulon on the complete genome of the strain IMT12185. The strain lacks the *nanXY* genes of the sialoregulon. b) Distribution of the human-associated *nan* gene cluster on core-genome phylogeny marked with STs with higher prevalence. c) The table indicates the prevalence of the human-associated *nan*-9 gene cluster in different STs in our collection and in the RefSeq *E. coli* genomes.

Association of novel nan genes with human host

GWAS revealed a strong association of nine contiguous genes, assigned to the group of *nan* genes with the human host (Fig 2B). Seven of these genes were annotated *in silico* as *nan* genes (Fig 3a) and the remaining two genes were annotated as being similar to *axeA1* of *Prevotella ruminicola* ATCC 19189 (Uniprot accession D5EV35). However, the amino acid sequences of the products of these *axeA1*-like genes only shared 19-20% similarity with AxeA1. Further investigation with EggNOG and CD search revealed an acetylesterase/lipase-encoding region (COG0657) in both genes and confirmed *nan* gene annotations. Previous evidence and the genomic location (i.e., between the *nan* genes; Fig 3a) suggest that these genes encode potential acetylesterases and may be analogous to sialyl esterases (NanS)[47]. Hence, these nine novel *nan* genes are collectively termed "human-associated *nan* gene cluster (*nan*-9)" (Fig 3a).

Distinct *nan* genes are present in *E. coli* and are also known as the sialoregulon (*nanRATEK-yhcH*, *nanXY [yjhBC],* and *nanCMS*; Fig 3a)[48]. The sialoregulon is known to be involved in metabolism of sialic acids[49–51], a diverse group of nine-carbon sugars, abundant in the glycocalyx of many animal tissues[52,53]. Sialic acids present on mucin proteins in the human gut are an essential energy source for many intestinal bacteria[54]. The proteins encoded by the seven genes of *nan*-9 (i.e. *nanAKTCMRS*) share 45-64% similarity with the corresponding *nan* genes of the sialoregulon in *E. coli* or the recently described phage-encoded *nanS*-p genes of enterohemorrhagic *E. coli*[55]. Both the human-associated *nan* gene cluster and the sialoregulon are located on the bacterial chromosome. The human-associated *nan* gene cluster was found in 7% of our isolate collection, whereas the genes comprising the sialoregulon were more common. In our collection, *nanXY* was identified in ~15% of isolates, *nanCMS* in ~93% of isolates, whilst *nanRATEKyhcH* was found in almost all (>99%) isolates.

The *nan*-9 cluster was detected in 86 isolates, mainly from phylogroups B2 and D (Fig 3b) and predominantly in isolates belonging to ST131, ST73, and ST69, both in our collection as well as across 17,994 RefSeq *E. coli* genomes (Fig 3c). The order and orientation of genes in the human-associated *nan* gene cluster were found to be identical in 82 out of 86 isolates (Fig. S5). In 63 isolates, insertion sequence (IS) 682 was found upstream, and in 23 isolates, IS2 was found downstream of this novel gene cluster (Fig. S5).

To further explore the function of the human-associated *nan-9* gene cluster, the entire cluster was knocked-out from strain IMT12185 (ST131), yielding strain IMT12185Δ*nan-9*. For comparison, an additional mutant, which lacked the *nanRATEK* locus from the sialoregulon (IMT12185Δ*nanRATEK*) was constructed from wild-type IMT12185. Correct gene deletion in both mutants was confirmed through WGS. No significant differences in carbon utilization and chemical sensitivity were observed between wild-type strain

IMT12185 and its mutant IMT12185Δ*nan-9* in Biolog phenotyping array experiments (PM1 and Gen III MicroPlates).

Deletion mutant IMT12185D*nan-9* was grown in MM with 0.2% 5-N-acetylneuraminic acid (Neu5Ac) or with 0.1% N-glycolylneuraminic acid (Neu5Gc) as sole carbon and energy source. Neu5Ac is the most common sialic acid of the glycocalyx of both humans and other mammals, whereas Neu5Gc is absent in humans. In the presence of Neu5Ac, mutant IMT12185D*nan-9* grew to a maximal OD_{600} of 1.34 comparable to that of parental strain IMT12185 (OD_{600} = 1.37). However, the mutant exhibited a delayed growth start of approximately three hours (Fig 4A). When Neu5Gc was offered as substrate, the mutant not only showed a similar growth start retardation, but also a slower growth rate and a lower maximal OD_{600} (1.31) in comparison with strain IMT12185 (OD_{600} = 1.43) (Fig 4B). Both Neu5Ac and Neu5Gc are degraded by the enzymatic activities of the enzymes NanRATEK, of which four, namely NanRATK, are encoded by redundant genes located on the determinants *nanRATEK* and *nan-9*. Deletion mutant IMT12185D*nanRATEK* was unable to grow with Neu5Ac (Fig 4C), demonstrating that *nan-9* alone is not sufficient for sialic acid degradation, probably due to a lack of *nanE* in the *nan-9* gene cluster. To exclude a pleiotropic effect of the *nan-9* deletion, parental strain IMT12185 and its

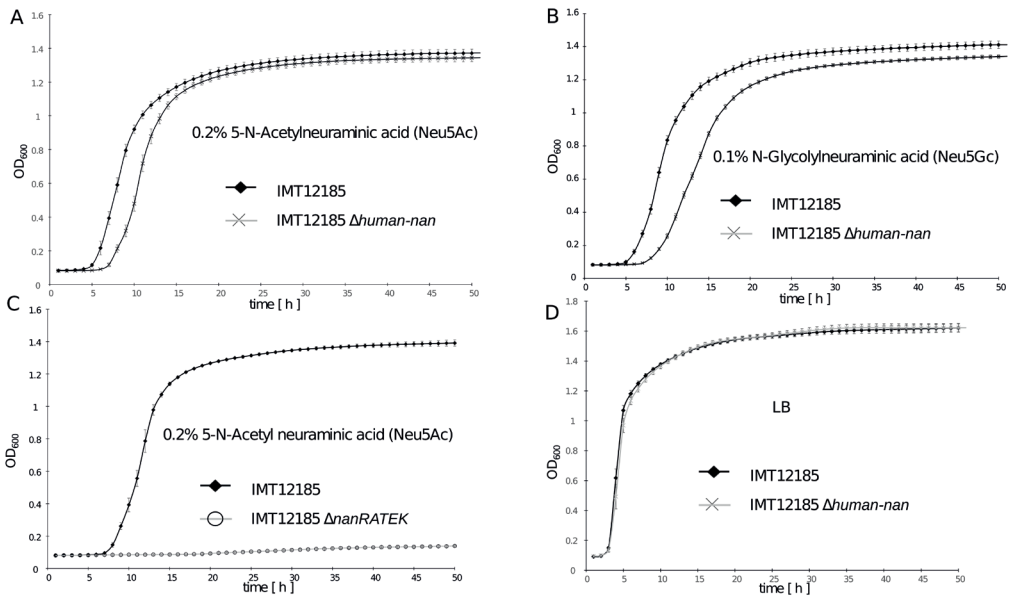


Figure 4. Growth curves of *E. coli* IMT12185 and its mutant derivatives in various media. a) Growth of IMT12185 and IMT12185 Δ*nan*-9 in M9 minimal medium with 0.2% 5-N-Acetylneuraminic acid (Neu5Ac) b) Growth of IM12185 and IMT12185 Δ*nan*-9 in M9 minimal medium with 0.1 5-N-Glycolylneuraminic acid (Neu5Gc) c) Growth of IMT12185 and IMT12185 Δ*nanRATEK* in M9 minimal medium with 0.2% 5-N-Acetylneuraminic acid (Neu5Ac) d) Growth of IMT12185 and IMT12185 Δ*nan*-9 in lysogeny broth (LB).

mutant IMT12185D*nan-9* were grown in LB medium. No significant difference was observed between the two growth curves (Fig 4D). These data demonstrate that the *nan-9* determinant of strain IMT12185 is biologically functional and contributes to the degradation of the sialic acids Neu5Ac and Neu5Gc.

Other genes associated with the human host

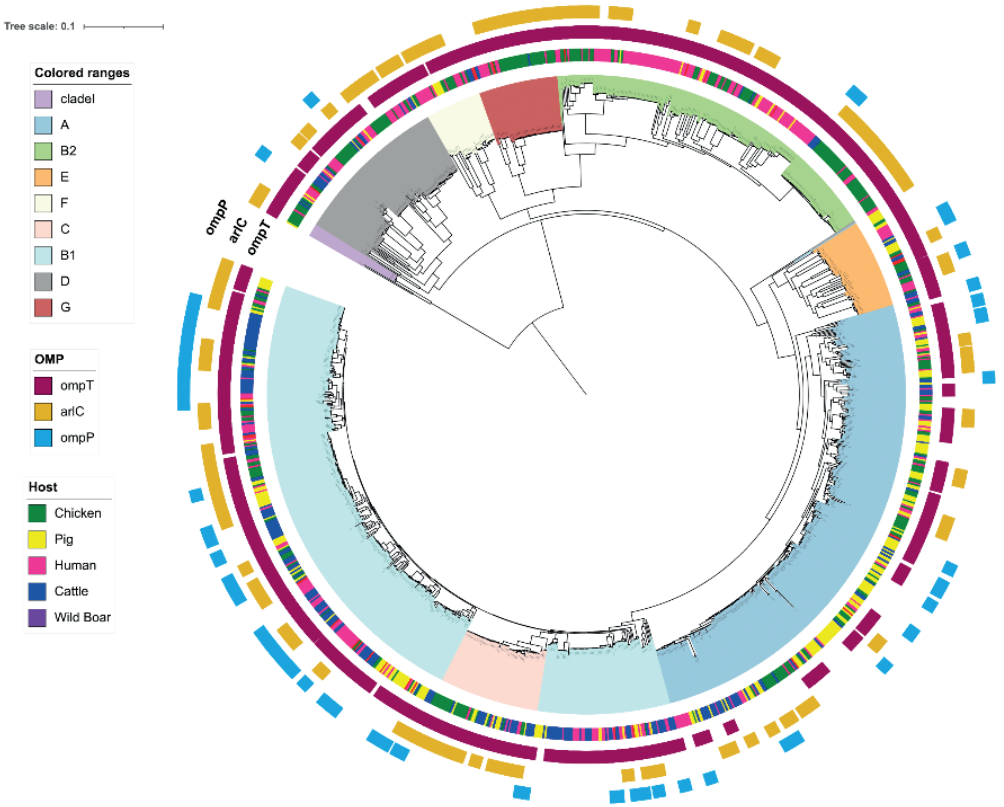
Several other genes associated with the human host were identified in the GWAS analysis, such as the *sat* gene encoding a serine protease autotransporter vacuolating toxin (Fig 2B)[56]. This gene was detected in 22.9% (n=75/327) of the human isolates in our collection and in only 0.59% (n=5/891) of the strains isolated from other hosts (Table S5). This gene was mainly detected in isolates belonging to specific lineages such as ST131, ST1193, and ST73 (Table S5). In addition, we found an association with two distinct homologs of the *macB* gene that encodes an ABC transporter[57] and is involved in many diverse processes, such as resistance to macrolides[58], lipoprotein trafficking[59], and cell division[60].

Association of distinct Omptins with the cattle and chicken hosts

We detected homologs of the *ompT* (encoding outer-membrane protease VII) gene, a member of the omptin family of proteases, in our dataset (Fig 2A & Fig 2C). Two homologs, *ompP* (UniProt accession P34210, sharing 70% amino acid identity with OmpT) and *arlC* (also referred to as *ompTp,* UniProt accession Q3L7I1, sharing 74% amino acid identity with OmpT), were found to be associated with the cattle and chicken hosts, respectively (Fig 5). In our collection, *ompP* was predominant in phylogroup B1 (n=68), whereas *arlC* was found in distinct phylogroups (such as B2, B1 and G) (Fig 5) and in isolates belonging to ST95 and ST117 (Table S6). A similar association was observed in 17,994 public *E. coli* genomes from RefSeq (Table S6). Previous studies have reported an increased prevalence of *arlC* (erroneously reported there as *ompT*) in a cluster of uropathogenic *E. coli* (UPEC) and avian pathogenic *E. coli* (APEC) classified as ST95[61]. Notably, *arlC* is associated with increased degradation of antimicrobial peptides (AMPs) in UPEC isolates[62]. OmpP is also able to degrade AMPs and displays a AMP cleavage specificity different from that of OmpT[63].

Association of genes involved in metal acquisition with the chicken host

GWAS analysis revealed an association of the *iroBCDEN* gene cluster (Fig 2C) with the chicken host, but not with other host species included in this study. The prevalence of the *iro* gene cluster was 24.3% (n=291/1198) in our collection, of which 61.5% (n=179/291) were from the chicken host. The gene cluster was found in different STs and with higher prevalence in STs such as ST117, ST95, ST23, and ST140 (Table S7). The chromosomal *iroBCDEN* gene cluster was first described in *Salmonella enterica* and is involved in uptake of catecholate-type siderophores, high-affinity iron-chelating molecules contributing to bacterial survival during infection by sequestering iron[64]. In *E. coli,* this gene cluster



Host	ompP(n)	arlC(n)	ompT(n)
Cattle (n=265)	7.2%(86)	1.3%(16)	12.3%(147)
Chicken (n=337)	0	17.4%(208)	18.9%(226)
Human (n=327)	0.4%(5)	2.1%(25)	19.9%(238)
Pig (n=240)	0.4%(5)	2.8%(33)	8.4%(101)
Wild Boar (n=29)	0.2%(2)	0	1.5%(18)
Total (n=1198)	8.2%(98)	23.5%(282)	60.9%(730)

Phylogroups	ompP prevalence(n)	arlC prevalence(n)
A (n=313)	15	29
B1 (n=366)	68	63
B2 (n=213)	4	80
C (n=62)	3	35
D (n=97)	2	19
E (n=55)	6	5
F (n=35)	0	9
G (n=49)	0	36
clade I (n=8)	0	6
Total (n=1198)	98	282

Figure 5. Distribution of *ompP*, *arlC*, and *ompT* genes in phylogroups and host across the phylogeny and their estimated prevalence.

has mainly been described in uropathogenic (UPEC) and avian pathogenic *E. coli* (APEC) and is regarded as a virulence factor[65]. The cluster has been reported on a chromosomal pathogenicity island, although in ExPEC, the cluster can also be located on ColV or ColBM virulence plasmids[66,67]. In addition, homologs of genes involved in zinc catabolism (*znuB*) and iron metabolism (*fes*) were found to be associated with the chicken host (Fig 2C).

Discussion

Escherichia coli can colonize many different ecological niches in a diverse range of host species, ranging from a commensal lifestyle to intra- or extra-intestinal infections. Presence of certain adhesin and other virulence-associated genes is well known to correlate with the relative ability of *E. coli* strains to colonize the intestinal tract of certain hosts (e.g., *ecp* for humans[68], F9 fimbriae and H7 flagellae for cattle[69,70] or Stg fimbriae for chickens[71]). Variations in host adaptation levels and their molecular basis in *E. coli* strains presumptively realizing a commensal-like lifestyle in the reservoir host are rarely described and poorly understood as of yet[72]. Commensal *E. coli* strains may be carriers of AMR and a source of mobile genetic elements conferring AMR to other bacteria including pathogenic strains in a shared microbiome, e.g. in the intestinal tract of animals including humans. We therefore collated an extensive and diverse dataset to identify genetic determinants of *E. coli* host adaptation. We observed significant enrichment of specific hosts within some phylogroups and STs in our collection. Furthermore, we unveiled correlations between the likelihood of genetically related isolates having been isolated from a certain host with the possession of distinctive genetic traits. Some of these traits, e.g. the *iroBCDEN* gene cluster, have been linked to *E. coli* and *Salmonella* virulence before, while others, in particular the human-associated *nan* gene cluster, are novel traits and have not been implicated in the infection and colonization process of *E. coli*. Of note, the latter gene cluster encodes for metabolic properties which have received little attention in bacterial infectious disease research. Specific metabolic properties have been linked to the relative ability of Shiga toxin-encoding *E. coli* (STEC) to asymptomatically colonize cattle, their reservoir host[73]. Unraveling the nutrient and energy flows in the complex interplay of intestinal bacteria, the surrounding microbiome and the host may open novel avenues to control the persistence and transmission of pathogenic and/or antimicrobial resistant bacteria[74].

We employed a *k-mer* based bacterial GWAS, applied in previous studies to associate multiple types of genetic variation with phenotypes[75,76]. In our study, we were able to associate a phenotype (i.e., isolates obtained from a certain host species) with the presence of specific genes, but not with sequence variation at the level of single nucleotide polymorphisms between genes. This lack of associations found at the SNP

level could possibly be explained by the fact that through our filtering approach to prevent false positive hits, we might have excluded *k-mers* that captured host-associated SNP variation. Secondly, it might be possible that since *E. coli* is genetically diverse, host-associated SNP variation is challenging to capture between unrelated strains. Finally, the absence of host-associated SNPs might be a biological observation, indicating that colonization of particular hosts is determined by gene presence or absence rather than minimal genetic variation within genetic elements. However, we were able to confirm previously published host associations, indicating the validity of our approach. For example, carriage of the salmochelin operon encoded by *iroBCDEN* and involved in iron metabolism was previously identified as associated with increased ability of *E. coli* strains to colonize chickens[65,77].

In addition to *iroBCDEN*, we found an association of omptin proteins (OmpP and ArlC) with chickens and cattle as hosts, respectively. Earlier studies using UPEC strains had demonstrated that these proteins are associated with cleavage and inactivation of cationic antimicrobial peptides (AMPs)[62]. Because AMPs are secreted as part of the host's innate immune response[78–80], these proteins may play a vital role in colonization. AMPs are also increasingly used as alternatives to antimicrobial agents in animal farming[81–83], further investigation into the contribution of these Omp variants to host colonization as well as to resistance to exogenous AMPs is warranted.

We did not identify any significant associations of *k-mers* with the pig host. Bacterial colonization of the porcine intestine by edema-disease *E. coli* (EDEC) is mediated by the ability of these bacteria to adhere to villous epithelial cells via their cytoadhesive F18 fimbriae[84]. The expression of receptors for these fimbriae on the apical enterocyte surface is inherited as a dominant trait among pigs and determines susceptibility to diseases caused by F18-fimbriated pathogenic *E. coli*[85]. Enterotoxigenic *E. coli* (ETEC) express F4 or F5 fimbriae with similar consequences[86]. However, we found only three, four and six isolates harbouring genes for F4, F5 and F18 fimbriae, respectively. Thus, we might not have had all *E. coli* pathovars associated with pig host sufficiently present in our collection, although we did observe an association between phylogroup A and pig colonization. An alternative reason might be that the association between phylogroup A and pig colonization complicated the identification of statistically significant *k-mers*. GWAS corrects for population structure, which means that if there is a strong association between lineage and phenotype, the genes harbored by that lineage will not be reported as having a strong association with the phenotype under study[87].

We identified a novel human host-associated *nan* gene cluster, distinct from the previously reported sialic acid (Sia) metabolic operon (*nanRATEK-yhcH*, *nanXY,* and *nanCMS)48*. This novel cluster is conserved and abundant in ExPEC lineages, such as ST131, ST73,

and ST69. The gene cluster is flanked by insertion sequences which might play a role in the horizontal exchange between different *E. coli* lineages. Knock-out *in vitro* studies indicated that this novel *nan-9* gene cluster contributes to catabolism of the sialic acids Neu5Ac and Neu5Gc, although it cannot replace the function of the *nanRATEK* locus which is abundant in *E. coli*. Hence, we hypothesize that *E. coli* harboring the *nan-9* gene cluster have an evolutionary advantage through either more efficient access to sialic acids or through access to more diverse sialic acids. The genes annotated as acetylxylan esterases are expected to represent novel sialyl esterases, as known sialyl esterases (*nanS* variants) have previously been mistaken for acetylxylan esterases[47]. Additional sialyl esterases – possibly with alternative deacetylation specificity – might provide a more efficient catabolism of acetylated sialic acids. Future studies should investigate the role of the human-associated *nan-9* gene cluster in the catabolism of differentially acetylated sialic acids and their relevance for the human host.

Approximately one-third of the isolates in our dataset were obtained from diseased hosts, while the remaining isolates were from healthy hosts. Many of the isolates in our dataset that originate from healthy hosts belong to ExPEC lineages which are typically considered to be pathogenic. In fact, the locus most strongly associated with the human host, the *nan-9* gene cluster, is abundant in ExPEC lineages. This does not necessarily mean that the *nan-9* gene cluster is associated with pathogenicity. In fact, this observation primarily supports the notion that these pathogenic *E. coli* are highly efficient colonizers of the human intestine[72]. Based on our results, we hypothesize that the human-associated *nan-9* gene cluster is one of the factors driving the adaptation of ExPEC to the human intestine.

Finally, we observed an association between the *sat* gene and human host colonization. Sat contributes to the pathogenicity of *E. coli* in the urinary tract[56]. The high prevalence of *sat* in previously studied *E. coli* isolates from the feces of healthy individuals suggests it may not act as a virulence factor in the human gut[88]. However, in our isolate collection, the *sat* gene was found in *E. coli* strains belonging to phylogroups A, B2, D, and F, which had been isolated from both healthy and diseased hosts (Table S5). Understanding the role of Sat in the colonization and adaptation of *E. coli* in healthy humans warrants further investigation.

Conclusion

Our study identified several distinct genetic determinants that may influence *E. coli* adaptation to different host species and provide an adaptive advantage. These findings are important as they aid the better understanding of the potential outcome of transmission events of *E. coli* between host species. This is particularly relevant for the control of the spread of antimicrobial resistant commensal and zoonotic *E. coli* strains within and across human and animal populations. The data generated here can also be used in risk analysis

and for diagnostic and monitoring purposes. More importantly, our study identified biological processes, including sialic acid catabolism, that should be investigated in more detail to better understand *E. coli* host adaptation.

Acknowledgements

The HECTOR research project was supported under the framework of the JPIAMR - Joint Programming Initiative on Antimicrobial Resistance – through the 3rd joint call, thanks to the generous funding by the Netherlands Organisation for Health Research and Development (ZonMw, grant number 547001012), the Federal Ministry of Education and Research (BMBF/DLR grant numbers 01KI1703A, 01KI1703C and 01KI1703B), the State Research Agency (AEI) of the Ministry of Science, Innovation and Universities (MINECO, grant number PCIN-2016-096), and the Medical Research Council (MRC, grant number MR/R002762/1).

Data availability

The raw-reads of the 1,090 *E. coli* isolates sequenced in this study were submitted to NCBI SRA with the Bioproject accession number PRJNA739205 and the SRA accession of 108 isolates, that were taken from other studies, were provided in supplement table S1.

References

1. Alteri, C. J. & Mobley, H. L. T. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. *Current Opinion in Microbiology* vol. 15 (2012).
2. Ewers, C., Bethe, A., Semmler, T., Guenther, S. & Wieler, L. H. Extended-spectrum ß-lactamase-producing and AmpC-producing *Escherichia coli* from livestock and companion animals, and their putative impact on public health: A global perspective. *Clinical Microbiology and Infection* (2012) doi:10.1111/j.1469-0691.2012.03850.x.
3. Bonnet, R. *et al.* Host Colonization as a Major Evolutionary Force Favoring the Diversity and the Emergence of the Worldwide Multidrug-Resistant *Escherichia coli* ST131 . *MBio* 12, (2021).
4. Lopatkin, A. J. *et al.* Persistence and reversal of plasmid-mediated antibiotic resistance. *Nat. Commun.* 8, (2017).
5. Pitout, J. D. D. & DeVinney, R. Escherichia coli ST131: A multidrug-resistant clone primed for global domination. *F1000Research* (2017) doi:10.12688/f1000research.10609.1.
6. Nicolas-Chanoine, M. H., Bertrand, X. & Madec, J. Y. *Escherichia coli* st131, an intriguing clonal group. *Clin. Microbiol. Rev.* (2014) doi:10.1128/CMR.00125-13.
7. Skurnik, D. *et al.* Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol. Biol. Evol.* (2016) doi:10.1093/molbev/msv280.
8. Riley, L. W. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. *Clinical Microbiology and Infection* vol. 20 (2014).
9. Cohen, E. *et al.* Pathoadaptation of the passerine-associated Salmonella enterica serovar Typhimurium lineage to the avian host. *PLoS Pathog.* 17, (2021).
10. Toft, C. & Andersson, S. G. E. Evolutionary microbial genomics: Insights into bacterial host adaptation. *Nature Reviews Genetics* (2010) doi:10.1038/nrg2798.
11. Sheppard, S. K. *et al.* Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc. Natl. Acad. Sci. U. S. A.* (2013) doi:10.1073/pnas.1305559110.
12. Mageiros, L. *et al.* Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. *Nat. Commun.* 12, (2021).
13. Salipante, S. J. *et al.* Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res.* (2015) doi:10.1101/gr.180190.114.
14. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel).* (2012) doi:10.3390/biology1030895.
15. Roehr, J. T., Dieterich, C. & Reinert, K. Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx330.
16. Andrews, S., Krueger, F., Seconds-Pichon, A., Biggins, F. & Wingett, S. FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics. *Babraham Institute* (2015).
17. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw354.
18. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* (2012) doi:10.1089/cmb.2012.0021.
19. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* (2013) doi:10.1093/bioinformatics/btt086.

20. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu153.

21. Beghain, J., Bridier-Nahmias, A., Nagard, H. Le, Denamur, E. & Clermont, O. ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb. Genomics* (2018) doi:10.1099/mgen.0.000192.

22. Page, A. J. *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv421.

23. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* (2014) doi:10.1007/978-1-62703-646-7_10.

24. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu033.

25. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* (2001) doi:10.1080/106351501753462876.

26. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* (2015) doi:10.1371/journal.pcbi.1004041.

27. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* (2013) doi:10.1093/molbev/mst028.

28. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. RhierBAPs: An R implementation of the population clustering algorithm hierbaps [version 1; referees: 2 approved]. *Wellcome Open Res.* (2018) doi:10.12688/wellcomeopenres.14694.1.

29. Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* (2008) doi:10.1007/s10479-011-0841-3.

30. Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* (2015).

31. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* (2007) doi:10.1093/bioinformatics/btl529.

32. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. genomics* (2016) doi:10.1099/mgen.0.000093.

33. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S Fourth edition by. World* vol. 53 (2002).

34. R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2020).

35. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: A comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty539.

36. Epping, L. *et al.* Genome-wide insights into population structure and host specificity of Campylobacter jejuni. *Sci. Rep.* 11, (2021).

37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp324.

38. Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* (2005) doi:10.1093/bioinformatics/bti610.

39. Marchler-Bauer, A. & Bryant, S. H. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* (2004) doi:10.1093/nar/gkh454.

40. Lu, S. *et al.* CDD / SPARCLE : the conserved domain database in 2020. 48, 265–268 (2020).

41. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* (2000) doi:10.1073/pnas.120163297.

42. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* (2017) doi:10.1371/journal.pcbi.1005595.

43. Chun, J. *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* (2018) doi:10.1099/ijsem.0.002516.

44. Liu, S. *et al.* Escherichia marmotae sp. nov., isolated from faeces of Marmota himalayana. *Int. J. Syst. Evol. Microbiol.* (2015) doi:10.1099/ijs.0.000228.

45. Ochman, H. & Selander, R. K. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* (1984) doi:10.1128/jb.157.2.690-693.1984.

46. Zhou, Z. *et al.* GrapeTree : visualization of core genomic relationships among 100 , 000 bacterial pathogens. 1395–1404 (2018) doi:10.1101/gr.232397.117.Freely.

47. Steenbergen, S. M., Jirik, J. L. & Vimr, E. R. YjhS (NanS) is required for *Escherichia coli* to grow on 9-O-acetylated N-acetylneuraminic acid. *J. Bacteriol.* (2009) doi:10.1128/JB.01000-09.

48. Kalivoda, K. A., Steenbergen, S. M. & Vimr, E. R. Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *J. Bacteriol.* (2013) doi:10.1128/JB.00692-13.

49. Vimr, E. R., Kalivoda, K. A., Deszo, E. L. & Steenbergen, S. M. Diversity of Microbial Sialic Acid Metabolism. *Microbiol. Mol. Biol. Rev.* (2004) doi:10.1128/mmbr.68.1.132-153.2004.

50. Vimr, E. R. & Troy, F. A. Identification of an inducible catabolic system for sialic acids (nan) in *Escherichia coli. J. Bacteriol.* (1985) doi:10.1128/jb.164.2.845-853.1985.

51. Bell, A. *et al.* Uncovering a novel molecular mechanism for scavenging sialic acids in bacteria. *J. Biol. Chem.* 295, (2020).

52. Vimr, E. R. Unified Theory of Bacterial Sialometabolism: How and Why Bacteria Metabolize Host Sialic Acids. *ISRN Microbiol.* (2013) doi:10.1155/2013/816713.

53. Severi, E., Hood, D. W. & Thomas, G. H. Sialic acid utilization by bacterial pathogens. *Microbiology* (2007) doi:10.1099/mic.0.2007/009480-0.

54. Haines-menges, B. L., Whitaker, W. B., Lubin, J. B. & Boyd, E. F. Host Sialic Acids: A Delicacy for the Pathogen with Discerning Taste. in *Metabolism and Bacterial Pathogenesis* (2015). doi:10.1128/microbiolspec.mbp-0005-2014.

55. Saile, N. *et al. Escherichia coli* O157:H7 strain EDL933 harbors multiple functional prophage-associated genes necessary for the utilization of 5-N-acetyl-9-O-acetyl neuraminic acid as a growth substrate. *Appl. Environ. Microbiol.* 82, (2016).

56. Guyer, D. M., Henderson, I. R., Nataro, J. P. & Mobley, H. L. T. Identification of Sat, an autotransporter toxin produced by uropathogenic *Escherichia coli*. *Mol. Microbiol.* (2000) doi:10.1046/j.1365-2958.2000.02110.x.

57. Kobayashi, N., Nishino, K. & Yamaguchi, A. Novel macrolide-specific ABC-type efflux transporter in *Escherichia coli*. *J. Bacteriol.* (2001) doi:10.1128/JB.183.19.5639-5644.2001.

58. Tikhonova, E. B., Devroy, V. K., Lau, S. Y. & Zgurskaya, H. I. Reconstitution of the *Escherichia coli* macrolide transporter: The periplasmic membrane fusion protein MacA stimulates the ATPase activity of MacB. *Mol. Microbiol.* (2007) doi:10.1111/j.1365-2958.2006.05549.x.

59. Khwaja, M., Ma, Q. & Saier, M. H. Topological analysis of integral membrane constituents of prokaryotic ABC efflux systems. *Res. Microbiol.* (2005) doi:10.1016/j.resmic.2004.07.010.

60. Yakushi, T., Masuda, K., Narita, S. I., Matsuyama, S. I. & Tokuda, H. A new ABC transporter mediating the detachment of lipid-modified proteins from membranes. *Nat. Cell Biol.* (2000) doi:10.1038/35008635.

61. Johnson, T. J. *et al.* Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. *Appl. Environ. Microbiol.* (2008) doi:10.1128/AEM.01395-08.

62. Desloges, I. *et al.* Identification and characterization of OmpT-like proteases in uropathogenic *Escherichia coli* clinical isolates. *Microbiologyopen* (2019) doi:10.1002/mbo3.915.

63. Hwang, B. Y. *et al.* Substrate specificity of the *Escherichia coli* outer membrane protease OmpP. *J. Bacteriol.* (2007) doi:10.1128/JB.01493-06.

64. Ratledge, C. & Dover, L. G. Iron Metabolism in Pathogenic Bacteria. *Annu. Rev. Microbiol.* (2000) doi:10.1146/annurev.micro.54.1.881.

65. Caza, M., Lépine, F., Milot, S. & Dozois, C. M. Specific roles of the iroBCDEN genes in virulence of an avian pathogenic *Escherichia coli* O78 strain and in production of salmochelins. *Infect. Immun.* (2008) doi:10.1128/IAI.00455-08.

66. Sorsa, L. J., Dufke, S., Heesemann, J. & Schubert, S. Characterization of an iroBCDEN gene cluster on a transmissible plasmid of uropathogenic *Escherichia coli*: Evidence for horizontal transfer of a chromosomal virulence factor. *Infect. Immun.* (2003) doi:10.1128/IAI.71.6.3285-3293.2003.

67. Dobrindt, U. *et al.* Genetic structure and distribution of four pathogenicity islands (PAI I536 to PAI IV536) of uropathogenic *Escherichia coli* strain 536. *Infect. Immun.* (2002) doi:10.1128/IAI.70.11.6365-6372.2002.

68. Rendón, M. A. *et al.* Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization. *Proc. Natl. Acad. Sci. U. S. A.* 104, (2007).

69. Low, A. S. *et al.* Cloning, expression, and characterization of fimbrial operon F9 from enterohemorrhagic *Escherichia coli* O157:H7. *Infect. Immun.* 74, (2006).

70. Mahajan, A. *et al.* An investigation of the expression and adhesin function of H7 flagella in the interaction of *Escherichia coli* O157: H7 with bovine intestinal epithelium. *Cell. Microbiol.* 11, (2009).

71. Lymberopoulos, M. H. *et al.* Characterization of Stg fimbriae from an avian pathogenic *Escherichia coli* O78:K80 Strain and assessment of their contribution to colonization of the chicken respiratory tract. *J. Bacteriol.* 188, (2006).

72. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nature Reviews Microbiology* (2010) doi:10.1038/nrmicro2298.

73. Barth, S. A. *et al.* Metabolic traits of bovine shiga toxin-producing *Escherichia coli* (STEC) strains with different colonization properties. *Toxins (Basel).* 12, (2020).

74. Stecher, B. Establishing causality in Salmonella-microbiota-host interaction: The use of gnotobiotic mouse models and synthetic microbial communities. *Int. J. Med. Microbiol.* 311, (2021).

75. Ma, K. C. *et al.* Adaptation to the cervical environment is associated with increased antibiotic susceptibility in Neisseria gonorrhoeae. *Nat. Commun.* (2020) doi:10.1038/s41467-020-17980-1.

76. Gröschel, M. I. *et al.* The phylogenetic landscape and nosocomial spread of the multidrug-resistant opportunist Stenotrophomonas maltophilia. *Nat. Commun.* (2020) doi:10.1038/s41467-020-15123-0.

77. Gao, Q. *et al.* Roles of iron acquisition systems in virulence of extraintestinal pathogenic *Escherichia coli*: Salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. *BMC Microbiol.* (2012) doi:10.1186/1471-2180-12-143.

78. McPhee, J. B. *et al.* Host defense peptide resistance contributes to colonization and maximal intestinal pathology by Crohn's disease-associated adherent-invasive *Escherichia coli*. *Infect. Immun.* (2014) doi:10.1128/IAI.01888-14.

79. Fjell, C. D. *et al.* Identification of novel host defense peptides and the absence of α-defensins in the bovine genome. *Proteins Struct. Funct. Genet.* (2008) doi:10.1002/prot.22059.

80. Lynn, D. J. *et al.* Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics* (2004) doi:10.1007/s00251-004-0675-0.

81. Li, Z., Hu, Y., Yang, Y., Lu, Z. & Wang, Y. Antimicrobial resistance in livestock: Antimicrobial peptides provide a new solution for a growing challenge. *Anim. Front.* (2018) doi:10.1093/af/vfy005.

82. Liu, Q. *et al.* Use of antimicrobial peptides as a feed additive for juvenile goats. *Sci. Rep.* (2017) doi:10.1038/s41598-017-12394-4.

83. Xiao, H. *et al.* The application of antimicrobial peptides as growth and health promoters for swine. *Journal of Animal Science and Biotechnology* (2015) doi:10.1186/s40104-015-0018-z.

84. Barth, S., Schwanitz, A. & Bauerfeind, R. Polymerase chain reaction-based method for the typing of f18 fimbriae and distribution of f18 fimbrial subtypes among porcine shiga toxin-encoding *Escherichia coli* in Germany. *J. Vet. Diagnostic Investig.* 23, (2011).

85. Frydendahl, K., Jensen, T. K., Andersen, J. S., Fredholm, M. & Evans, G. Association between the porcine *Escherichia coli* F18 receptor genotype and phenotype and susceptibility to colonisation and postweaning diarrhoea caused by E. coli O138:F18. *Vet. Microbiol.* 93, (2003).

86. Barth, S. *et al.* Virulence and fitness gene patterns of Shiga toxin-encoding *Escherichia coli* isolated from pigs with edema disease or diarrhea in Germany. *Berl. Munch. Tierarztl. Wochenschr.* 120, (2007).

87. Power, R. A., Parkhill, J. & De Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.* 18, 41–50 (2016).

88. Toloza, L. *et al.* The secreted autotransporter toxin (Sat) does not act as a virulence factor in the probiotic *Escherichia coli* strain Nissle 1917. *BMC Microbiol.* (2015) doi:10.1186/s12866-015-0591-5.

Supplementary material

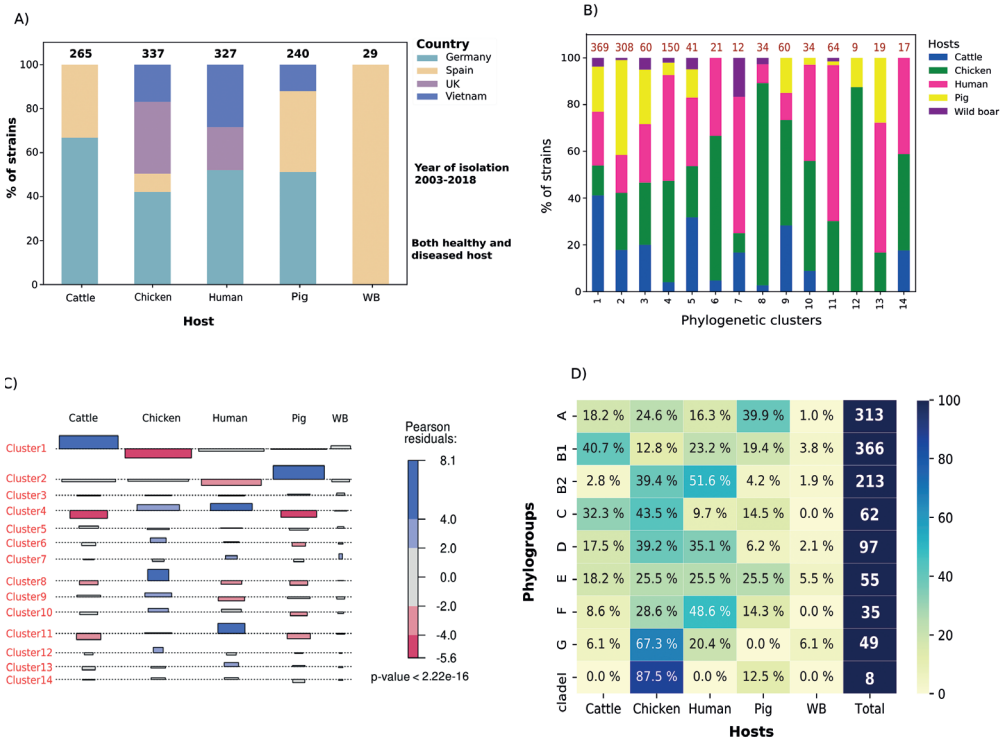


Figure S1. Distribution of 1,198 isolates and enrichment analysis: **A)** The plot represents the proportion of *E. coli* strains isolated from hosts in four countries. The number above each plot indicates the total number of isolates per host. **B)** Proportion of *E. coli* strains obtained from the five host species contributing to 14 phylogenetic clusters. The number on top of each bar represents the total number of isolates per cluster. **C)** Phylogenetic clusters enriched with a different host (Pearson residual > 0 represents positive correlation indicating the enrichment of certain host-species in distinct clusters at *p-value* <2.22e^{-16}). **D)** The proportion of *E. coli* isolates from different host species contributing to different phylogroups.

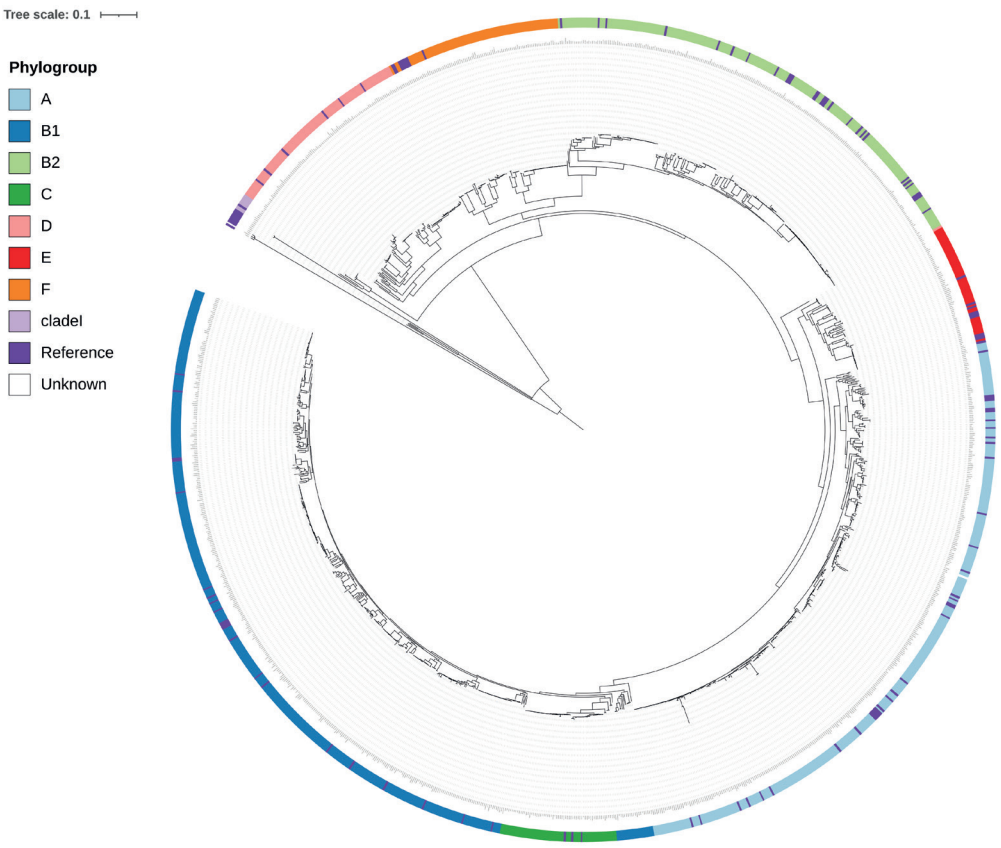


Figure S2. Core genome phylogeny of *E. coli* isolates of our collection (n=1,198) and reference strains (n=146) from the ECOR collection, RefSeq and cryptic clades annotated with their phylogroups (phylogroups were determined by ClermonTyper v. 1.3).

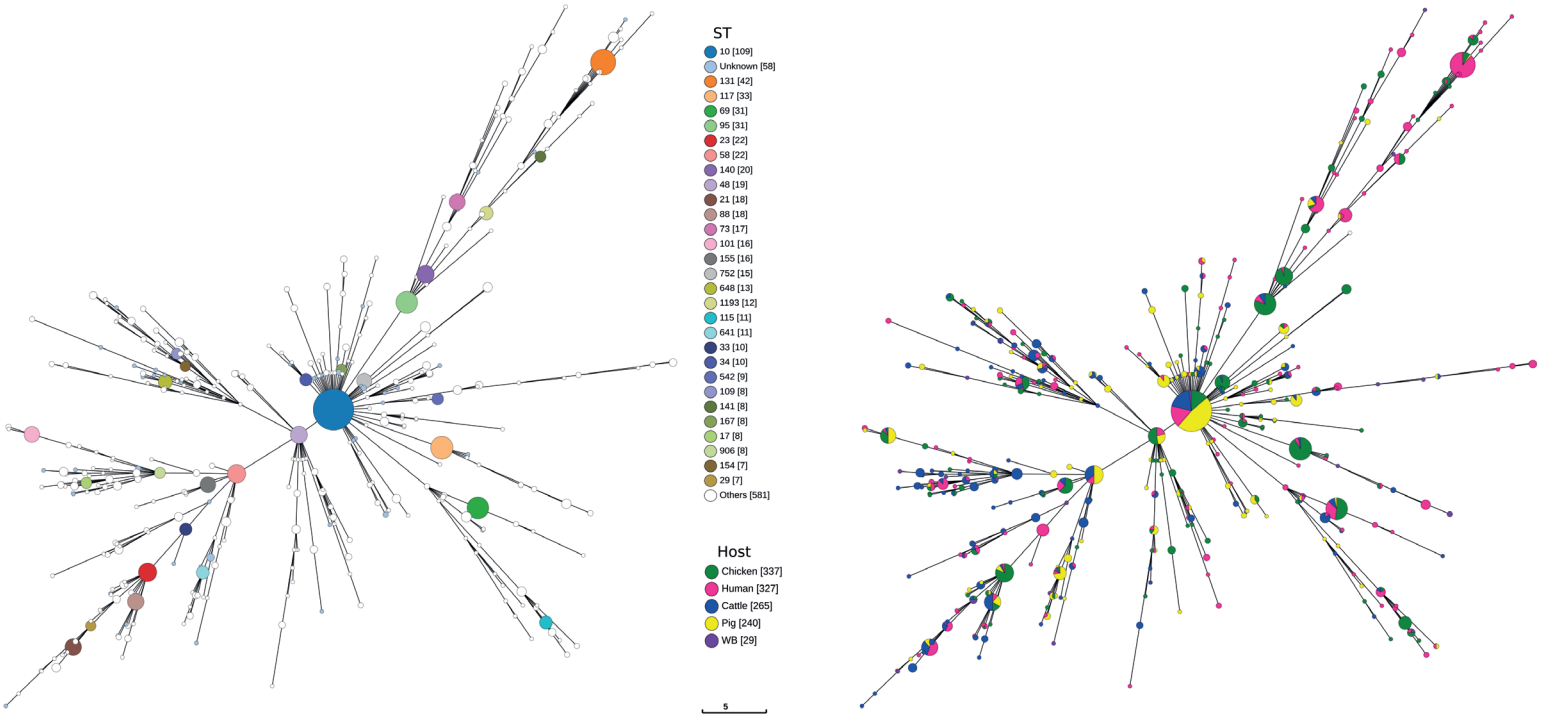


Figure S3. Minimum-spanning tree of MLST profiles of 1,198 *E. coli* isolates. Left: the number of isolates constituting an ST; Right: the proportion of hosts in each ST.

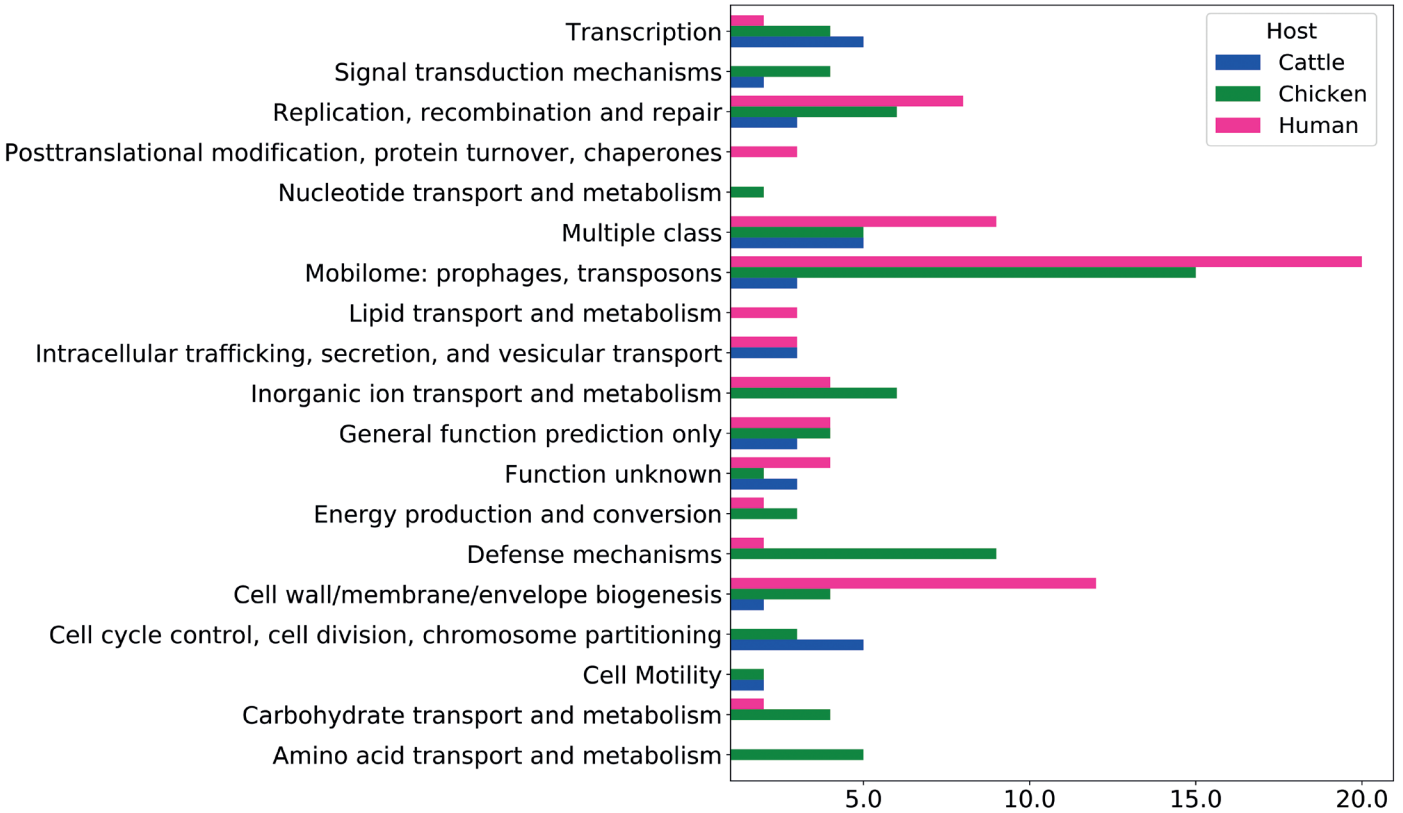


Figure S4. COG classification of genes or gene variants associated with cattle, chickens, and humans. Y-axis: Indicates the molecular function, and the x-axis indicates the number of genes in each functional class from each host.

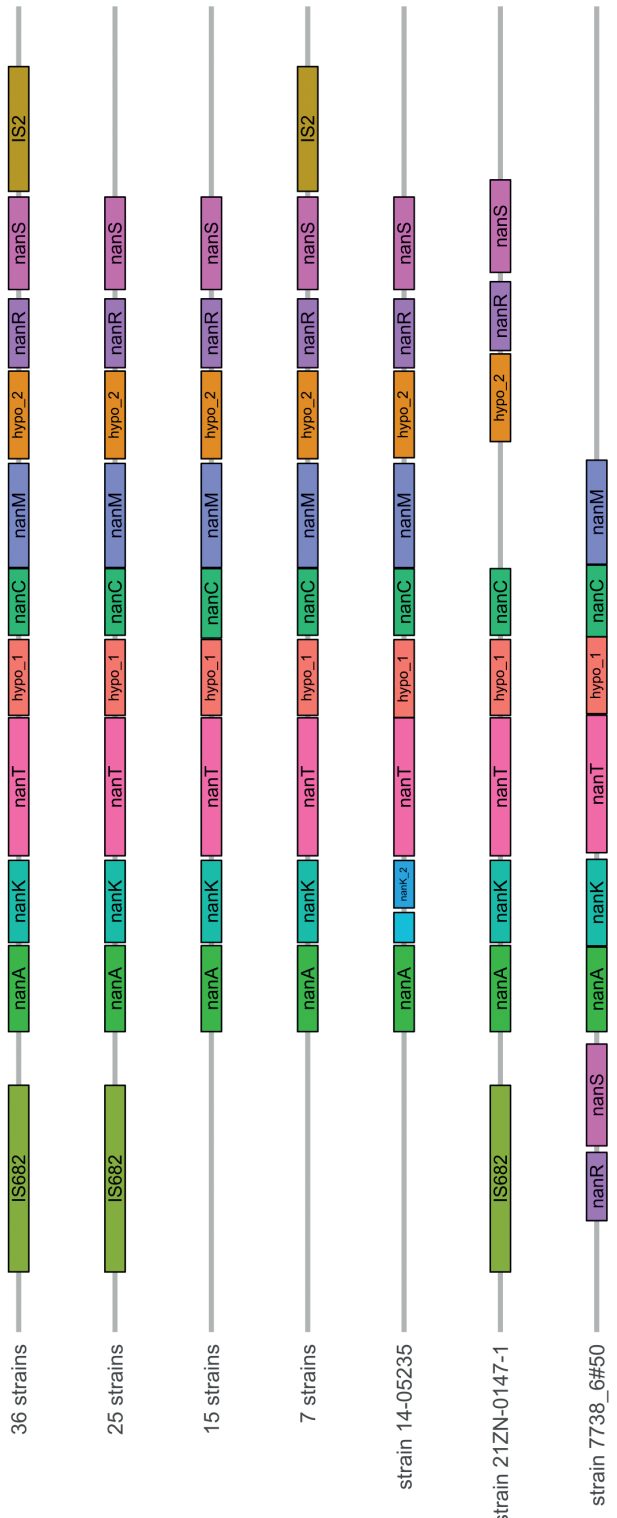


Figure S5. Genetic surroundings of the human-associated *nan* gene cluster in the genomes of all isolates in which it was identified in.

Table S1. Metadata of 1,198 *E. coli* isolates.

This table (1198 rows by 14 columns) is available from Figshare (tinyurl.com/48054cf).

Table S2. Genomes from our dataset used for annotating host-associated *k-mers*.

Host	Genome
Human	21224_3#267
	15-04719
	21224_3#237
	14-03445
	21ZN-0147-1
Cattle	ZTA1500328-1EC
	IMT33253
	IMT13376
	IMT30910
Chicken	14756
	21225_2#181
	Sap638

Table S3. Genes and genetic variants associated with different hosts.

This table (1521 rows by 10 columns) is available from Figshare (tinyurl.com/48054cf).

Table S4. Functional re-annotation of host-associated *E. coli* genes using Blast2go.

This table (246 rows by 5 columns) is available from Figshare (tinyurl.com/48054cf).

Table S5. The *E. coli* isolates (n=81) harboring the *sat* gene in our collection.

Strain	Country	Health	Host	year of isolation	ST	Phylogenetic cluster	Phylogroups
35110	Germany	Diseased	Human	2015	69	10	D
06-03041	Germany	Diseased	Human	2006	38	6	D
14-03445	Germany	Healthy	Human	2014	62	13	F
14-04252	Germany	Diseased	Human	2014	975	2	A
20222_6-191	Vietnam	Healthy	Human	NA	648	14	F
20222_7-277	Vietnam	Healthy	Human	NA	131	11	B2
20222_8-126	Vietnam	Healthy	Human	NA	73	4	B2
21224_2-340	Vietnam	Healthy	Human	NA	131	11	B2
21224_2-344	Vietnam	Healthy	Human	NA	131	11	B2
21224_2-88	Vietnam	Healthy	Human	NA	10	2	A
21224_3-256	Vietnam	Healthy	Human	NA	131	11	B2
21224_3-267	Vietnam	Healthy	Human	NA	1177	6	D
21224_3-370	Vietnam	Healthy	Human	NA	1193	4	B2
21225_2-129	Vietnam	Healthy	Human	NA	131	11	B2
21225_2-274	Vietnam	Healthy	Human	NA	1177	6	D
21225_2-28	Vietnam	Healthy	Human	NA	131	11	B2
21225_2-31	Vietnam	Healthy	Human	NA	131	11	B2
21225_2-74	Vietnam	Healthy	Human	NA	131	11	B2
21225_2-89	Vietnam	Healthy	Human	NA	10	2	A
21ZN-0026-2	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0037-2	Vietnam	Diseased	Human	NA	131	11	B2
21ZN-0052-2	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0071-2	Vietnam	Diseased	Human	NA	69	10	D
21ZN-0073-1	Vietnam	Diseased	Human	NA	131	11	B2
21ZN-0089-1	Vietnam	Diseased	Human	NA	131	11	B2
21ZN-0105-1	Vietnam	Diseased	Human	NA	38	6	D
21ZN-0134-1	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0147-1	Vietnam	Diseased	Human	NA	69	10	D
21ZN-0151-1	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0152-1	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0154-2	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0188-1	Vietnam	Diseased	Human	NA	69	10	D
21ZN-0188-2	Vietnam	Diseased	Human	NA	69	10	D
21ZN-0196-2	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0198-2	Vietnam	Diseased	Human	NA	131	11	B2
21ZN-0201-2	Vietnam	Diseased	Human	NA	131	11	B2
21ZN-0206-2	Vietnam	Diseased	Human	NA	1193	4	B2
21ZN-0220-2	Vietnam	Diseased	Human	NA	1193	4	B2
9352_7-29	Germany	Healthy	Human	2004	10	2	A
9352_7-48	Germany	Healthy	Human	2004	10	2	A
IMT12185	Germany	Diseased	Human	2006	131	11	B2

Strain	Country	Health	Host	year of isolation	ST	Phylogenetic cluster	Phylogroups
IMT12490	Germany	Diseased	Chicken	2007	73	4	B2
IMT13784	Germany	Healthy	Human	2007	10	2	A
IMT13798	Germany	Diseased	Human	2007	73	4	B2
IMT13800	Germany	Diseased	Human	2007	73	4	B2
IMT13882	Germany	Healthy	Human	2007	73	4	B2
IMT15474	Germany	Diseased	Pig	2008	393	5	D
IMT16218	Germany	Healthy	Pig	2003	73	4	B2
IMT16220	Germany	Healthy	Pig	2003	73	4	B2
IMT20024	Germany	Diseased	Cattle	2009	73	4	B2
IMT9270	Germany	Diseased	Human	2004	62	13	F
IMT9687	Germany	Diseased	Human	2004	34	2	A
SAP1372	UK	Diseased	Human	2017	131	11	B2
SAP1468	UK	Diseased	Human	2017	394	10	D
SAP1515	UK	Diseased	Human	2017	73	4	B2
SAP1597	UK	Diseased	Human	2017	69	10	D
SAP1609	UK	Diseased	Human	2017	73	4	B2
SAP1614	UK	Diseased	Human	2017	14	4	B2
SAP1621	UK	Diseased	Human	2017	131	11	B2
SAP1632	UK	Diseased	Human	2017	38	6	D
SAP1781	UK	Diseased	Human	2017	8455	11	B2
SAP1836	UK	Diseased	Human	2017	131	11	B2
SAP1847	UK	Diseased	Human	2017	131	11	B2
SAP1852	UK	Diseased	Human	2017	131	11	B2
SAP1858	UK	Diseased	Human	2017	131	11	B2
SAP1873	UK	Diseased	Human	2017	131	11	B2
SAP1887	UK	Diseased	Human	2017	131	11	B2
SAP1913	UK	Diseased	Human	2017	1193	4	B2
SAP1926	UK	Diseased	Human	2017	73	4	B2
SAP1953	UK	Diseased	Human	2017	131	11	B2
SAP2068	UK	Healthy	Human	2017	131	11	B2
SAP2072	UK	Healthy	Human	2017	62	13	F
SAP2081	UK	Healthy	Human	2017	1193	4	B2
SAP2082	UK	Healthy	Human	2017	10	2	A
SAP2089	UK	Healthy	Human	2017	131	11	B2
SAP2093	UK	Healthy	Human	2017	415	13	F
SAP2096	UK	Healthy	Human	2017	131	11	B2
SAP2098	UK	Healthy	Human	2017	73	4	B2
SAP2131	UK	Healthy	Human	2018	59	13	F
SAP2148	UK	Healthy	Human	2018	131	11	B2

Table S6. Prevalence of *ompP*, *arlC* and *ompT* in different STs in the RefSeq collection (n=17,994) and in our collection of *E. coli* isolates (n=1,198).

This table (1799 rows by 9 columns) is available from Figshare (tinyurl.com/48054cf).

Table S7. Prevalence of the *iroBCDEN* gene cluster in sequence types and associated hosts. Details are shown for sequence types (STs) with at least ten isolates harboring *iroBCDEN*.

ST	# Isolates harboring *iroBCDEN*	# Total isolates	%
117	32	33	96.9
95	21	31	67.7
23	20	22	90.9
140	18	20	90
69	17	31	54.8
73	14	17	82.3
101	10	19	52.6
58	10	22	45.4
Other STs	149	1003	14.6
Total	291	1198	24.3

Part III

Genomic Resources and Methodologies

Chapter 6

Benchmarking topological accuracy of bacterial phylogenomic workflows using *in silico* evolution

Boas C.L. van der Putten[$], Niek A.H. Huijsmans[$], Daniel R. Mende, Constance Schultsz

[$]Equal contribution

A revised version of this manuscript was published in *Microbial Genomics*, Volume 8, Issue 3, March 2022, https://doi.org/10.1099/mgen.0.000790

Abstract

Phylogenetic analyses are widely used in microbiological research, for example to trace the progression of bacterial outbreaks based on whole-genome sequencing data. In practice, multiple analysis steps such as *de novo* assembly, alignment and phylogenetic inference are combined to form phylogenetic workflows. Comprehensive benchmarking of the accuracy of complete phylogenetic workflows is lacking.

To benchmark different phylogenetic workflows, we simulated bacterial evolution under a wide range of evolutionary models, varying the relative rates of substitution, insertion, deletion, gene duplication, gene loss and lateral gene transfer events. The generated datasets corresponded to a genetic diversity usually observed within bacterial species (≥95% average nucleotide identity). We replicated each simulation three times to assess replicability. In total, we benchmarked seventeen distinct phylogenetic workflows using 8 different simulated datasets.

We found that recently developed k-mer alignment methods such as kSNP and SKA achieve similar accuracy as reference mapping. The high accuracy of k-mer alignment methods can be explained by the large fractions of genomes these methods can align, relative to other approaches. We also found that the choice of *de novo* assembly algorithm influences the accuracy of phylogenetic reconstruction, with workflows employing SPAdes or SKESA outperforming those employing Velvet. Finally, we found that the results of phylogenetic benchmarking are highly variable between replicates.

We conclude that for phylogenomic reconstruction k-mer alignment methods are relevant alternatives to reference mapping at species level, especially in the absence of suitable reference genomes. We show *de novo* genome assembly accuracy to be an underappreciated parameter required for accurate phylogenomic reconstruction.

Impact statement

Phylogenetic analyses are crucial to understand the evolution and spread of microbes. Among their many applications is the reconstruction of transmission events which can provide information on the progression of pathogen outbreaks. For example, to investigate foodborne outbreaks such as the 2011 outbreak of *Escherichia coli* O104:H4 across Europe. As different microbes evolve differently, it is important to know which phylogenetic workflows are most accurate when working with diverse bacterial data. However, benchmarks usually consider only a limited dataset. We therefore employed a range of simulated evolutionary scenarios and benchmarked seventeen phylogenetic

workflows on these simulated datasets. An advantage of our simulation approach is that we know *a priori* what the outcome of the analyses should be, allowing us to benchmark accuracy. We found significant differences between phylogenetic workflows and were able to dissect which factors contribute to phylogenetic analysis accuracy. Taken together, this new information will hopefully enable more accurate phylogenetic analysis of bacterial outbreaks.

Data summary

A Zenodo repository is available at https://doi.org/10.5281/zenodo.5036179 containing all simulated genomes, all alignments produced by phylogenetic workflows and .csv files summarising the topological accuracies of phylogenies produced based on these alignments. Code is available at https://github.com/niekh-13/phylogenetic_workflows.

Introduction

Phylogenetic analyses are crucial to assess the relatedness within a population of micro-organisms. These analyses provide information on the speciation, evolution and spread of microbes. Within clinical settings they can be used to identify microbial outbreaks and transmission events[1]. With the introduction of cost-efficient whole-genome sequencing (WGS), bacterial outbreak tracing is increasingly based on whole genome data, instead of on a small section of the genome such as 16S rRNA genes or a set of universal genes[2]. Whole-genome phylogenetic analysis can be applied by various pipelines or workflows, often composed of multiple separate tools. Common differences between workflows are which genomic loci are considered in the analysis (only protein-coding genes or also intergenic regions), how genetic features are defined (genes, k-mers, single nucleotide variants, etc.), but also how genomes are assembled. Benchmarking is necessary to make sense out of the plethora of bioinformatic methodologies available. Although previous benchmarks of bacterial phylogenetic reconstruction have generated important insights[3–5], some gaps remain. For example, the usefulness of recently developed k-mer alignment methods has not been fully explored in previous benchmark exercises. Additionally, the role of using different *de novo* assembly methods prior to comparative analysis has received little attention (especially in combination with the aforementioned k-mer alignment methods). Other methodological choices (e.g. choice of phylogenetic tree inference) have been amply studied before[3].

Benchmarking phylogenetic workflows requires knowledge of the true phylogenetic tree, as benchmarking results need to be compared to this reference. The true phylogenetic

tree is typically not known in real world settings. As such, various approaches have been proposed to determine or estimate the true phylogenetic tree of a set of strains. Some previous studies have assumed that the consensus of all phylogenies produced by the studied methods is close to the true phylogeny. Alternatively, studies have collated benchmark data sets where the epidemiological data was concordant with the phylogenomic analyses[4]. Because this approach uses real-life data, little is known about the underlying genetic events, and it does not allow to experimentally vary evolutionary parameters. Another approach is to have a mutant strain with an increased mutation rate evolve *in vitro*, and determine the structure of the true phylogeny from the experimental evolution controlled in the lab[5]. This approach provides a good grasp of the true phylogeny and allows the sampling of ancestral strains, but the method is costly and time-consuming and evolutionary parameters cannot be easily controlled. Finally, some studies have used *in silico* evolution to produce realistic sequencing data together with an *a priori* defined true phylogeny[3,4,6–9]. This approach offers the possibility to increase or decrease the rate of a range of evolutionary events, such as point mutations, indels, gene duplication, gene loss, gene translocation and lateral gene transfer. Additionally, genomic regions can be evolved under different evolutionary models, as is typical in real life scenarios (e.g. protein-coding genes vs. intergenic regions). Finally, this approach allows a comparison to the true phylogeny, which is not possible with other methods.

Several *in silico* evolution frameworks have been developed, with differing goals and strengths[3,7–13]. In the current study, we aimed to select a simulation strategy producing complete, haploid bacterial genomes. As lateral gene transfer is a common phenomenon in bacteria, simulation of lateral gene transfer should be included during *in silico* evolution. As we aim to compare against a true tree, the *in silico* evolution should be guided by a user-provided phylogenetic tree. We surveyed the Genetic Data Simulator database (https://surveillance.cancer.gov/genetic-simulation-resources/) and previously published manuscripts[3,4,7–11,13]. The workflow used by Lees *et al*. (2018)[3] was used as it satisfied all our criteria. The workflow combines ALF and DAWG software and enabled easy tuning of evolutionary parameters and setting simulation seeds for reproducible analysis.

In this study, we aim to assess which bioinformatic workflows are able to reconstruct the true phylogeny accurately under diverse evolutionary scenarios. We consider simulating evolution *in silico* to be the optimal approach to achieve this. We simulated the evolution of *Escherichia coli* genomes *in silico* under eight different scenarios, varying the rates of indels, gene duplication, gene loss and lateral gene transfer. We used these simulated datasets to assess the topological accuracy of seventeen phylogenetic reconstruction workflows, including *de novo* genome assembly, alignment or mapping, and finally phylogenetic tree inference. We included six alignment or mapping methods to identify single nucleotide polymorphisms (SNPs) between samples which can be subdivided

into k-mer alignment, reference mapping and gene-by-gene alignment methods. We also included three different *de novo* assembly approaches as the impact of this pre-processing step on phylogenomic accuracy is understudied.

Methods

Study design

This study consists of two main parts: simulation of *in silico* genome evolution (Fig 1A) and application of phylogenetic workflows on the simulated data sets (Fig 1B). A total of eight sets of parameters were used to simulate a variety of evolutionary processes on genic and intergenic regions separately, using the same phylogeny every time (Table S1). Each simulation was repeated three times with different random seeds to obtain technical replicates. From the *in silico* evolved genomes, short sequencing reads were generated. These sequencing reads were then used as input for the 17 phylogenetic workflows. We tested three *de novo* assembly algorithms in the workflows (Velvet, SKESA, SPAdes), alongside six methods for alignment or mapping (Snippy, Roary, PIRATE, SKA, kSNP, mlst-check). A total of seventeen phylogenetic workflows were tested (Table S2). All phylogenies have been inferred from alignments using IQ-Tree and ModelFinder. As the same phylogeny was used for each simulation, but the parameters for genetic events changed between simulations, each simulated dataset is expected to yield the same genetic distance between isolates (governed by the phylogeny), although the genetic events that have led to this identical genetic distance could be different (governed by the parameters).

In silico evolution

All code is available as a Snakemake v5.8.1[15] pipeline at https://github.com/niekh-13/phylogenetic_workflows. All tools were run using default parameters, unless otherwise noted. The complete chromosome of *Escherichia coli* K-12 MG1655 (RefSeq Assembly GCF_000005845.2) was used as the ancestral genome in all simulations. Evolution was simulated according to the phylogeny described in Kremer *et al.*[16]. The general approach used in this study was based on the approach described by Lees *et al.*[3]. The ancestral genome was annotated using Prokka v1.14.6[17] and subsequently divided into protein-coding genes and intergenic regions (all sequences not annotated as protein-coding gene). Protein-coding regions were *in silico* evolved using Artificial Life Framework v1.0 (ALF)[12] while intergenic regions were *in silico* evolved using DAWG v2.0.beta1[13].

ALF simulations were run using an empirical codon model, using a standard indel rate of 0.0252, a gene duplication and gene loss rate of 0.05, lateral gene transfer rates of 0.04 for single genes and 0.16 for groups of genes, and no spontaneous gene inversion or

gene translocation, based on previous bacterial simulations[3]. Complete specifications for the default run are available from https://github.com/niekh-13/phylogenetic_workflows/blob/master/input/alf_protein_sim.drw. Seven additional simulation were performed (Table S1): "Indel x 0.5" (halved indel rate), "Indel x 2" (doubled indel rate), "gene duplication x 2" (doubled gene duplication rate), "gene loss x 2" (doubled gene loss rate), "gene duplication x 2 & gene loss x 2" (doubled gene duplication and gene loss rates), "lateral gene transfer x 0.5" (halved lateral gene transfer rate for single genes and groups of genes), "lateral gene transfer x 2" (doubled lateral gene transfer rate for single genes and groups of genes).

DAWG simulations were run using a default indel rate of 0.00175 and evolved under a general time-reversible (GTR) model with rates A↔C: 0.91770, A↔G:4.47316, A↔T:1.10375, C↔G:0.56499, C↔T:6.01846, G↔T:1.00000, based on the GTR matrix inferred from dataset of nearly 1200 *E. coli* strains isolated from various host species (HECTOR study, manuscript in preparation). For simulations "indel x 0.5" and "indel x 2" the indel rate was appropriately changed (Table S1).

Per simulation, ALF and DAWG *in silico* evolution yielded protein-coding genes and intergenic regions for 96 *in silico* evolved genomes. These were assembled into 96 complete genomes. As stop codons are removed during ALF simulation, stop codons were inserted at the ends of genes. Paired-end sequencing reads in FASTQ format were simulated using ART v2016.06.05[18], based on a Illumina HiSeq 2500 profile with 30X depth, read length of 150 bp and a mean DNA fragment size of 600 bp with a standard deviation of 10 bp, using seed 21 (flags "-ss HS25 -na -rs 21 -p -l 150 -f 30 -m 600 -s 10").

For the generation of clonal datasets, we divided branch lengths of the true tree by factor 3, 30 and 100 corresponding to a median average nucleotide identity of 99.0%, 99.5% and 99.9% between genomes, respectively. Clonal datasets were generated using the standard rates for indel, gene duplication, gene deletion and lateral gene transfer events.

Comparing pipelines

From the simulated Illumina sequencing reads, phylogenies were reconstructed through seventeen workflows (Table S2). Assemblies were created using the Shovill v1.1.0 (https://github.com/tseemann/shovill) wrapper for Velvet v1.2.10[19], SPAdes v3.14.0 using "--isolate" mode[20] and SKESA v2.3.0[21]. Contigs were retained if they were 500bp or larger for all *de novo* assembly algorithms. Assembly quality metrics were assessed using Quast v5.0.2[22] and all versus all average nucleotide identity (ANI) comparisons were made using fastANI v1.2[23]. K-mer alignment methods kSNP v3.1[24] and SKA v1.0[25] were used on all assemblies, and SKA was additionally run on sequencing reads. In our study, both tools were used to extract k-mers of 31 bp from assemblies or sequencing reads.

Subsequently, these tools aligned k-mers of which the first and last 15 bp were identical, thus allowing only the middle base to vary between aligned k-mers. This k-mer alignment produced SNP alignments, which can be used for phylogenetic inference. Important to note is that although SKA and kSNP also employ k-mer-based methods, these methods are conceptually distinct from other k-mer-based tools such as Mash (https://github.com/marbl/Mash). The bacterial mapping pipeline Snippy v4.6.0 (https://github.com/tseemann/snippy) was used on sequencing reads alone, using the *Escherichia coli* K-12 MG1655 chromosome as reference (RefSeq Assembly GCF_000005845.2). As all genomes in the current study are simulated from this chromosome, this represents the most suitable reference. Gene-by-gene methods Roary v3.13.0[26] and PIRATE v1.0.3[27] were used on annotations produced by Prokka v1.14.6. Finally, alignments were constructed from multi-locus sequence type genes using mlst-check v2.1.1706216[28] and realigned using ClustalO v1.2.4[29]. All methods, including k-mer alignment methods, produce nucleotide alignments, which were subsequently used to infer phylogenies using IQ-tree v2.0.3[30] and ModelFinder[31] packaged with IQ-tree. Differences between the ground truth phylogeny and produced phylogenies were assessed using Robinson-Foulds distance calculation implemented in ape v5.4[32] and Kendall-Colijn distance calculation implemented in treespace v1.1.3.2[33]. All simulations and pipelines were run three times, with seeds 1, 42 and 1704 in ALF simulation. Alignment lengths were extracted using snp-sites v2.5.1[34].

Visual and statistical analysis

Parsing of results was performed using the pandas library v0.25.3[35] in Python v3.8.3 and using the tidyverse v1.3.0[36] and rstatix v0.6.0 (https://cran.r-project.org/package=rstatix) libraries in R v4.0.1. Results were plotted using ggplot2 v3.3.1[37], ggpubr v0.4.0 (https://cran.r-project.org/package=ggpubr), ggthemes v4.2.0[38], patchwork v1.0.1[39] and using SuperPlotsOfData[40]. Tests for statistical significance were carried out using the scipy library[41] using paired Wilcoxon ranked sum tests where indicated. Bonferroni correction for multiple testing was applied where applicable.

Results

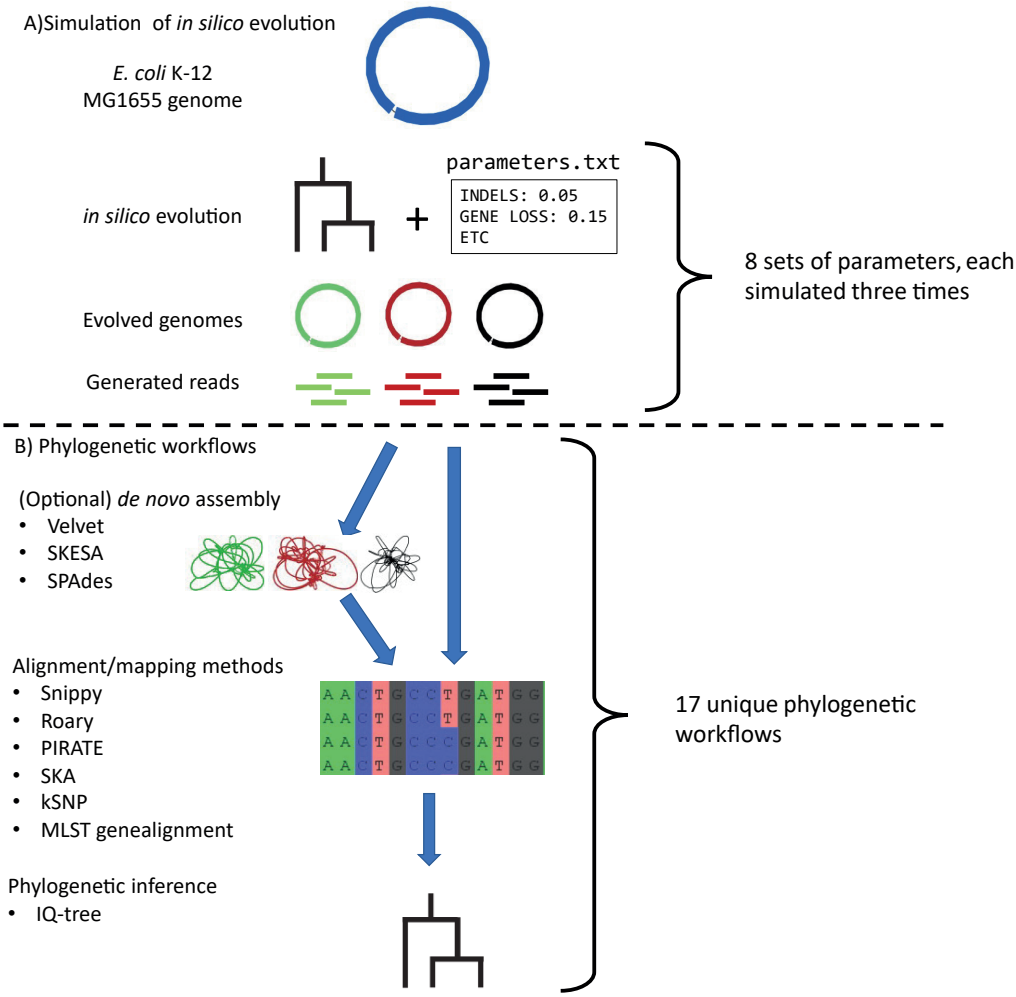


Figure 1. Overview of this study. **A)** Simulation of the in silico evolution. The *E. coli* K-12 MG1655 genome is evolved *in silico* according to a phylogeny (providing genetic distances) and a set of parameters controlling the rates of genetic events (providing which genetic events result in the genetic distance provided by the phylogeny). The resulting genomes are depicted by coloured complete genome graphs visualised in Bandage[42]. The complete genomes are subsequently shredded into sequencing reads. **B)** Phylogenetic workflows. Generated sequencing reads are assembled into draft genomes (coloured draft genome graphs) or directly mapped onto the ancestral genome. From alignments, phylogenetic trees are inferred using IQ-Tree.

Reference-based mapping and k-mer alignment methods yield phylogenetic trees most similar to ground truth

The *in silico* evolution yielded isolate sets with a genetic diversity comparable to a single bacterial species (≥95% average nucleotide identity[43], Figures 1 and S1). The same level of genetic diversity was attained between simulations, although these simulations included different rates of simulated genetic events (substitutions, indels, lateral gene transfer, etc., Table S3).

The optimal phylogenetic workflow should produce a phylogeny identical to the phylogeny which was used in the simulation process (the "ground truth phylogeny"). Per workflow, we calculated tree distance between the phylogeny produced by the workflow and the ground truth phylogeny. Tree distances were expressed in the Robinson-Foulds distance and the Kendall-Colijn metric.

The workflow showing the lowest tree distances across simulations employed SPAdes *de novo* assembly and subsequently SKA for k-mer alignment. After Bonferroni correction for multiple testing, the Kendall-Colijn metric of this workflow was significantly lower than all other workflows except Snippy, SPAdes + kSNP, SKESA + kSNP and SKESA + SKA (Fig 2 and Table S4). Notably, core gene alignment methods and methods employing Velvet for *de novo* assembly performed worse in our study. MLST gene alignment methods showed the highest deviation from the ground truth phylogeny as measured by Kendall Colijn metric and Robinson Foulds distance (Fig S2).

We also simulated more clonal datasets with a median ANI of 99.0%, 99.5% and 99.9%. As expected, the true tree was reconstructed less accurately when simulated genomes were more similar (Fig S3). The workflows which showed low Kendall-Colijn metrics between reconstructed phylogenies and the true tree showed a similar pattern in the clonal datasets, although differences are less clear than in Figure 2.

De novo assembly algorithms have a strong influence on accuracy of phylogenetic reconstruction

Next, we compared the accuracy of phylogenetic reconstruction between workflows employing different *de novo* assembly algorithms (Fig 3 and Table S5). Across eight simulations, workflows employing SPAdes and SKESA both resulted in significantly lower Kendall-Colijn metric values compared to the same workflows employing Velvet. In other words, workflows employing SPAdes and SKESA reconstruct phylogenies more accurately than the same workflows employing Velvet.

To gain insights in the *de novo* genome assembly quality, we compared the assemblies produced by Velvet, SKESA and SPAdes to the *in silico* evolved genomes from which

the sequencing reads were generated, using detailed assembly quality metrics such as total genome fraction, NGA50 (N50 of all blocks correctly aligned to the reference genome and corrected for reference genome length[22]), and the number of misassemblies alongside standard quality metrics such as number of contigs or total assembly size. We observed that although Velvet produced genome assemblies with a relatively high NGA50, Velvet also produced the highest number of misassemblies compared to SKESA or SPAdes (Table S6 and Fig S4). SPAdes seemed to perform best across multiple assembly quality metrics, reconstructing a large part of the original genome in few contigs (NGA50, genome fraction reconstructed, number of contigs), with a low number of errors (number of misassemblies).

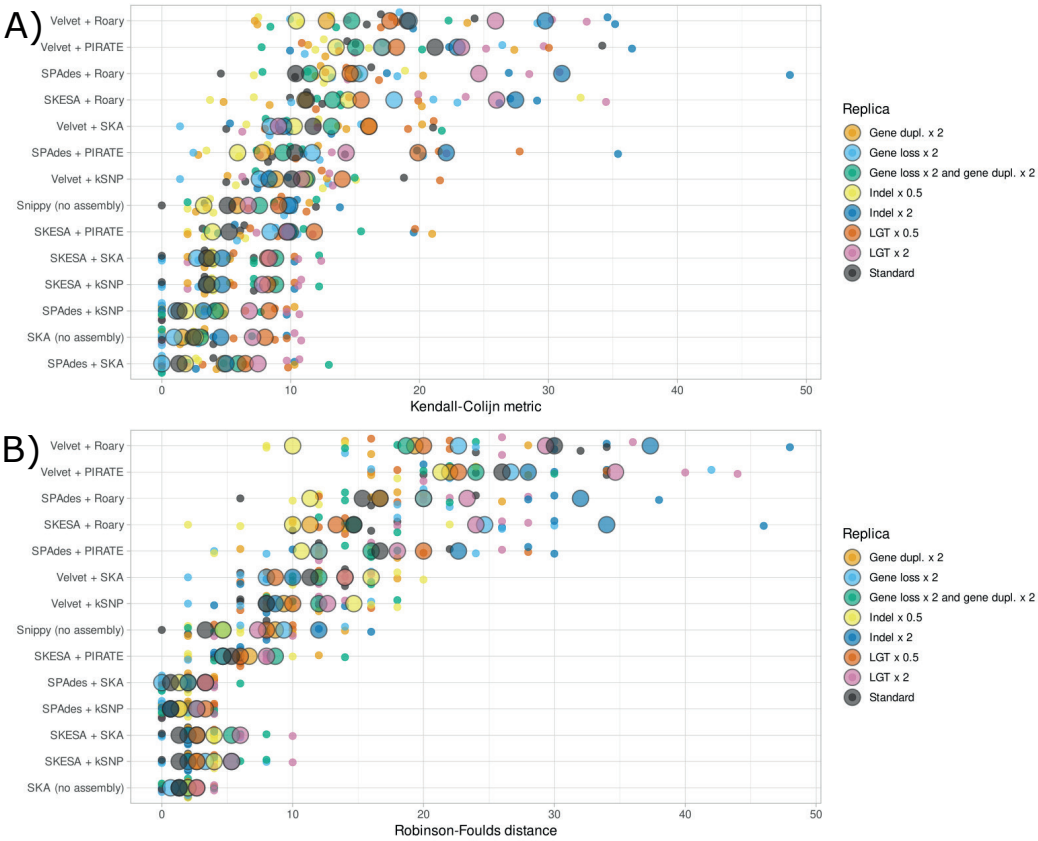


Figure 2. Kendall-Colijn metrics and Robinson-Foulds distances per phylogenetic workflow across eight simulations. Displayed distances are calculated between the ground truth phylogeny and the phylogeny produced by the relevant workflow. Generated using SuperPlotsOfData, ordered by median. Large circles indicate median of replicates. Small circles indicate separate measurements for replica.

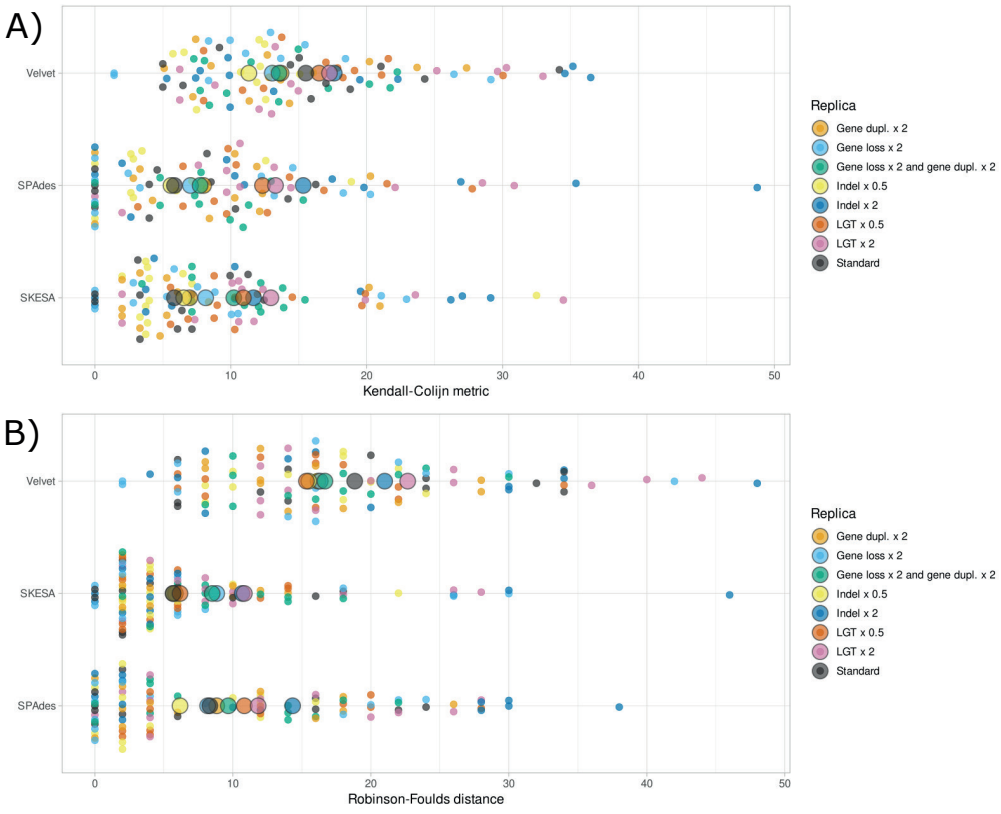


Figure 3. Kendall-Colijn metrics and Robinson-Foulds distances per *de novo* assembly algorithm used in workflows, across eight simulations. Displayed distances are calculated between the ground truth phylogeny and the phylogeny produced by the relevant workflow. Generated using SuperPlotsOfData, ordered alphabetically. Large circles indicate median of replicates. Small circles indicate separate measurements for replica.

Accuracy of phylogenetic reconstruction is associated with number of informative sites in the alignment

We hypothesised that the workflows using a larger part of the genome in the comparative analysis would yield larger alignments and more accurate phylogenetic reconstruction. To assess this, we extracted the alignment length produced per workflow. We found that the alignment length shows a strong negative correlation with the Kendall-Colijn metric and explains approximately 32% of variance in KC metric (R^2, Fig 4). This indicates that the methods that included a larger fraction of the genomes under study produced more accurate phylogenies. When the workflows employing MLST alignments were included, this negative correlation was even stronger (Fig S5).

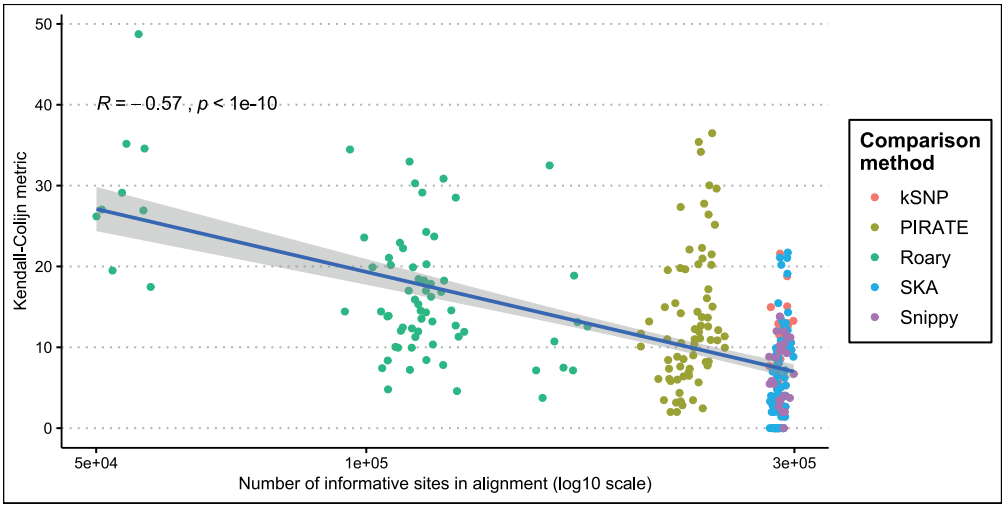


Figure 4. Count of informative sites in alignment plotted against Kendall-Colijn metric, with a linear model fitted (shading indicates 95% confidence interval). Pearson's Rho and associated p-value are shown.

Phylogenetic benchmarking shows a high variability between replicates

Repeating each of the eight simulations three times allows us to assess the reproducibility of this analysis. We see extensive variability in the accuracy of phylogenetic reconstruction even when comparing identical workflows across identical simulations, where only the starting seed for simulation differed (Fig 5). The largest difference between technical replicates reached a 31 point different in the Kendall-Colijn metric (SPAdes + Roary, simulation with double indel rate). Over 22% of Kendall-Colijn metric calculations were off more than 10 points between technical replicates.

Discussion

We present a systematic analysis of the accuracy of phylogenetic reconstruction of several workflows, based on simulated bacterial whole-genome data. We have included seventeen phylogenetic workflows. These were each benchmarked using eight simulation scenarios with three independent replicates.

First, we show that k-mer alignment methods provide a good alternative to reference-based mapping in species-level phylogenetic reconstruction. The high accuracy of workflows employing k-mer alignment seems to be due to the large fraction of genomes that can be utilised in these workflows, reflected by the high number of informative sites

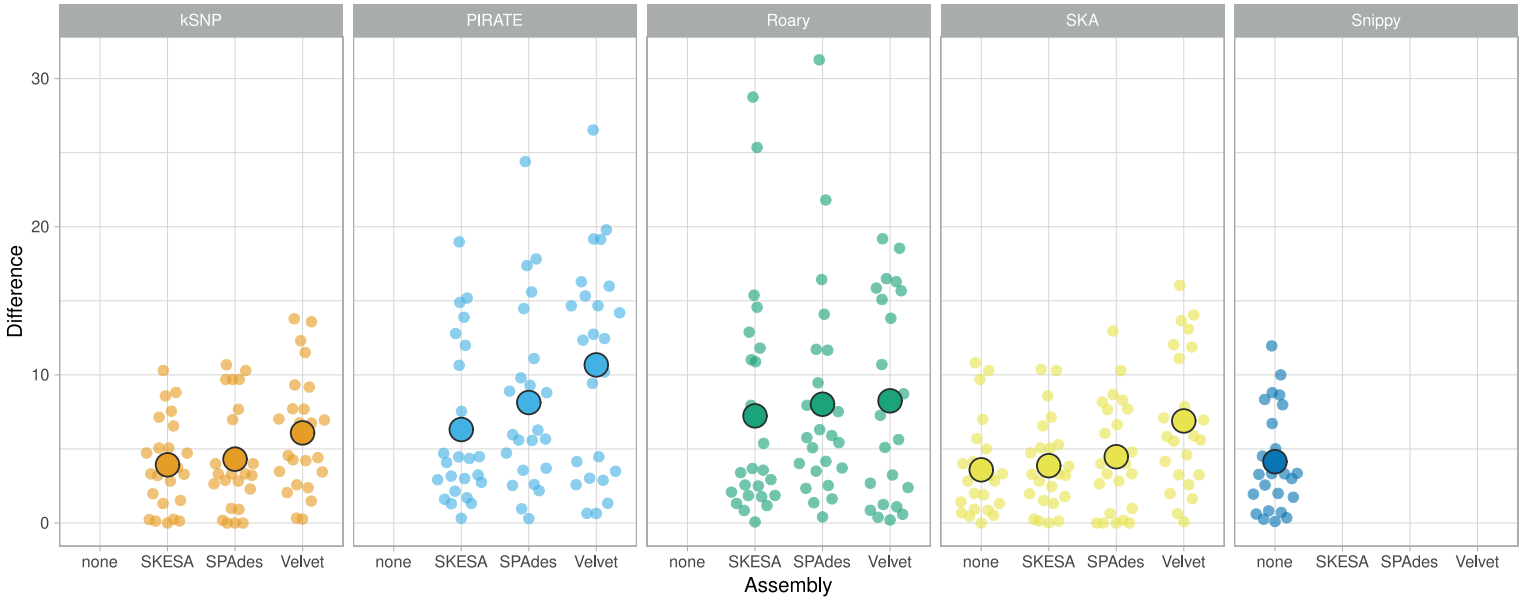


Figure 5. Differences between technical replicates for identical workflows across identical simulations, only differing in starting seed for the simulation. Workflows including MLST were excluded. Generated using SuperPlotsOfData.

in alignments produced by k-mer methods. In more clonal datasets, k-mer alignment methods also performed well. Through including eight simulation scenarios we were able to determine a clear influence of the *de novo* assembly algorithm on phylogenetic accuracy. Based on assembly quality evaluation, we hypothesise that an increased rate of misassemblies has a detrimental effect on phylogenetic accuracy. This also applies to k-mer alignment methods, which performed best when combined with either SPAdes or SKESA.

Surprisingly, we observed a high variability between replicates of phylogenetic workflows. Over one-fifth of comparisons showed differences of 10 points or more in the Kendall-Colijn metric. To contextualise, the difference in median Kendall-Colijn metric between the best and worst workflows in Fig 2A was 14.6 points. Generally, workflows using core gene alignment methods such as Roary or PIRATE displayed the highest discrepancies between replicates. This might be because core gene alignment methods need to employ heuristics to compare genes in an all-versus-all manner, which could introduce variability in their results.

Across seventeen phylogenetic workflows, eight simulations and three replicates, we reconstructed a total of 408 phylogenies. By including multiple workflows, simulations and replicates, this number increases quickly. We were able to limit computational workload by selecting only a single method (IQtree) to infer phylogenies from alignments. We chose to include only IQ-Tree because there was little difference between IQ-Tree, RAxML or other approaches in earlier studies[3], because IQ-Tree is widely used and thus represents an established method to infer phylogenies, and finally because IQ-Tree offers the identification of an optimal substitution model through ModelFinder. For the reference-based mapping analysis, we only used the Snippy pipeline in the current study. In a recent study by Bush *et al*.[44], Snippy was identified as the method which identified SNPs most accurately when reads were mapped to a reference genome representative for a species. For a larger genetic distance between reads and reference genome, Smalt and NextGenMap were identified as highly accurate read mapping algorithms (Table S9 of Bush *et al*. (2020)[44]). Given the small genetic distances between reads and reference in our study, which were smaller than in the study of Bush *et al*., we did not include Smalt and NextGenMap in our approach.

One of the challenges in benchmarking studies is to employ all methods in such a way that these can be compared sensibly. For k-mer alignment methods SKA and kSNP, we observed that configuring the desired k-mer length differs between tools. To obtain aligned k-mers of 31 bp, SKA requires to set k-mer length (flag "-k") to 15, resulting in the alignment of two split k-mers of 15 bp with a middle base, amounting to a total aligned k-mer of 31 bp. However, for kSNP the k-mer length (flag "-k") should be set to 31, to obtain

a 31 bp aligned k-mers of which the middle base may vary. Configuring the k-mer length correctly resulted in a highly similar accuracy of SKA and kSNP, while previous studies did not establish similar performance due to discrepancies in k-mer length configuration[25].

Determining the exact rates of genetic events such as point mutations or indels is challenging. In this study, we have evolved bacterial genomes across a range of evolutionary scenarios which means our results should be interpreted as generalisable findings, rather than findings specific to *Escherichia coli* and its evolutionary mechanisms.

Here we simulated datasets which exhibited a limited genetic diversity, similar to the genetic diversity observed within species (at least ~95% ANI[43]). In the context of more diverse datasets, for example comparing different species or genera, we expect that k-mer alignment methods would perform worse as these methods typically perform best with limited genetic diversity[25]. In accord with our results, we theorise that this is due to a faster decrease in informative sites with increasing evolutionary distance.

The current study focuses on the analysis of short sequencing reads specifically. However, previous studies have investigated the applicability of long read sequencing (especially Oxford Nanopore Technologies) for outbreak analysis[45,46]. Analysing long read sequence data uses fundamentally different algorithms and approaches than short read sequence data analysis,. Future studies could focus on the parameters that influence accuracy of phylogenetic reconstruction based on long read sequence data.

This study illustrates how phylogenetic reconstruction methods based on bacterial whole genome data compare. The simulations cover diverse evolutionary scenarios for bacterial species, providing detailed insight into the performance of phylogenetic reconstruction methods valid across diverse sets of bacterial strains. Recently developed k-mer alignment methods achieved similar accuracy as the gold standard (reference mapping) and thus seems to be a useful alternative when no suitable reference genome is available. Every microbe evolves according to different evolutionary parameters, so phylogenetic workflows need to be able to resolve many different evolutionary scenarios. Our study provides data on the accuracy of existing phylogenetic workflows and a framework to assess future phylogenetic workflows.

Author statements

Authors and contributors

Conceptualization: BCLP, DRM, CS. Data curation: BCLP, NAHH. Formal Analysis: BCLP, NAHH. Funding acquisition: CS. Investigation: BCLP, NAHH. Methodology: BCLP, NAHH,

DRM. Project administration: BCLP, DRM, CS. Software: BCLP, NAHH. Supervision: DRM, CS. Validation: BCLP, NAHH. Visualization: BCLP, NAHH. Writing – original draft: BCLP, NAHH. Writing – review & editing: BCLP, NAHH, DRM, CS.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Funding information

BP was supported through an internal AMC grant ("Flexibele OiO beurs"). The HECTOR research project was supported under the framework of the JPIAMR - Joint Programming Initiative on Antimicrobial Resistance – through the 3rd joint call, thanks to the generous funding by the Netherlands Organisation for Health Research and Development (ZonMw, grant number 547001012), the Federal Ministry of Education and Research (BMBF/DLR grant numbers 01KI1703A, 01KI1703C and 01KI703B), the State Research Agency (AEI) of the Ministry of Science, Innovation and Universities (MINECO, grant number PCIN-2016-096), and the Medical Research Council (MRC, grant number MR/R002762/1).

Acknowledgements

We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster. We thank the members of the HECTOR consortium for the use of the HECTOR data in inferring the GTR matrix used in the *in silico* evolution.

References

1. Harris SR, Feil EJ, Holden MTG, *et al.* Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science*. 2010;327(5964):469-474. doi:10.1126/science.1182395
2. Quainoo S, Coolen JPM, Hijum SAFT van, *et al.* Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin Microbiol Rev*. 2017;30(4):1015-1063. doi:10.1128/CMR.00016-17
3. Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, Harris SR. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res*. 2018;3. doi:10.12688/wellcomeopenres.14265.2
4. Timme RE, Rand H, Shumway M, *et al.* Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*. 2017;5:e3893. doi:10.7717/peerj.3893
5. Ahrenfeldt J, Skaarup C, Hasman H, Pedersen AG, Aarestrup FM, Lund O. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. *BMC Genomics*. 2017;18(1):19. doi:10.1186/s12864-016-3407-6
6. Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. *mBio*. 2014;5(6). doi:10.1128/mBio.02158-14
7. McTavish EJ, Pettengill J, Davis S, *et al.* TreeToReads - a pipeline for simulating raw reads from phylogenies. *BMC Bioinformatics*. 2017;18(1):178. doi:10.1186/s12859-017-1592-1
8. Nell LA. jackalope: A swift, versatile phylogenomic and high-throughput sequencing simulator. *Mol Ecol Resour*. 2020;20(4):1132-1140. doi:10.1111/1755-0998.13173
9. Escalona M, Rocha S, Posada D. NGSphy: phylogenomic simulation of next-generation sequencing data. *Bioinformatics*. 2018;34(14):2506-2507. doi:10.1093/bioinformatics/bty146
10. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genomics*. 2020;6(3):e000337. doi:10.1099/mgen.0.000337
11. Davín AA, Tricou T, Tannier E, de Vienne DM, Szöllősi GJ. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages. *Bioinformatics*. 2020;36(4):1286-1288. doi:10.1093/bioinformatics/btz710
12. Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C. ALF—A Simulation Framework for Genome Evolution. *Mol Biol Evol*. 2012;29(4):1115-1123. doi:10.1093/molbev/msr268
13. Cartwright RA. DNA assembly with gaps (Dawg): simulating sequence evolution. *Bioinformatics*. 2005;21(Suppl_3):iii31-iii38. doi:10.1093/bioinformatics/bti1200
14. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. doi:10.1093/bioinformatics/bts480
15. Mölder F, Jablonski KP, Letcher B, *et al.* Sustainable data analysis with Snakemake. *F1000Research*. 2021;10:33. doi:10.12688/f1000research.29032.1
16. Kremer PHC, Lees JA, Koopmans MM, *et al.* Benzalkonium tolerance genes and outcome in *Listeria monocytogenes* meningitis. *Clin Microbiol Infect*. 2017;23(4):265.e1-265.e7. doi:10.1016/j.cmi.2016.12.008
17. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153
18. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28(4):593-594. doi:10.1093/bioinformatics/btr708

19. Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829. doi:10.1101/gr.074492.107

20. Bankevich A, Nurk S, Antipov D, *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

21. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol*. 2018;19(1):153. doi:10.1186/s13059-018-1540-z

22. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086

23. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114. doi:10.1038/s41467-018-07641-9

24. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*. 2015;31(17):2877-2878. doi:10.1093/bioinformatics/btv271

25. Harris SR. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. *bioRxiv*. Published online October 25, 2018:453142. doi:10.1101/453142

26. Page AJ, Cummins CA, Hunt M, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31(22):3691-3693. doi:10.1093/bioinformatics/btv421

27. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *GigaScience*. 2019;8(giz119). doi:10.1093/gigascience/giz119

28. Page AJ, Taylor B, Keane JA. Multilocus sequence typing by blast from *de novo* assemblies against PubMLST. *J Open Source Softw*. 2016;1(8):118. doi:10.21105/joss.00118

29. Sievers F, Wilm A, Dineen D, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7(1):539. doi:10.1038/msb.2011.75

30. Minh BQ, Schmidt HA, Chernomor O, *et al*. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*. 2020;37(5):1530-1534. doi:10.1093/molbev/msaa015

31. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587-589. doi:10.1038/nmeth.4285

32. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2019;35(3):526-528. doi:10.1093/bioinformatics/bty633

33. Jombart T, Kendall M, Almagro-Garcia J, Colijn C. treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour*. 2017;17(6):1385-1392. doi:https://doi.org/10.1111/1755-0998.12676

34. Page AJ, Taylor B, Delaney AJ, *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genomics*. 2016;2(4). doi:10.1099/mgen.0.000056

35. pandas development Team. *Pandas-Dev/Pandas: Pandas*. Zenodo; 2019. doi:10.5281/zenodo.3509134

36. Wickham H, Averick M, Bryan J, *et al*. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686

37. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. Springer; 2016.

38. Arnold JB. ggthemes: Extra Themes, Scales and Geoms for "ggplot2." *R Package Version*. 2017;3(0).

39. Pedersen TL. patchwork: The Composer of ggplots. *R Package Version 00*. 2017;1.

40. Goedhart J. SuperPlotsOfData—a web app for the transparent display and quantitative comparison of continuous data from different conditions. *Mol Biol Cell*. 2021;32(6):470-474. doi:10.1091/mbc.E20-09-0583

41. Jones E, Oliphant T, Peterson P, others. *SciPy: Open Source Scientific Tools for Python*.; 2001. http://www.scipy.org/

42. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics*. 2015;31(20):3350-3352. doi:10.1093/bioinformatics/btv383

43. Chun J, Oren A, Ventosa A, *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol*. 68(1):461-466. doi:10.1099/ijsem.0.002516

44. Bush SJ, Foster D, Eyre DW, *et al.* Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. *GigaScience*. 2020;9(giaa007). doi:10.1093/gigascience/giaa007

45. Greig DR, Jenkins C, Gharbia SE, Dallman TJ. Analysis of a small outbreak of Shiga toxin-producing *Escherichia coli* O157:H7 using long-read sequencing. *Microb Genomics*. 2021;7(3):mgen000545. doi:10.1099/mgen.0.000545

46. Quick J, Ashton P, Calus S, *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol*. 2015;16(1):114. doi:10.1186/s13059-015-0677-2

Supplementary material

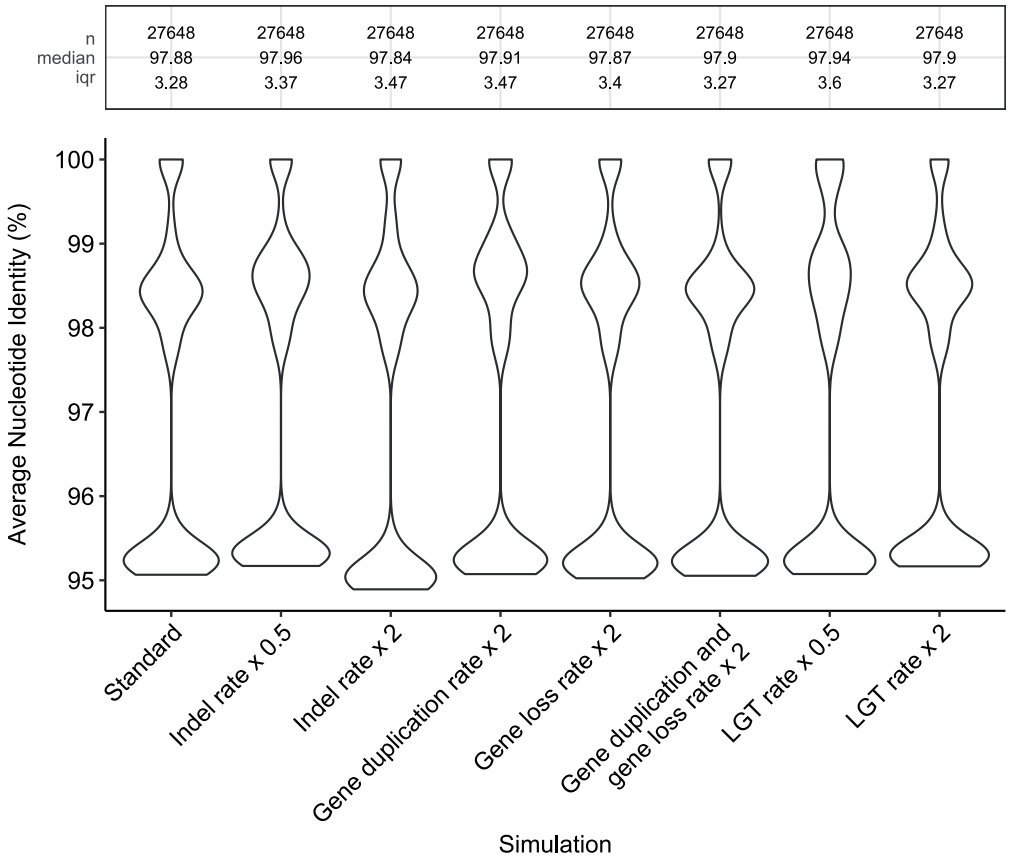


Figure S1. Violin plots of ANI comparisons made using FastANI, per simulation. From each simulation replicate (three replicates for eight simulations), all 96 in silico evolved genomes were compared in an all vs. all fashion.

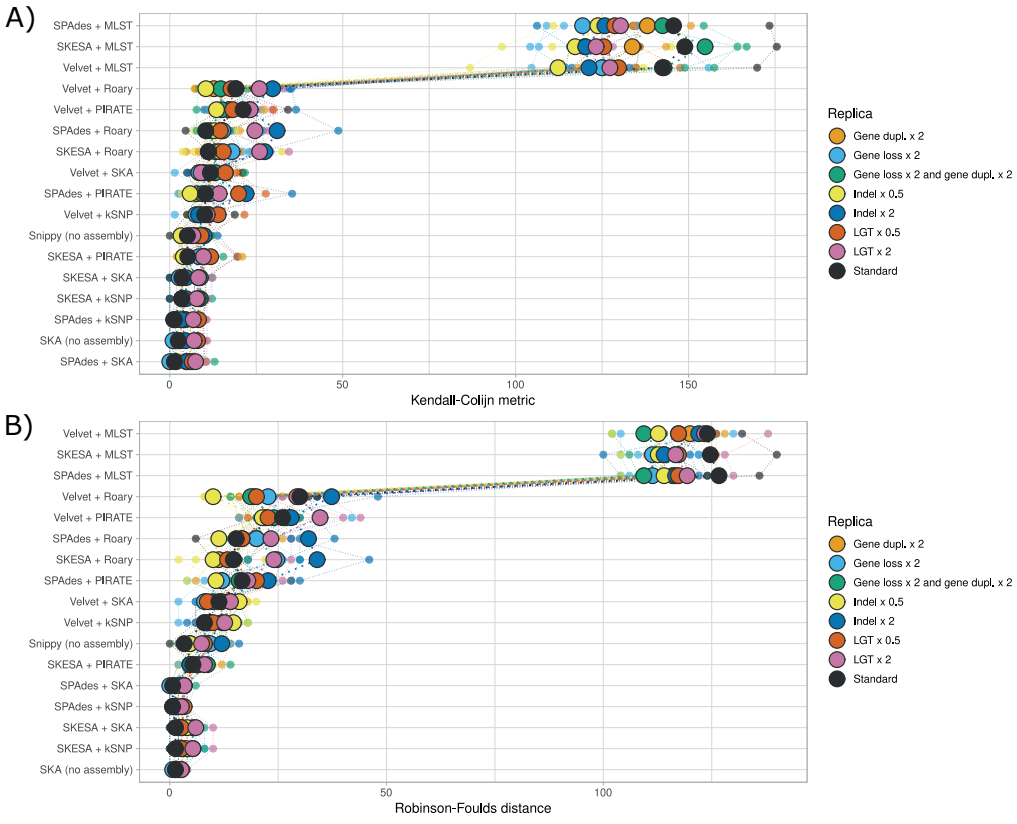


Figure S2. Kendall-Colijn metrics and Robinson-Foulds distances between the ground truth phylogeny and phylogenies produced by workflows, across eight simulations.

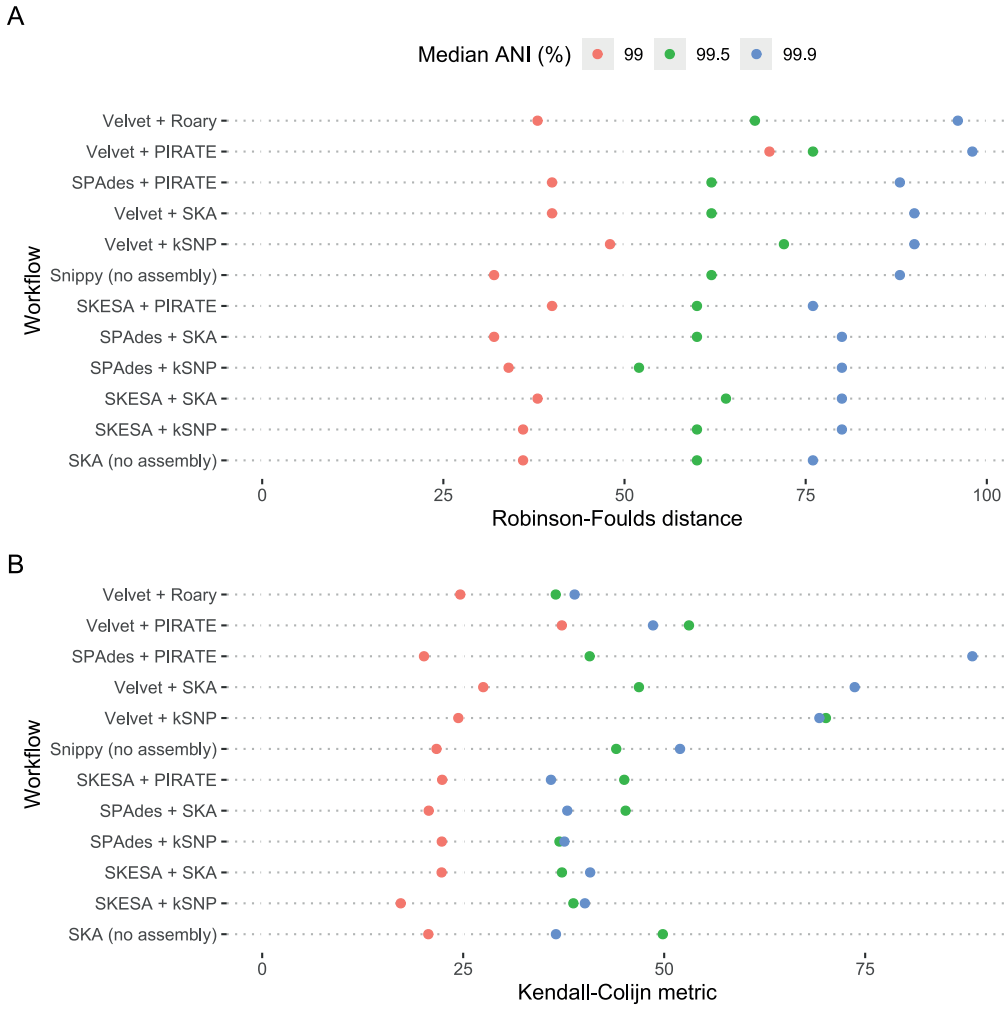


Figure S3. Kendall-Colijn metrics and Robinson-Foulds distances per phylogenetic workflow for clonal simulated data. Displayed distances are calculated between the ground truth phylogeny and the phylogeny produced by the relevant workflow.

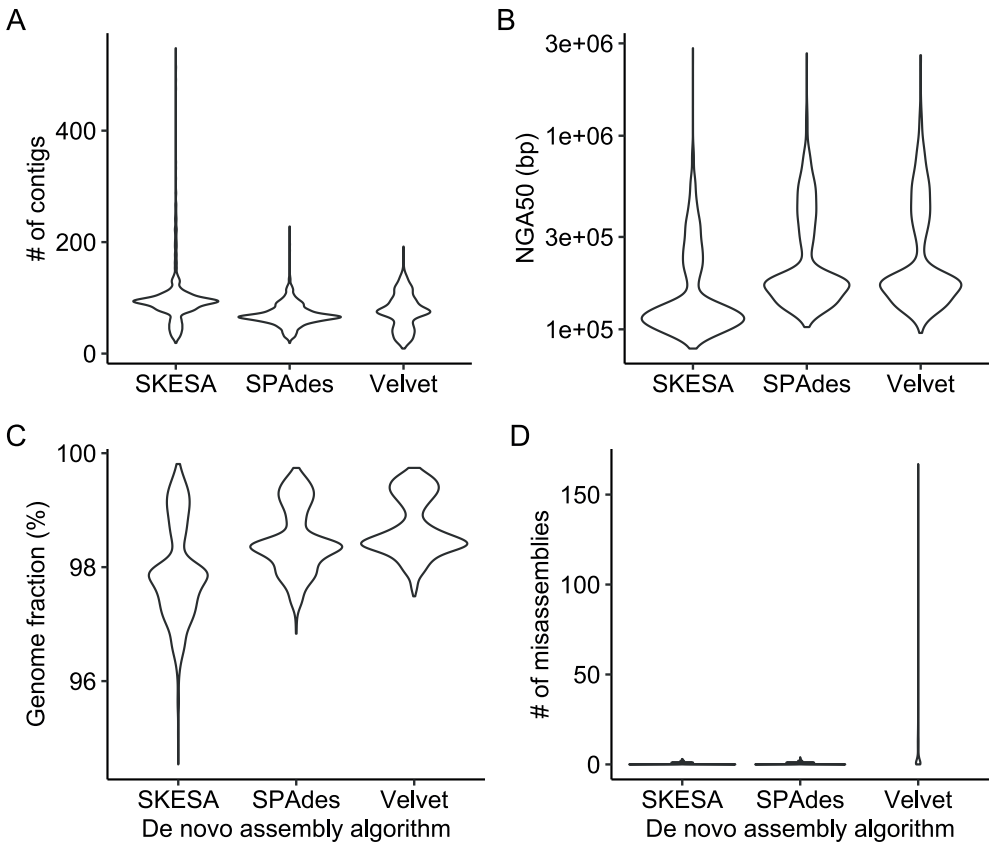


Figure S4. Comparison of SKESA, SPAdes and Velvet algorithms for de novo genome assembly, based on number of contigs, NGA50, genome fraction reconstructed and number of misassemblies.

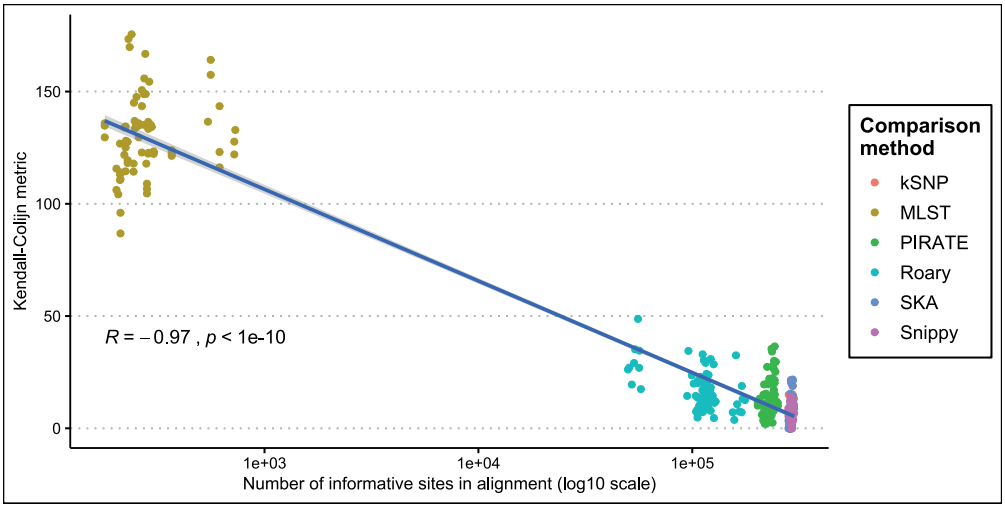


Figure S5. Count of informative sites in alignment plotter against Kendall-Colijn metric, with a linear model fitted (shading indicates 95% confidence interval). Pearson's Rho and associated p-value are shown.

Table S1. Parameters which differed per simulation. All other parameters were kept stable.

	Genes (ALF)					Intergenic regions (DAWG)
Simulation	**Indel rate**	**geneDuplRate**	**geneLossRate**	**lgtRate**	**lgtGRate**	**Indel rate**
Standard	0.0252	0.05	0.05	0.04	0.16	0.00175
Indel x 0.5	0.0126	0.05	0.05	0.04	0.16	0.000875
Indel x 2	0.0504	0.05	0.05	0.04	0.16	0.0035
LGT x 0.5	0.0252	0.05	0.05	0.02	0.08	0.00175
LGT x 2	0.0252	0.05	0.05	0.08	0.32	0.00175
Gene dupl. x 2	0.0252	0.1	0.05	0.04	0.16	0.00175
Gene loss x 2	0.0252	0.05	0.1	0.04	0.16	0.00175
Gene loss x 2 and gene dupl. x 2	0.0252	0.1	0.1	0.04	0.16	0.00175

Table S2. Seventeen phylogenetic workflows included in this study, with software versions.

Workflow	*de novo* genome assembly algorithm	Annotation method	Comparison method
Snippy (no assembly)	NA	NA	Snippy vs.4.6.0
Velvet + kSNP	Shovill v1.1.0 using Velvet v1.2.10	NA	kSNP v3.1
SPAdes + kSNP	SPAdes v3.14.0	NA	kSNP v3.1
SKESA + kSNP	SKESA v2.3.0	NA	kSNP v3.1
Velvet + PIRATE	Shovill v1.1.0 using Velvet v1.2.10	Prokka v1.14.6	PIRATE v1.0.3
SPAdes + PIRATE	SPAdes v3.14.0	Prokka v1.14.6	PIRATE v1.0.3
SKESA + PIRATE	SKESA v2.3.0	Prokka v1.14.6	PIRATE v1.0.3
Velvet + Roary	Shovill v1.1.0 using Velvet v1.2.10	Prokka v1.14.6	Roary v3.13.0
SPAdes + Roary	SPAdes v3.14.0	Prokka v1.14.6	Roary v3.13.0
SKESA + Roary	SKESA v2.3.0	Prokka v1.14.6	Roary v3.13.0
Velvet + SKA	Shovill v1.1.0 using Velvet v1.2.10	NA	SKA v1.0
SPAdes + SKA	SPAdes v3.14.0	NA	SKA v1.0
SKESA + SKA	SKESA v2.3.0	NA	SKA v1.0
SKA (no assembly)	NA	NA	SKA v1.0
Velvet + MLST	Shovill v1.1.0 using Velvet v1.2.10	NA	mlst-check v2.1.1706216
SPAdes + MLST	SPAdes v3.14.0	NA	mlst-check v2.1.1706216
SKESA + MLST	SKESA v2.3.0	NA	mlst-check v2.1.1706216

Table S3. Counts of genetic events across replicates of eight simulations.

Simulation	Replicate	# of lateral gene transfers	# of insertions	# of deletions	# of genes lost	# of genes duplicated
Standard	Run 1	575	7750	7567	167	144
Standard	Run 2	515	7838	7662	142	161
Standard	Run 3	534	7790	7721	145	114
Indel x 0.5	Run 1	575	3864	3890	171	152
Indel x 0.5	Run 2	515	3909	3829	142	161
Indel x 0.5	Run 3	534	3838	3851	145	114
Indel x 2	Run 1	575	15302	15644	171	152
Indel x 2	Run 2	515	15239	15507	142	161
Indel x 2	Run 3	534	15128	15480	145	114
LGT x 0.5	Run 1	256	7620	7705	135	194
LGT x 0.5	Run 2	308	7810	7663	126	164
LGT x 0.5	Run 3	289	7638	7589	129	126
LGT x 2	Run 1	1189	8046	8321	157	133
LGT x 2	Run 2	1206	7947	7966	152	157
LGT x 2	Run 3	1240	8236	8163	181	146
Gene dupl. x 2	Run 1	544	8018	8001	114	356
Gene dupl. x 2	Run 2	575	7815	7743	125	266
Gene dupl. x 2	Run 3	585	7754	7784	183	238
Gene loss x 2	Run 1	504	7776	7727	323	155
Gene loss x 2	Run 2	571	7655	7698	251	154
Gene loss x 2	Run 3	576	7688	7596	304	124
Gene loss x 2 and gene dupl. x 2	Run 1	585	7747	7815	265	306
Gene loss x 2 and gene dupl. x 2	Run 2	615	7596	7832	209	226
Gene loss x 2 and gene dupl. x 2	Run 3	601	7679	7800	290	273

Table S4. Results of statistical significance tests using Wilcoxon ranked sum test between Kendall-Colijn metrics of SPADES + SKA vs other workflows. The p-value threshold after Bonferroni correction for multiple testing is 3.85×10^{-3}. Statistically significant differences are marked with an asterisk.

Condition	Median of differences	P-value
SPAdes + SKA vs Velvet + PIRATE	12.53	1.192E-07*
SPAdes + SKA vs Velvet + kSNP	5.10	2.384E-07*
SPAdes + SKA vs Velvet + SKA	7.98	2.384E-07*
SPAdes + SKA vs Velvet + Roary	10.92	1.192E-07*
SPAdes + SKA vs SKESA + PIRATE	3.65	8.345E-07*
SPAdes + SKA vs SKESA + Roary	10.15	1.192E-07*
SPAdes + SKA vs SPAdes + Roary	11.30	1.192E-07*
SPAdes + SKA vs SPAdes + PIRATE	6.73	5.126E-06*
SPAdes + SKA vs Snippy (no assembly)	2.66	1.135E-03*
SPAdes + SKA vs SKESA + SKA	1.00	1.447E-02
SPAdes + SKA vs SKESA + kSNP	0.90	1.962E-02
SPAdes + SKA vs SKA (no assembly)	0.00	0.862
SPAdes + SKA vs SPAdes + kSNP	0.00	0.441

Table S5. Results of statistical significance tests using paired Wilcoxon ranked sum tests between Kendall-Colijn metrics of workflows employing Velvet vs SPAdes and SKESA. The p-value threshold after Bonferroni correction for multiple testing is 2.5×10^{-2}. Statistically significant differences are marked with an asterisk.

Condition	Median of differences	P-value
Velvet vs SKESA	-5.65	1.393E-27*
Velvet vs SPAdes	-5.42	9.684E-24*

Chapter 7

Software testing in microbial bioinformatics: a call to action

Boas C.L. van der Putten[$], C. Ines Mendes[$], Brooke M. Talbot, Jolinda de Korne-Elenbaas, Rafael Mamede, Pedro Vila-Cerqueira, Luis Pedro Coelho, Christopher A. Gulvik, Lee S. Katz, The ASM NGS 2020 hackathon participants*

[$]Equal contribution
*Full list of members available in Acknowledgments

Microbial Genomics, Volume 8, Issue 3, March 2022, https://doi.org/10.1099/mgen.0.000799

Abstract

Computational algorithms have become an essential component of research, with great efforts of the scientific community to raise standards on development and distribution of code. Despite these efforts, sustainability and reproducibility are major issues since continued validation through software testing is still not a widely adopted practice. Here, we report seven recommendations that help researchers implement software testing in microbial bioinformatics. We have developed these recommendations based on our experience from a collaborative hackathon organised prior to the American Society for Microbiology Next Generation Sequencing (ASM NGS) 2020 conference. We also present a repository hosting examples and guidelines for testing, available from https://github.com/microbinfie-hackathon2020/CSIS.

Impact Statement

In the field of microbial bioinformatics, good software engineering practises are not yet widely adopted. Many microbial bioinformaticians start out as (micro)biologists and subsequently learn how to code. Without abundant formal training, a lot of education about good software engineering practices comes down to an exchange of information within the microbial bioinformatics community. This paper serves as a resource that could help microbial bioinformaticians get started with software testing if they have not had formal training.

Background

Computational algorithms, software, and workflows have enhanced the breadth and depth of microbiological research and expanded the capacity of infectious disease surveillance in public health practice. Scientists now have a wealth of bioinformatic tools for addressing pertinent questions quickly and keeping pace with the availability of larger and more complex biological datasets. Despite these advances, we are finding ourselves in a crisis of computational reproducibility[1].

Modern software engineering advocates reliable software testing standards and best practices. Different approaches are employed: from unit testing to system testing[2], going from testing every individual component to testing a tool as a whole (Fig 1). The extent of testing is a balance between the resources available and increasing sustainability and reproducibility. Continuous Integration (CI), where code changes are frequently integrated and assertion of the new code's correctness before integration is often

automatedly performed through tests, provides a robust approach for ensuring the reproducibility of scientific results without requiring human interaction. Comprehensive testing of scientific software might prevent computational errors which subsequently lead to erroneous results and retractions[3,4]. However, the role of testing extends beyond that, as it also provides a way to measure software coverage, and therefore its robustness, allowing for reported issues to be converted into testable actions (regression tests), and the expansion and refactoring of existing code without compromising its function.

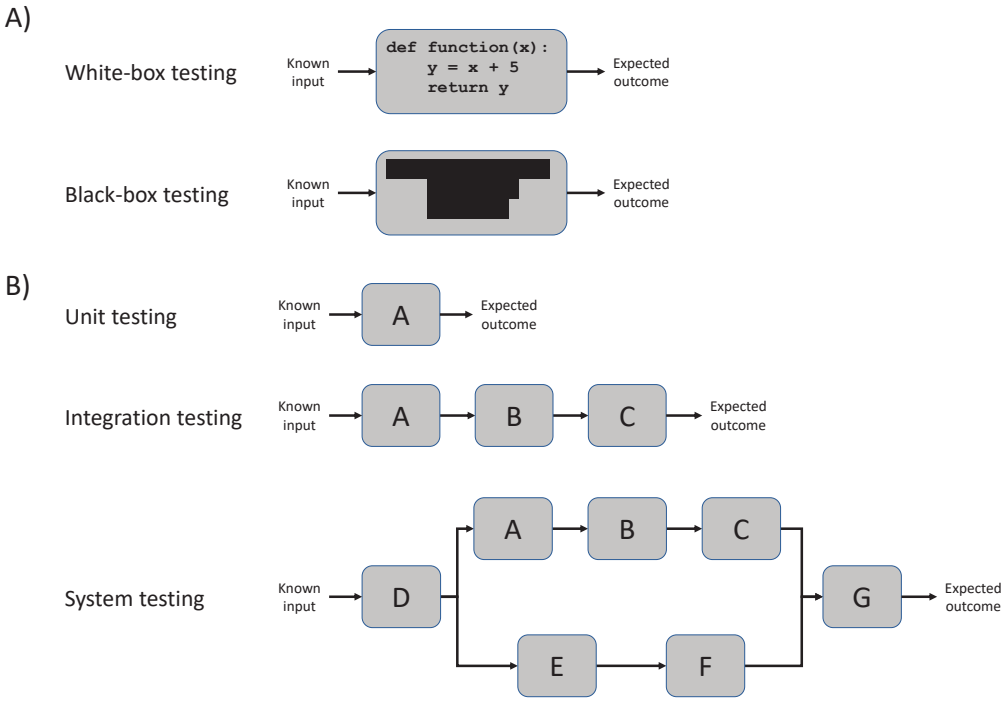


Figure 1. Testing strategies. A) White-box vs. black-box testing. In white-box testing, the tester knows the underlying code and structure of the software, where the tester does not know this in black-box testing. Note that this distinction is not strictly dichotomous and is considered less useful nowadays B) Unit vs. integration vs. system testing. When software comprises several modules, it is possible to test each single module (unit testing), groups of related modules (integration testing) or all modules (system testing). Note that the terms white-box testing and unit testing are sometimes used interchangeably but relate to different concepts.

Software testing among peers across fields aligns with previous efforts of hackathons to create a more unified and informed bioinformatics software community[5]. In this context, we hosted a cooperative hackathon prior to the ASM NGS conference in 2020, demonstrating that the microbial bioinformatics community can contribute to software sustainability using a collaborative platform. From this experience, we would like to propose collaborative software testing as an opportunity to continuously engage software users,

developers, and students to unify scientific work across domains. We have outlined the following recommendations for ensuring software sustainability through testing and offer a repository of automated test knowledge and examples at the Code Safety Inspection Service (CSIS) repository on GitHub (https://github.com/microbinfie-hackathon2020/CSIS).

Recommendations

Based on our experiences from the ASM NGS 2020 hackathon, we developed seven recommendations that can be followed during software development.

1. Establish software needs and testing goals

Manually testing the functionality of a tool is feasible in early development but can become laborious as the software matures. Developers may establish software needs and testing goals during the planning and designing stages to ensure an efficient testing structure. Table 1 provides an overview of testing methodologies and can serve as a guide to developers that aim to implement testing practises. A minimal test set could address the validation of core components or the program as a whole (system testing) and gradually progress toward verification of key functions which can accommodate code changes over time (unit testing, Fig 1). Ideally, testing should be implemented from the early stages of software development (test-driven development). Defining the scope of testing is important before developing tests. For pipeline development, testing of each individual component can be laborious and can be expedited if those components already implement testing of their own. Testing of the pipeline itself should take priority.

2. Input test files: the good, the bad, and the ugly

When testing, it is important to include test files with known expected outcomes for a successful run. However, it is equally important to include files or other inputs on which the tool is expected to fail. For example, some tools should recognize and report an empty input file or a wrong input format. Therefore, the test dataset should be small enough to be easily deployed (see recommendation #4) but as large as necessary to cover all intended test cases. Data provenance should be disclosed, either if it's from real data or originated *in silico*. Typically, a small test data is packaged with the software. Examples of valid and invalid file formats are available through the BioJulia project (https://github.com/BioJulia/BioFmtSpecimens). The nf-core project (https://nf-co.re/) provides a repository with test data for a myriad of cases (https://github.com/nf-core/test-datasets).

3. Use an established framework to implement testing

Understanding the test workflow can not only ensure continued software development but also the integrity of the project for developers and users. Testing frameworks improve test development and efficiency. Examples include unittest (https://docs.python.org/3/

Table 1 Overview of testing approaches. Software testing can be separated into three types: installation, functionality and destructive. Each component is described, followed by an example on a real-life application on *Software X*, a hypothetical nucleotide sequence annotation tool.

Name	Description	Example
Installation testing: can the software be invoked on different setups?		
Installation testing	Can the software be installed on different platforms?	*Test whether Software X can be installed using apt-get, pip, conda and from source.*
Configuration testing	With which dependencies can the software be used?	*Test whether Software X can be used with different versions of BLAST+.*
Implementation testing	Do different implementations work similarly enough?	*Test whether Software X works the same between the standalone and webserver versions.*
Compatibility testing	Are newer versions compatible with previous input/output?	*Test whether Software X can be used with older versions of the UniProtKB database.*
Static testing	Is the source code syntactically correct?	*Check whether all opening braces have corresponding closing braces or whether code is indented correctly in Software X.*
Standard functionality testing: does the software do what it should in daily use?		
Use case testing	Can the software do what it is supposed to do regularly?	*Test whether Software X can annotate different FASTA files: with spaces in the header, without a header, an empty file, with spaces in the sequence, with unknown characters in the sequences, et cetera.*
Workflow testing	Can the software successfully traverse each path in the analysis?	*Test whether Software X works in different modes (using fast mode or using one dependency over the other).*
Sanity testing	Can the software be invoked without errors?	*Test whether Software X works correctly without flags, or when checking dependencies or displaying help info.*
Destructive testing: what makes the software fail?		
Mutation testing	How do the current tests handle harmful alterations to the software?	*Test whether changing a single addition to a subtraction within Software X causes the test suite to fail.*
Load testing	At what input size does the software fail?	*Test whether Software X can annotate a small plasmid (10 Kbp), a medium-size genome (2 Mbp) or an unrealistically large genome for a prokaryote (1 Gbp).*
Fault injection	Does the software fail if faults are introduced and how is this handled?	*Test whether Software X fails if nonsense functions are introduced in the gene calling code.*

library/unittest.html) or pytest (https://docs.pytest.org/en/stable/) for Python, and testthat (https://testthat.r-lib.org/) for R, testing interfaces such as TAP (http://testanything.org/), or built-in test attributes such as in Rust. Although many tests can be implemented using a combination of frameworks, personal preferences (e.g. amount of boilerplate code required) might drive your choice. Additionally, in Github Actions the formulas of each test block can be explicitly stated using the standardised and easy-to-follow YAML (https://

yaml.org/, Supplementary Figure S1), already adopted by most continuous integration platforms (Recommendation 4). For containerised software, testing considerations differ slightly and have been covered previously by Gruening *et al*. (2019)[6].

4. Testing is good, automated testing is better

When designing tests, planning for automation saves development time. Whether your tests are small or comprehensive, automatic triggering of tests will help reduce your workload. Many platforms trigger tests automatically based on a set of user-defined conditions. Platforms such as GitHub Actions (https://github.com/features/actions) and GitLab CI (https://about.gitlab.com/stages-devops-lifecycle/continuous-integration) offer straightforward automated testing of code seamlessly upon deployment. A typical workflow, consisting of a minimal testing framework (see recommendation #1 and #3) and a small test dataset (see recommendation #2), can then be directly integrated within your project hosted on a version control system, such as GitHub (https://github.com/), and directly integrated with a continuous integration provider, such as GitHub Actions in GitHub. Testing considerations for containerised software has been covered previously by Gruening *et al*. (2019)[6].

5. Ensure portability by testing on several platforms

The result of an automated test in the context of one computational workspace does not ensure the same result will be obtained in a different setup. It is important to ensure your software can be installed and used across supported platforms. One way to ensure this is to test on different environments, with varying dependency versions (*e.g.*, multiple Python versions, instead of only the most recent one). Developers can gain increased benefits of testing if tests are run on different setups automatically (see recommendation #4 and Supplementary Figure S1).

6. Showcase the tests

For prospective users, it is good to know whether you have tested your software and, if so, which tests you have included. This can be done by displaying a badge[7] (see https://github.com/microbinfie-hackathon2020/CSIS/blob/main/README.md#example-software-testing), or linking to your defined testing strategy e.g. a GitHub Actions YAML, (see recommendation #2, Supplementary Figure S1). Documenting the testing goal and process enables end-users to easily check tool functionality and the level of testing[8].

It may be helpful to contact the authors, directly or through issues in the code repository, whose software you have tested to share successful outcomes or if you encountered abnormal behaviour or component failures. An external perspective can be useful to find bugs that the authors are unaware of. A set of issue templates for various situations is available in the CSIS repository on GitHub (https://github.com/microbinfie-hackathon2020/CSIS/tree/main/templates).

7. Encourage others to test your software

Software testing can be crowdsourced, as showcased by the ASM NGS 2020 hackathon. Software suites such as Pangolin (https://github.com/cov-lineages/pangolin)[9] and chewBBACA (https://github.com/B-UMMI/chewBBACA)[10] have implemented automated testing developed during the hackathon. For developers, crowdsourcing offers the benefits of fresh eyes on your software. Feedback and contributions from users can expedite the implementation of software testing practices. It also contributes to software sustainability by creating community buy-in, which ultimately helps the software maintainers keep pace with dependency changes and identify current user needs.

Conclusions

Testing is a critical aspect of scientific software development, but automated software testing remains underused in scientific software. In this hackathon, we demonstrated the usefulness of testing and developed a set of recommendations that should improve the development of tests. We also demonstrated the feasibility of producing test suites for already-established microbial bioinformatics software.

Acknowledgements

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC). The mention of company names or products does not constitute an endorsement by the CDC.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Hackathon participants

In addition to the authors, the following participants were responsible for automating tests for bioinformatic tools and contributing a community resource for identifying software that can pass unit tests, available at https://github.com/microbinfie-hackathon2020/CSIS. Participants are listed alphabetically: Áine O'Toole, Amit Yadav, Justin Payne, Mario Ramirez, Peter van Heusden, Robert A. Petit III, Verity Hill, Yvette Unoarumhi.

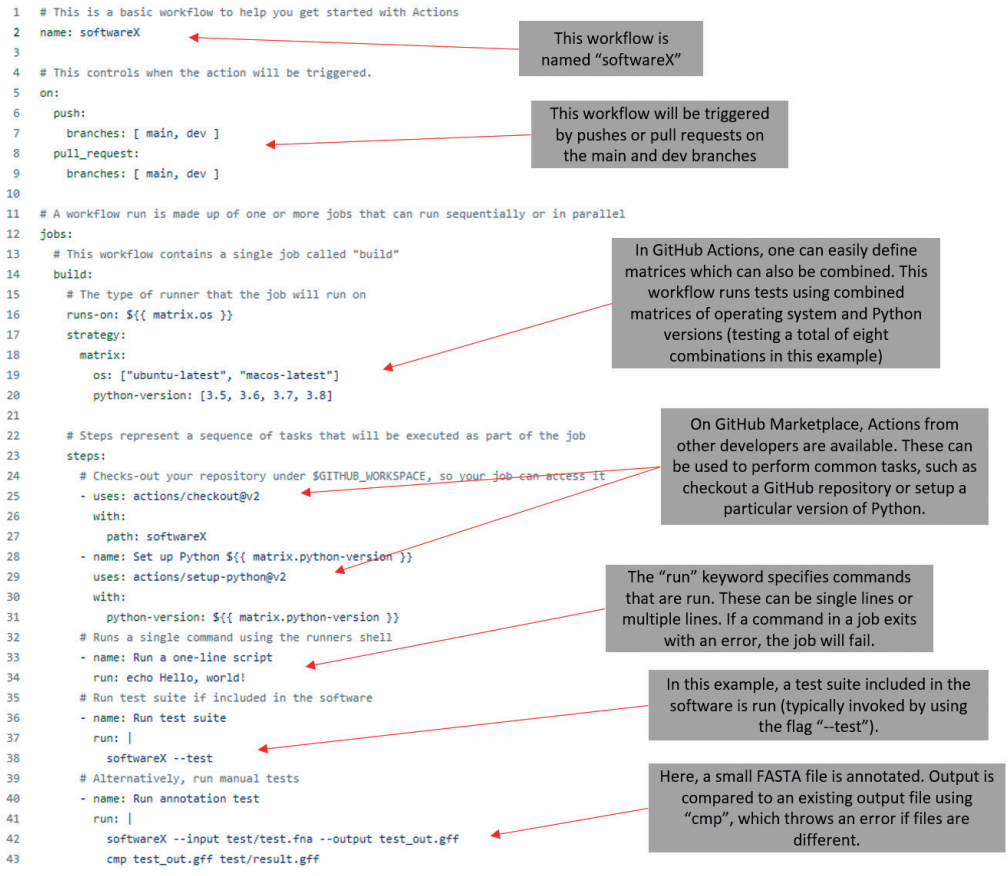
Funding

C.I.M. was supported by the Fundação para a Ciência e Tecnologia (grant SFRH/BD/129483/2017). L.P.C. was partially supported by Shanghai Municipal Science and Technology Major Project (2018SHZDZX01) and ZJLab. R. M. was supported by the Fundação para a Ciência e Tecnologia (grant 2020.08493.BD).

References

1. Stodden V, Seiler J, Ma Z. 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc Natl Acad Sci* USA 115:2584–2589.
2. Krafczyk M, Shi A, Bhaskar A, Marinov D, Stodden V. 2019. Scientific Tests and Continuous Integration Strategies to Enhance Reproducibility in the Scientific Software Context, p. 23–28. *In Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*. Association for Computing Machinery, Phoenix, AZ, USA.
3. Chang G, Roth CB, Reyes CL, Pornillos O, Chen Y-J, Chen AP. 2006. Retraction. *Science* 314:1875.2-1875.
4. Hall BG, Salipante SJ. 2007. Retraction: Measures of Clade Confidence Do Not Correlate with Accuracy of Phylogenetic Trees. *PLoS Comput Biol* 3:e158.
5. Busby B, Lesko M, Federer L. 2016. Closing gaps between open software and public data in a hackathon setting: User-centered software prototyping. *F1000Res* 5:672.
6. Gruening B, Sallou O, Moreno P, da Veiga Leprevost F, Ménager H, Søndergaard D, Röst H, Sachsenberg T, O'Connor B, Madeira F, Dominguez Del Angel V, Crusoe MR, Varma S, Blankenberg D, Jimenez RC, BioContainers Community, Perez-Riverol Y. 2019. Recommendations for the packaging and containerizing of bioinformatics software. *F1000Res* 7:742.
7. Trockman A, Zhou S, Kästner C, Vasilescu B. 2018. Adding sparkle to social coding: an empirical study of repository badges in the *npm* ecosystem, p. 511–522. *In Proceedings of the 40th International Conference on Software Engineering*. Association for Computing Machinery, Gothenburg, Sweden.
8. Karimzadeh M, Hoffman MM. 2018. Top considerations for creating bioinformatics software documentation. *Briefings in Bioinformatics* 19:693–699.
9. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, Yeats C, Du Plessis L, Maloney D, Medd N, Attwood SW, Aanensen DM, Holmes MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom* 4.

Supplementary material




```yaml
# This is a basic workflow to help you get started with Actions
name: softwareX

# This controls when the action will be triggered.
on:
  push:
    branches: [ main, dev ]
  pull_request:
    branches: [ main, dev ]

# A workflow run is made up of one or more jobs that can run sequentially or in parallel
jobs:
  # This workflow contains a single job called "build"
  build:
    # The type of runner that the job will run on
    runs-on: ${{ matrix.os }}
    strategy:
      matrix:
        os: ["ubuntu-latest", "macos-latest"]
        python-version: [3.5, 3.6, 3.7, 3.8]

    # Steps represent a sequence of tasks that will be executed as part of the job
    steps:
      # Checks-out your repository under $GITHUB_WORKSPACE, so your job can access it
      - uses: actions/checkout@v2
        with:
          path: softwareX
      - name: Set up Python ${{ matrix.python-version }}
        uses: actions/setup-python@v2
        with:
          python-version: ${{ matrix.python-version }}
      # Runs a single command using the runners shell
      - name: Run a one-line script
        run: echo Hello, world!
      # Run test suite if included in the software
      - name: Run test suite
        run: |
          softwareX --test
      # Alternatively, run manual tests
      - name: Run annotation test
        run: |
          softwareX --input test/test.fna --output test_out.gff
          cmp test_out.gff test/result.gff
```

Figure S1. Example YAML file for a GitHub Actions workflow.

Table S1. Software tested during the ASM NGS 2020 hackathon.

Software Name	Software (URL)	Test File (URL)	Literature Citation (DOI)
BUSCO	https://gitlab.com/ezlab/busco	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/busco.yml	10.1093/bioinformatics/btv351
Centrifuge	https://github.com/DaehwanKimLab/centrifuge	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/centrifuge.yml	10.1101/gr.210641.116
CheckM	https://github.com/Ecogenomics/CheckM	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/checkm.yml	10.1101/gr.186072.114
chewBBACA	https://github.com/B-UMMI/chewBBACA	https://github.com/B-UMMI/chewBBACA/blob/master/.github/workflows/chewbbaca.yml	10.1099/mgen.0.000166
CSIS	https://github.com/microbinfie-hackathon2020/CSIS	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/CSIS.yml	this manuscript
Genotyphi	https://github.com/katholt/genotyphi	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/genotyphi.yml	10.1038/ncomms12827
Kraken	https://github.com/DerrickWood/kraken	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/kraken.yml	10.1186/gb-2014-15-3-r46
Kraken2	https://github.com/DerrickWood/kraken2	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/kraken2.yml	10.1186/s13059-019-1891-0
KrakenUniq	https://github.com/fbreitwieser/krakenuniq	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/krakenuniq.yml	10.1186/s13059-018-1568-0
Pangolin	https://github.com/cov-lineages/pangolin	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/pangolin.yml	10.1093/ve/veab064
Prokka	https://github.com/tseemann/prokka	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/prokka.yml	10.1093/bioinformatics/btu153
Quast	https://github.com/ablab/quast	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/quast.yml	10.1093/bioinformatics/btt086
Shovill	https://github.com/tseemann/shovill	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/shovill.yml	absent
SKESA	https://github.com/ncbi/SKESA	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/skesa.yml	10.1186/s13059-018-1540-z
Trycycler	https://github.com/rrwick/Trycycler	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/trycycler.yml	10.5281/zenodo.4430941
Unicycler	https://github.com/rrwick/Unicycler	https://github.com/microbinfie-hackathon2020/CSIS/blob/main/.github/workflows/unicycler.yml	10.1371/journal.pcbi.1005595

Chapter 8

Five complete genome sequences spanning the Dutch *Streptococcus suis* serotype 2 and serotype 9 population

Boas C.L. van der Putten[$], Thomas J. Roodsant[$], Martin A. Haagmans, Constance Schultsz, Kees C. H. van der Ark

[$]Equal contribution

Microbiology Resource Announcements, Volume 9, Issue 6, February 2020, e01439-19, https://doi.org/10.1128/MRA.01439-19

Abstract

The zoonotic pathogen *Streptococcus suis* can cause septicemia and meningitis in humans. We report five complete genomes of *Streptococcus suis* serotype 2 and serotype 9, covering the complete phylogeny of serotype 9 Dutch porcine isolates and zoonotic isolates. The isolates include the model strain S10 and Dutch emerging zoonotic lineage.

Main

Streptococcus suis is an opportunistic pathogen in pigs, which can cause zoonotic infections. Human infections are predominantly caused by *S. suis* serotype 2[1] and can lead to septicemia and meningitis[2]. We recently identified a zoonotic *S. suis* serotype 2 clone belonging to Clonal Complex (CC) 20 which emerged from a non-zoonotic serotype 9 CC16 clone[3], in the Netherlands. To facilitate further research on the zoonotic potential of *S. suis*, we sequenced the genomes of *S. suis* serotype 9 CC16 and CC20 strains, isolated from diseased pigs, and three serotype 2 strains, including strain S10 (CC1, pig) and two CC20 strains, one each from human and porcine infection (Table 1). Data were generated using Illumina and Nanopore MinION sequencing technologies.

S. suis was grown overnight in THY and genomic DNA was isolated using the Qiagen MagAttract HMW DNA extraction kit. The sequence library was constructed using the native barcoding (EXP-NBD114) and ligation sequencing (SQK-LSK109) kit (Oxford Nanopore). DNA was repaired and A-tailed using NEBNext FFPE DNA Repair Mix and the NEBNext Ultra II End Repair/dA-Tailing Module (New England BioLabs). A barcode was ligated to the A-tailed DNA using Blunt/TA Ligase Master Mix (New England Biolabs). Sequence adapters were ligated to barcoded samples pooled by equal mass with Quick T4 DNA Ligase (New England BioLabs). The library was loaded on the flow cell (FLO-MIN 106D R9) and sequenced using MinKNOW fast basecalling version 3.5.5. Default parameters were used for all tools except where noted otherwise. Illumina data were available from our previous study (Table 1)[3].

Illumina read filtering was performed using fastp version 0.20.0[4]. MinION reads were filtered on quality and length using filtlong version 0.2.0[5], using the filtered Illumina reads as reference. FastQC was used for quality control version 0.11.8[6]. Illumina and MinION reads were used in hybrid assembly using Unicycler version 0.4.8, which also performs assembly trimming, circularizing and rotating[7]. Assembly statistics were collected using Quast version 4.6.3[8]. Coverage was assessed using Minimap2 version 2.17[9], Samtools version 1.9[10] and bedtools version 2.29.0[11]. The complete genomes were annotated using prokka version 1.14.0[12]. MLST was performed using mlst version 2.17.6[13]. For workflow

management, Snakemake version 5.7.1[14] was used. The pipeline is freely available from https://github.com/boasvdp/MRA_Streptococcus_suis.

Genomes of all five strains consisted of a single chromosome ranging from 2,042,889 to 2,292,626 bp with a GC content between 41,10% and 41,43 and a coverage between 23-72x using Nanopore data (Table 1).

Draft assemblies of the five strains were between 46-74 kbp smaller than the complete genomes. Mapping the draft genomes to the complete genomes revealed no missing regions in the draft genomes. The draft genomes are likely smaller than the complete genomes due to the collapse of repeats, which has been described before[15].

Accession numbers

Nanopore .fastq and .fast5 data, as well as assembled genomes have been deposited in ENA under accession numbers listed in Table 1 and study number PRJEB35407.

Acknowledgements

This study was funded through EU-Horizon2020 grant 727966 (PIGSs) and an Amsterdam UMC PhD grant. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Table 1. Isolate details, genome information and accession numbers.

Isolate	Isolation source	Serotype	Clonal Complex	Genome length (bp)	GC content (%)	Total CDS	Nanopore read N50	Nanopore reads	Coverage Nanopore	Nanopore run accessions	Illumina run accessions	Assembly accessions
861160	Human CSF	2	20	2,148,824	41.10%	2029	13,589	7557	23	ERR3664732	ERR1055554	GCA_902702745
GD-0001	Diseased pig	2	20	2,125,468	41.24%	2014	25,831	10,490	54	ERR3664733	ERR1055586	GCA_902702785
9401240	Diseased pig	9	20	2,195,215	41.43%	2036	11,192	30,418	60	ERR3664735	ERR1055578	GCA_902702775
GD-0088	Diseased pig	9	16	2,298,012	41.20%	2213	7657	15,321	27	ERR3664734	ERR1055627	GCA_902702765
S10	Diseased pig	2	1	2,048,275	41.32%	1952	15,208	20,251	72	ERR3664731	ERR1055646	GCA_902702755

References

1. Huong VTL, Ha N, Huy NT, Horby P, Nghia HDT, Thiem VD, Zhu X, Hoa NT, Hien TT, Zamora J, Schultsz C, Wertheim HFL, Hirayama K. 2014. Epidemiology, clinical manifestations, and outcomes of *Streptococcus suis* infection in humans. *Emerging infectious diseases* 20:1105-1114.
2. Wertheim HFL, Nghia HDT, Taylor W, Schultsz C. 2009. *Streptococcus suis*: An Emerging Human Pathogen. *Clinical Infectious Diseases* 48:617-625.
3. Willemse N, Howell KJ, Weinert LA, Heuvelink A, Pannekoek Y, Wagenaar JA, Smith HE, van der Ende A, Schultsz C. 2016. An emerging zoonotic clone in the Netherlands provides clues to virulence and zoonotic potential of *Streptococcus suis*. *Scientific Reports* 6:28984.
4. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884-i890.
5. Wick RR. 2019. Filtlong, https://github.com/rrwick/Filtlong.
6. Andrew S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
7. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* 13:e1005595.
8. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072-5.
9. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094-3100.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England) 25:2078-2079.
11. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
12. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-9.
13. Seemann T. mlst, Github, https://github.com/tseemann/mlst.
14. Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 28:2520-2522.
15. Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews Genetics* 13:36-46.

Chapter 9

Identification of *Streptococcus suis* putative zoonotic virulence factors: A systematic review and genomic meta-analysis

Thomas J. Roodsant, Boas C.L. van der Putten, Sara M. Tamminga, Constance Schultsz, Kees C.H. van der Ark

Virulence, Volume 12, Issue 1, November 2021, Pages 2787-2797, https://doi.org/10.1080/21505594.2021.1985760

Abstract

Streptococcus suis is an emerging zoonotic pathogen. Over 100 putative virulence factors have been described, but it is unclear to what extent these virulence factors could contribute to zoonotic potential of *S. suis*. We identified all *S. suis* virulence factors studied in experimental models of human origin in a systematic review and assessed their contribution to zoonotic potential in a subsequent genomic meta-analysis. PubMed and Scopus were searched for English-language articles that studied *S. suis* virulence published until 31 March 2021. Articles that analyzed a virulence factor by knockout mutation, purified protein, and/or recombinant protein in a model of human origin, were included. Data on virulence factor, strain characteristics, used human models and experimental outcomes were extracted. All publicly available *S. suis* genomes with available metadata on host, disease status and country of origin, were included in a genomic meta-analysis. We calculated the ratio of the prevalence of each virulence factor in human and pig isolates. We included 130 articles and 1703 *S. suis* genomes in the analysis. We identified 53 putative virulence factors that were encoded by genes which are part of the *S. suis* core genome and 26 factors that were at least twice as prevalent in human isolates as in pig isolates. Hhly3 and NisK/R were particularly enriched in human isolates, after stratification by genetic lineage and country of isolation. This systematic review and genomic meta-analysis have identified virulence factors that are likely to contribute to the zoonotic potential of *S. suis*.

Introduction

Streptococcus suis is an opportunistic pathogen in pigs and can cause zoonotic infections that often result in meningitis[1,2]. *S. suis* zoonotic infections occur worldwide with the highest reported incidence in Thailand, Vietnam and the Netherlands[1]. Close contact with pigs and consumption of undercooked pork have been identified as important risk factors for zoonotic *S. suis* infections[1]. The emergence of zoonotic clones has been demonstrated and led to new insights in the evolution of *S. suis*' population structure[3], but the virulence factors involved in zoonotic potential of *S. suis* are not well understood.

S. suis of multiple serotypes from different phylogenetic groups (clonal complexes) are found in healthy and diseased pigs, but human infections are predominantly caused by strains from clonal complex 1 and serotypes 2 or 14[1,4]. Distinct stages in the pathogenesis of *S. suis* infections in humans include the adhesion and translocation across mucosal surface particularly in case of foodborne infection, survival in blood, and translocation across the blood brain barrier in case of meningitis[5]. Over 100 putative *S. suis* virulence factors have been described that may contribute to the pathogenesis of infection in pigs[4,6]. Although many of these virulence factors were identified in *in vitro* models of human origin, their contribution to *S. suis* zoonotic potential has not been studied.

We performed a systematic review of *S. suis* virulence factors studied in *in vitro* models of human origin. In a subsequent genomic meta-analysis we determined if these putative virulence factors are encoded by the *S. suis* core or accessory genome and identified those virulence factors that may contribute to the pathogenesis of zoonotic infection, designated putative zoonotic virulence factors (PZVFs).

Methods

Definitions

Virulence factors can be defined as "molecules produced by pathogens that contribute to the pathogenicity of the organism by allowing its establishment, replication, dissemination and persistence in the host"[4]. Here, we define a PZVF as a virulence factor of a bacterial pathogen from an animal reservoir that contributes to pathogenicity in the human host specifically. We define human models as *in vitro* models of human origin, including cell lines in continuous culture of human origin, human primary cells, human blood, human blood components, human extracellular matrix proteins, and the zebrafish human streptococcal infection model[7].

Search strategy and selection criteria

The systematic review was performed according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[8]. TR searched PubMed and Scopus for primary research articles published until 31 March 2021 describing *S. suis* and virulence in the title and/or abstract using Pubmed PubReMiner to generate the search query (appendix S1 p1)[9]. References were downloaded and duplicates were removed using Endnote (9.3.3), Mendeley (1.19.8) and a manual search. TR and KA independently screened all titles and abstracts and selected articles that mentioned a host (e.g. human or pig) and *S. suis* and both agreed on the final selection for full text screening, which was done by TR. Studies were included when a virulence factor was evaluated in a human model and the virulence factor was studied in an isogenic knockout (KO) mutant, as recombinant protein, and/or as purified protein. Articles were excluded when the full text was unavailable in English. Experimental outcomes included bacterial binding of host proteins, adhesion, invasion, translocation, survival and immune cell responses.

Data extraction

TR and ST extracted data from the included articles in a pre-specified table in Microsoft Excel 2016 (appendix S1 p2, appendix S2), followed by an overall curation of extracted data by KA. In short, we extracted information on virulence factor analysis approach (KO, recombinant protein and/or purified protein), *S. suis* strain characteristics, applied *in vitro* models, experimental outcomes and NCBI protein ID. If NCBI protein ID was not stated, the NCBI protein ID was searched manually using available data such as primer sequences, gene names or protein sequences. Experimental outcomes for single virulence factors studied in at least 5 articles were summarized and compared. As part of a critical appraisal, data on growth rates of wildtype, isogenic KO and complementation mutants were extracted from the articles or articles' references. In addition, the number of *S. suis* strains analyzed in each study was recorded.

Bacterial genome meta-analysis

We downloaded all BioSample records from NCBI mentioning "*Streptococcus suis*" (final date 31–01-2020). Missing metadata were searched in the corresponding publications and pubMLST[10], and added. Genomes were included if at least metadata on host, host health status, and country of origin were available (see appendix 1 p3). The curated set of assembled genomes with corresponding metadata was deposited on Zenodo (10.5281/zenodo.4686597).

The presence of a virulence factor in *S. suis* isolates was determined by mapping its protein sequence with a minimal protein identity of 95% and query coverage of 60% on the translated *S. suis* genome assemblies (see appendix 1 p3[11–14]). We defined the core genome as all genes present in ≥95% of the isolates whilst the remaining genes constitute the accessory genome.

We calculated the ratio of the prevalence of each virulence factor in *S. suis* populations isolated from human, and healthy and diseased pigs respectively. A virulence factor was considered a PZVF if the prevalence ratio > 2. A stratified analysis was performed for the main zoonotic *S. suis* lineage (clonal complex 1) and the countries contributing most human isolates (China, Vietnam).

Results

Title and abstract of 713 unique records were screened and 411 articles were selected for full text screening. Of these 411, 268 articles did not meet inclusion criteria and 13 were excluded due to unavailability of full text in English. The 130 included articles described 124 different putative virulence factors (Figure 1).

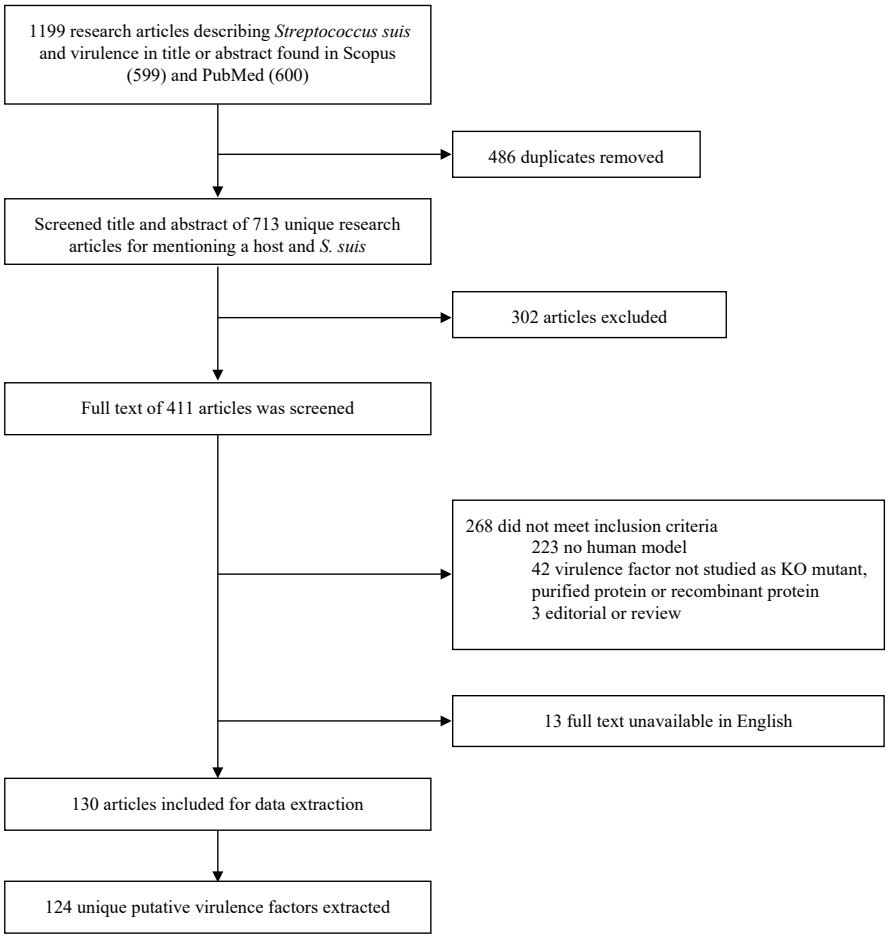


Figure 1. PRISMA flow diagram.

Putative virulence factors were studied as purified protein (3), as recombinant protein (51), as (partial) isogenic KO (152), by blocking protein function with antibodies (3) or as a combination of these. For 56/152 (37%) of the isogenic KO mutants, changes in growth rate compared to parental wildtype were not assessed. For 72/152 (47%) growth rate of KO mutants was reported as unaffected and for 24/152 (16%) impaired growth was observed for the KO mutant. In only 43 (28%) studies the KO mutant was genetically complemented and three articles (2%) described complementation with a recombinant protein.

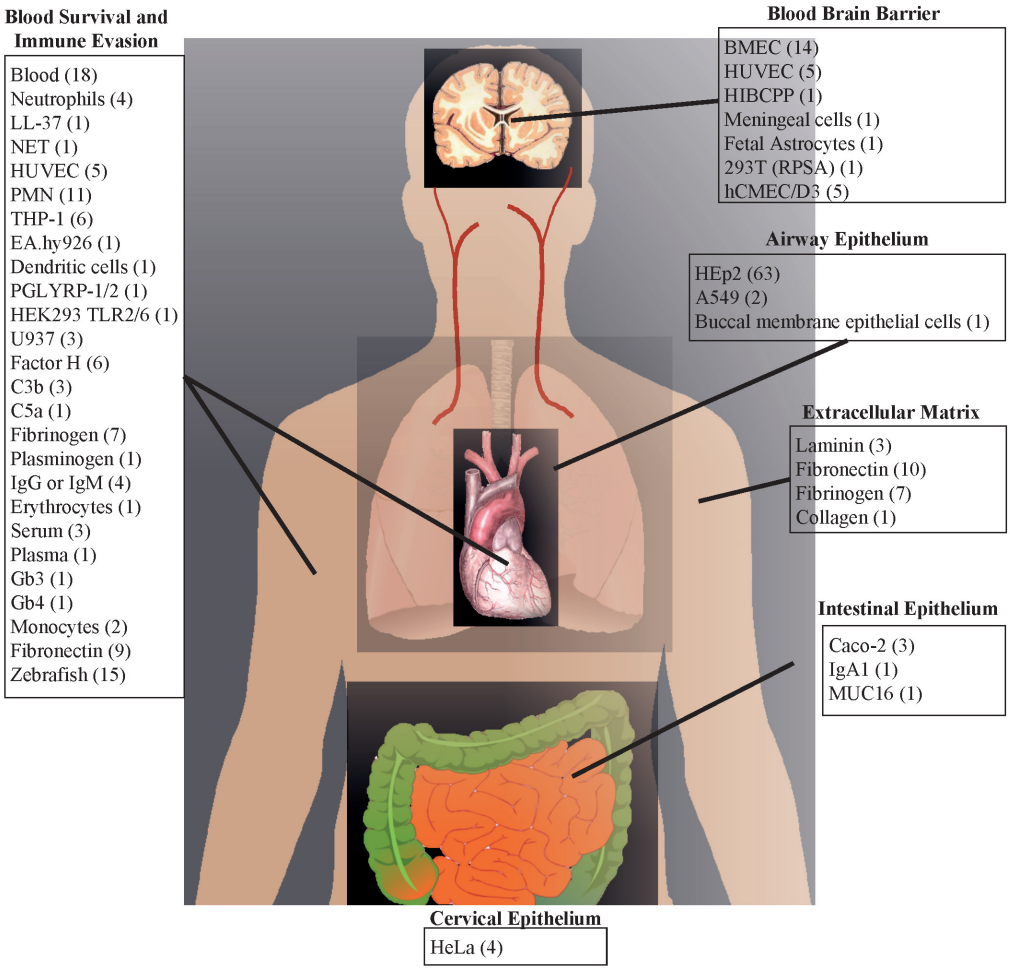


Figure 2. Grouping of human models to their respective human *S. suis* infection site. Number of articles per model is indicated between brackets.

Models used to evaluate putative virulence factors were grouped based on the human body sites from which the model originated (Figure 2). The human epithelial HEp2 cell line was used in 63 out of 72 articles that studied adhesion, invasion or cell lysis induced by *S. suis* in a human epithelial model. Adhesion to extracellular matrix was studied in 13 different articles that used laminin (3), collagen (1), fibronectin (10) and/or fibrinogen (7). Survival in blood was studied in 70 articles using a diverse set of models, of which human whole blood (19), human (polymorphonuclear) neutrophils (15) and zebrafish (15) were most frequent. Human brain microvascular endothelial cells (BMECs) were used in 14 out of 25 articles that studied the role of a virulence factor in crossing the blood-brain barrier (BBB).

Experimental outcomes

Five out of 124 (4%) putative virulence factors (appendix S1 p2, appendix S2) were studied in at least 5 articles and the experimental outcomes were summarized and compared for each factor to evaluate their contribution to zoonotic potential (appendix S1 p4-6).

Capsular polysaccharide (CPS)

The CPS forms the outer layer of bacteria and consists of repeating oligosaccharide subunits that differ in composition and linkage between *S. suis* serotypes[15,16]. The CPS decreases *S. suis* adherence to and invasion of human epithelial cells[17–21]. One study reported no effect of the CPS on adherence to cervical epithelial HeLa cells[22]. The CPS contributes to *S. suis* blood survival[21,23] and to immune evasion by decreasing phagocytosis[21,22,24–29], increasing intracellular survival in PMN[23], dampening the innate immune response and decreasing complement activation[25,26,28–30]. The CPS decreases *S. suis* adherence to and invasion of meningeal cells and fetal astrocytes[31] and the adherence to human umbilical vein endothelial cells (HUVEC)[21]. A CPS KO showed a trend of increased translocation across human choroid plexus papilloma cells (HIBCCP)[32]. The CPS decreases IL-6 and IL-8 secretion by BMECs, but MCP-1 secretion is unaffected[33]. Moreover, purified CPS induces PGE2 and MMP-9 secretion in macrophage-like U937 cells which increases BBB leakiness[25]. The CPS also binds fibrinogen[24].

Suilysin (Sly)

Suilysin is a cholesterol-dependent hemolysin secreted by *S. suis* that can form pores in eukaryotic cells by oligomerization in cellular membranes[34,35]. Sly can induce HEp2 cell lysis[36] and at subcytolitic concentrations contributes to HEp2 invasion[37]. Sly does not contribute to HEp2 cell adherence[37] or translocation across an intestinal epithelial Caco-2 monolayer[20]. In human blood, Sly induces TNFα release by monocytes[38], PMN degranulation[39], platelet-neutrophil complex formation[40] and its hemolytic activity causes inflammasome activation in macrophage-like THP-1 cells[28]. Sly induces the release of arachidonic acid in BMEC, which can enhance BBB permeability[41]. Sly does not affect

adherence to meningeal cells, adherence to astrocytes or invasion of meningeal cells but does increase astrocytes invasion[31].

Muramidase release protein (MRP)

MRP is a cell-wall anchored protein similar to the fibronectin-binding protein of *Staphylococcus aureus*[42–45]. MRP can directly bind to HEp2 cells[46] and contributes to *S. suis* adhesion to HEp2[47]. MRP was shown to bind fibrinogen[48–50] and MRP binding of fibrinogen contributes to blood survival, decreases PMN killing[48,49] and in brain microvascular endothelial hCMEC/D3 cells increases adhesion and translocation[51]. MRP with bound fibrinogen also decreases the adherens junction protein p120-catenin in hCMEC/D3 thus potentially increasing BBB permeability[51]. Besides fibrinogen, MRP was shown to bind factor H[50,52] and fibronectin[50].

Factor H binding protein (Fhb)

Fhb also named Streptococcal adhesin protein (SadP) is anchored in the cell wall and secreted[53]. Fhb can bind proteins from the host complement system as well as glycans[53–55]. Fhb can bind to Gb3 on human erythrocytes and a specific allele of Fhb (SadPn) can also bind to Gb4[56]. Fhb contributes to *S. suis* adhesion to and translocation across a Caco-2 monolayer by binding to Gb3[57]. Fhb can bind human factor H, which increases *S. suis* adherence to airway epithelial A549 cells[29]. A Fhb KO showed decreased binding to vascular endothelial EA.hy926 cells[56]. Fhb contributes to *S. suis* survival in whole blood[53] and intracellular survival in PMN[53,55]. Fhb can bind factor H[52,53,55] and C3 simultaneously[53] and a Fhb KO showed decreased factor H binding and increased C3b/iC3b deposition[53,55]. Secreted Fhb lowers C3b/C3b deposition on a Fhb KO mutant and restores PMN intracellular survival of *S. suis*[53]. However, a Fhb KO in a different strain still bound factor H and degraded C3b, whilst THP-1 phagocytosis of the KO mutant was unaffected[29]. Translocation across and adhesion to hCMEC/D3 cells is decreased in a Fhb KO mutant[58] and factor H binding by Fhb increases adherence to BMEC cells[29]. Fhb was shown to bind fibrinogen[48].

Enolase

Enolase is a multifunctional protein with glycolytic functions and plasminogen binding abilities, and is found in many organisms[59]. In *S. suis*, enolase was found within the cytoplasm and on the cell surface of *S. suis*, although lacking a LPXTG-motif[60]. Blocking enolase functioning with recombinant protein or polyclonal antibodies was shown to decrease adherence to HEp2 cells[61–63]. Enolase was shown to bind fibronectin[61], laminin[61] and factor H[52]. 40S ribosomal protein SA (RPSA), a protein involved in BBB integrity, was shown to increase at the cell surface of hCMEC/D3 cells when treated with enolase[64]. In transfected HEK-293 T cells it was demonstrated that enolase can interact with RPSA[64,65]. Enolase can induce apoptosis in HEK-293 T cells[65] and in hCMEC/D3 cells by interacting

with RSPA[64,65]. The apoptosis induced by enolase is inhibited by caveolae, a type of lipid raft[64].

Identification of putative zoonotic virulence factors

Out of 3307 *S. suis* BioSample records in the NCBI database, 315 human, 896 pig diseased and 492 pig healthy *S. suis* genome assemblies were included (appendix S1 p7, appendix S3), including human isolates from Vietnam (45%), China (31%), the Netherlands (9%), Thailand (5%) and Togo (5%). For 1012 (31%) BioSample records, the information on host was unavailable (appendix S1 p7).

Out of 111 unique protein sequences, including multiple alleles for four proteins, 53 proteins were encoded by genes which are part of the *S. suis* core genome (appendix S1 p8). The remaining 58 proteins were encoded by genes which are part of the accessory genome. The presence of these 58 accessory proteins together with isolate metadata was plotted against a clustered core genome alignment[11–14] (Figure 3a) and the human-pig prevalence ratio was calculated (Figure 3b, appendix 1 p9). Six proteins (Atl1, Atl2, AtlAss, CPS9E, KAR and PK) had a prevalence ratio below 1. For 26 proteins, including MRP, Sly and CPS2B/E/F/G/J/L which form a single operon[15], the prevalence ratio was above 2 and three of these proteins, Fhb_1, NisK and NisR were at least ten times more prevalent in human isolates than in pig isolates.

Ninety percent of human *S. suis* isolates had the same genetic background (CC1) while the pig isolates were genetically more diverse (Figure 3a). To adjust for potential lineage effects, we repeated our analysis for the 52 proteins with prevalence ratio above 1 in the first analysis, but restricted to CC1 isolates. Of these 52 proteins, 35 were encoded by genes which are part of the CC1 core genome, including Sly and CPS2B/L. Four proteins (nisin dependent two-component signal transduction system [NisK/R], putative hemolysin-III-related protein [Hhly3] and Fhb_1) had a prevalence ratio of at least 2 (appendix S1 p10-11). NisK/R and Hhly3 were initially discovered on the 89 K pathogenicity island found in Chinese human *S. suis* outbreak isolates belonging to ST7[66,67]. Outside ST7 but within CC1, both PZVFs were also present in 110 Vietnamese human isolates from ST 1 (105), ST144 (3), ST869 (1) and ST951 (1), and in 4 Chinese human isolates from ST1 (2), ST665 (1) and ST658 (1).

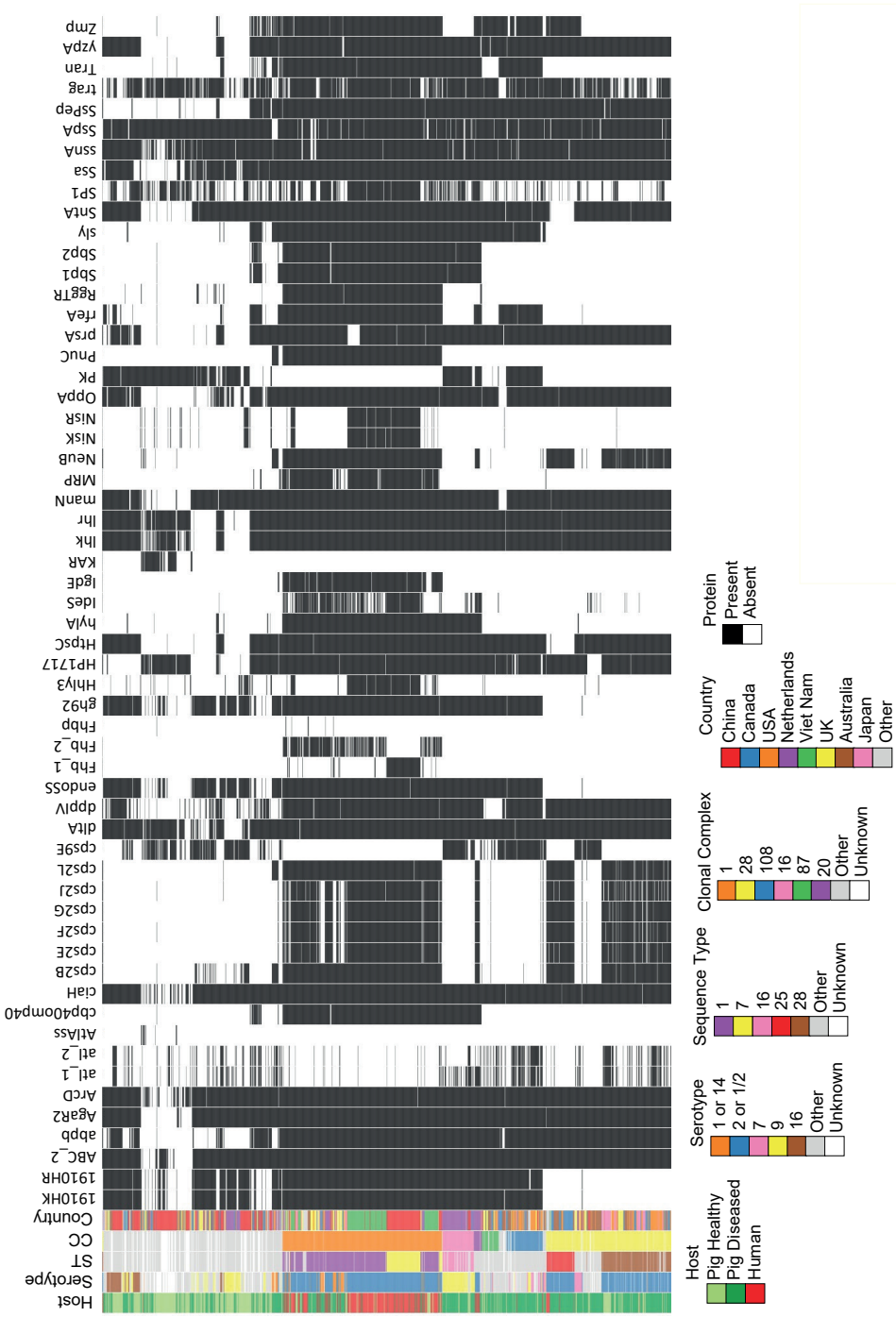


Figure 3. Presence of virulence factors in *S. suis* isolates and the corresponding virulence factor prevalence ratio in human isolates compared to pig isolates.

Geographical clustering of *S. suis* lineages may explain the presence of NisK/R and Hhly3 in zoonotic isolates from certain countries. Therefore, we determined the prevalence ratio of these proteins per country of origin. The prevalence ratio within Chinese isolates was 6·4 for NisK/R and 5·5 for Hhly3. In addition, inclusion of multiple strains belonging to a single outbreak may cause confounding. When isolates from the Chinese outbreak in 2005[68], which all except one harbored NisK/R and Hhly3, were excluded, the prevalence ratio within Chinese isolates was 3·7 for NisK/R and 3·1 for Hhly3. In Vietnamese isolates the prevalence ratio for NisK/R was 1·5 and for Hhly3 1·4. NisK/R and Hhly3 were not detected in human isolates from other countries than China and Vietnam.

Discussion

We identified 124 *S. suis* putative virulence factors studied in a human model in our systematic review. In our subsequent genomic meta-analysis, we identified 26 putative virulence factors with prevalence at least two times higher in human isolates than in pig isolates, which were therefore considered as PZVFs.

The five virulence factors most studied in *in vitro* models of human origin were CPS, Sly, MRP, Fhb and enolase. The contribution of these five virulence factors to *S. suis* virulence has also been studied *in vivo* in pig and mouse infection models. In a review of studies of Sly, MRP and Fhb, these putative virulence factors were found not to be critical for virulence in all models[4]. CPS was shown to contribute to *S. suis* virulence *in vivo* in pig and mice[15,21,69,70,71]. Both Sly and MRP contributed to virulence in mice[50,51,28,72,73], but a Sly or MRP KO did not show decreased virulence in pigs[38,43,74]. Fhb was shown to contribute to virulence in pigs[55] and to be essential to cross the BBB via Gb3 in mice[58]. Enolase was only tested in mice and increased the BBB permeability[75]. Pig and mouse *in vivo* infection models appear to yield different outcomes for certain virulence factors. A similar observation was made for the difference in virulence of different *S. suis* serotype 2 strains, observed after experimental infections in pig and mouse[2]. These data indicate that, although we can learn much from these *in vivo* models, the translation of mice or pig infection studies to the human *S. suis* pathogenesis can be challenging.

In our genomic meta-analysis, proteins involved in the serotype 2 capsular polysaccharide biosynthesis were more prevalent in zoonotic isolates, confirming epidemiological observations[1]. Sly, MRP, and Fhb were identified as PZVF, while enolase was found to be part of the *S. suis* core genome and therefore not identified as PZVF. Only Fhb_1 remained more prevalent in human than in pig isolates within the CC1 lineage. In a previous genomic analysis, a comparison between human and pig isolates from Vietnam and pig isolates from the UK did not find a substantial enrichment of specific accessory

genes in human isolates[76]. The prevalence of virulence factors was higher in clinical pig isolates than in non-clinical isolates from the UK[76]. Putative virulence factors were more abundant in Dutch zoonotic isolates than in non-zoonotic isolates[3]. Zoonotic and non-zoonotic strains could only be separated based on their accessory genome and not based on their core genome[3]. Moreover, zoonotic isolates with dissimilar core genomes showed similarity in their accessory genome[3], implying that PZVFs are most likely part of the accessory genome. Here, 53 of the putative virulence factors were encoded by genes which are part of the *S. suis* core genome. Given their function (appendix S2), many of these putative virulence factors are likely to be involved in *S. suis* metabolism although a role in pathogenesis cannot be ruled out. As was noted before and was also observed in this study, many *S. suis* putative virulence factors have not yet been thoroughly characterized[4]. Most virulence factors were studied in a single isolate instead of multiple isolates, introducing potential bias[77]. An additional concern is that isogenic KO mutants used to study the virulence factors were not always properly characterized. In 37% of the studies that used an isogenic KO mutant, the impact of the mutation on growth rate was not verified and therefore a direct effect of the KO on the experimental outcome due to changes in growth rate, instead of or in addition to a potential functional effect, cannot be ruled out.

Independent parallel genomic acquisition events can introduce different PZVFs that could drive the emergence of a zoonotic *S. suis* lineage, as observed in the Dutch zoonotic CC20 lineage[3]. Such acquisition event could explain why NisK/R or Hhly3 are not present in all human *S. suis* isolates. These findings suggest that these specific PZVFs are not essential for zoonotic potential per se, as the acquisition of other genes could confer zoonotic potential as well. However, within the zoonotic CC1 lineage or after stratification by country of origin, NisK/R and Hhly3, as well as Fhb_1 are still more prevalent in human isolates than in pig isolates suggesting that these PZVFs contribute to zoonotic potential.

Hhly3 is a cholesterol-independent hemolysin first discovered in the foodborne pathogen *Bacillus cereus*[78] and later also identified in the foodborne pathogen *Vibrio vulnificus*[79]. Hhly3 monomers bind in a temperature-dependent fashion to host cell membranes and form 3–3·5 mm pores after multimerization[80]. The cholesterol independency of Hhly3 could give *S. suis* the ability to induce pores in membranes with low cholesterol or unavailable cholesterol, such as endosomes[81]. The contribution of Hhly3 to *S. suis* virulence has not been studied in *in vivo* pig or mouse infection models yet.

Nisin is al antibiotic produced by several *Lactococcus* and *Streptococcus* species with antimicrobial properties against Gram-positive and Gram-negative bacteria[82]. Operons conferring nisin resistance in strains that cannot produce nisin themselves have mainly been found in human pathogenic strains, including Streptococcus mutants and

Streptococcus agalactiae[82]. In *S. suis*, three independent acquisitions of nisin resistance genes have been reported. A complete nisin production and resistance locus including NisK/R was found on two different pathogenicity islands in two unrelated strains[83,84]. NisK/R was also present on the 89 K pathogenicity island in a CC1/ST7 strain from China[66]. Here we also detected NisK/R in CC1/ST1 strains from Vietnam. Besides conferring nisin resistance, NisK/R could potentially contribute to zoonotic potential by regulating gene expression[85]. NisK/R was demonstrated to contribute to *S. suis* virulence in mice[66]. A NisK/R KO mutant was shown to have decreased hemolytic activity and decreased adhesion to and invasion of HeLa cells[66].

Our study has several limitations. To determine the presence of the putative virulence factors in *S. suis* genomes, we mapped the proteins to the assembled genomes using a minimal identity of 95%. Although this cutoff can distinguish between virulent and avirulent MRP[50], it cannot distinguish small differences at the amino acid level. However, a single amino acid change can affect the function of a putative virulence factor, as for example recently shown for SadP[86]. Additionally, we included two articles in the systematic review that studied sRNAs[87,88], but our protein mapping approach did not permit meta-analysis of regulatory RNA molecules or regulatory non-coding DNA sequences that could contribute to virulence. Moreover, we determined the presence of single virulence factors and did not study a potential combined effect of virulence factors. For the proteins encoded by genes of the accessory genome we attempted to compare their prevalence in human and pig isolates per study, which would allow for a combined statistical analysis comparable to an individual patient data meta-analysis. However, only a single study systematically sampled both pig and human isolates[3], precluding such meta-analysis. Whilst we included all *S. suis* genomes with accompanying metadata present in NCBI BioSample, for 31% of BioSample records metadata were lacking, likely introducing further bias. We tried to overcome this limitation partly by performing our analysis within genomic lineage CC1 and for individual countries.

Genomic determinants associated with particular bacterial traits are increasingly identified using genome-wide association studies. Such studies require confirmation of biological relevance of genes with significant association. Here, we used a different approach by starting with a systematic approach toward identification of functional proteins and subsequent estimation of their relative frequency in genomes of strains representing different *S. suis* populations. The collected metadata with corresponding assembled genomes and the list of PZVF are valuable tools for further research into zoonotic potential of *S. suis*, the pathogenesis of zoonotic *S. suis* infections, and for early detection of emerging zoonotic lineages.

Acknowledgments

This work was funded through EU-Horizon2020 grant 727966 (PIGSs).

Funding Statement

This research has received funding from the European Union's Horizon 2020 research and innovationprogramme under grant agreement No 727966 (https://cordis.europa.eu/project/id/727966).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Contributors

CS, TR and KA conceived the study. TR and KA did the abstract screening. Full text was read by TR and data was extracted by ST and TR. KA curated included articles and data extraction. TR collected metadata from BioSample records and divided records over groups. Genomic analysis was performed by BP. TR made visualizations and drafted the manuscript. All authors contributed to the final version of the manuscript.

Data Availability statement

"The data that support the findings of this study are available in Zenodo at https://doi.org/10.5281/zenodo.4686597. These data were derived from the following resources available in the public domain: NCBI (https://www.ncbi.nlm.nih.gov/)."

References

1. Huong VTL, Ha N, Huy NT, *et al.* Epidemiology, clinical manifestations, and outcomes of *Streptococcus suis* infection in humans. *Emerg Infect Dis J*. 2014;20:1105.
2. Vecht U, Stockhofe-Zurwieden N, Tetenburg BJ, *et al.* Virulence of *Streptococcus suis* type 2 for mice and pigs appeared host- specific. *Vet Microbiol*. 1997;58:53–60.
3. Willemse N, Howell KJ, Weinert LA, *et al.* An emerging zoonotic clone in the Netherlands provides clues to virulence and zoonotic potential of *Streptococcus suis*. *Sci Rep*. 2016;6:28984.
4. Segura M, Fittipaldi N, Calzas C, *et al.* Critical *Streptococcus suis* virulence factors: are they all really critical? *Trends Microbiol*. 2017;25:585–599.
5. Fulde M , Valentin-Weigand P. Epidemiology and pathogenicity of zoonotic streptococci. *Curr Top Microbiol Immunol*. 2013;368:49-81. DOI: 10.1007/82_2012_277.PMID:23192319
6. Arenas J, Zomer A, Harders-Westerveen J, *et al.* Identification of conditionally essential genes for *Streptococcus suis* infection in pigs. *Virulence*. 2020;11:446–464.
7. Saralahti, A. and Rämet, M. Zebrafish and Streptococcal Infections. *Scand J Immunol*. 2015;82:174–183.
8. Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4:1.
9. Koster J. Pubmed Pubreminer. AMC, UvA; 2004. http://hgserver2.amc.nl/cgi-bin/miner/miner2.cgi
10. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: bIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124.
11. Minh BQ, Schmidt HA, Chernomor O, *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic Era. *Mol Biol Evol*. 2020;37:1530–1534.
12. Page AJ, Cummins CA, Hunt M, *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691–3693.
13. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–2069.
14. Hadfield J, Croucher NJ, Goater RJ, *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*. 2017;34:292–293.
15. Smith HE, Damman M, van der Velde J, *et al.* Identification and characterization of the cps locus of *Streptococcus suis* serotype 2: the capsule protects against phagocytosis and is an important virulence factor. *Infect Immun*. 1999;67:1750–1756.
16. Okura M, Takamatsu D, Maruyama F, *et al.* Genetic analysis of capsular polysaccharide synthesis gene clusters from all serotypes of *Streptococcus suis*: potential mechanisms for generation of capsular variation. *Appl Environ Microbiol*. 2013;79:2796–2806.
17. Lalonde M, Segura M, Lacouture S, *et al.* Interactions between *Streptococcus suis* serotype 2 and different epithelial cell lines. *Microbiology-Sgm*. 2000;146:1913–1921.
18. Zhang Y, Ding D, Liu M, *et al.* Effect of the glycosyltransferases on the capsular polysaccharide synthesis of *Streptococcus suis* serotype 2. *Microbiol Res*. 2016;185:45–54.
19. Benga L, Goethe R, Rohde M, *et al.* Non-encapsulated strains reveal novel insights in invasion and survival of *Streptococcus suis* in epithelial cells. *Cell Microbiol*. 2004;6:867–881.
20. Ferrando ML, De Greeff A, van Rooijen WJ, *et al.* Host-pathogen Interaction at the intestinal mucosa correlates with zoonotic potential of *Streptococcus suis*. *J Infect Dis*. 2015;212:95–105.

21. Feng Y, Cao M, Shi J, *et al.* Attenuation of *Streptococcus suis* virulence by the alteration of bacterial surface architecture. *Sci Rep*. 2012;2:710.

22. Salasia SI, Lammler C, Herrmann G. Properties of a *Streptococcus suis* isolate of serotype 2 and two capsular mutants. *Vet Microbiol*. 1995;45:151–156.

23. Huang W, Chen Y, Li Q, *et al.* LytR plays a role in normal septum formation and contributes to full virulence in *Streptococcus suis*. *Vet Microbiol*. 2021;254:109003.

24. Esgleas M, Lacouture S, Gottschalk M. *Streptococcus suis* serotype 2 binding to extracellular matrix proteins. *FEMS Microbiol Lett*. 2005;244:33–40.

25. Jobin MC, Gottschalk M, Grenier D. Upregulation of prostaglandin E2 and matrix metalloproteinase 9 production by human macrophage-like cells: synergistic effect of capsular material and cell wall from *Streptococcus suis*. *Microb Pathog*. 2006;40:29–34.

26. Meijerink M, Ferrando ML, Lammers G, *et al.* Immunomodulatory effects of *Streptococcus suis* capsule type on human dendritic cell responses, phagocytosis and intracellular survival. *PLoS One*. 2012;7:e35849.

27. Zaccaria E, Cao R, Wells JM, *et al.* Model to assess virulence of porcine *Streptococcus suis* strains. *PLoS One*. 2016;11:e0151623.

28. Lin L, Xu L, Lv W, *et al.* An NLRP3 inflammasome-triggered cytokine storm contributes to streptococcal toxic shock-like syndrome (STSLS). *PLoS Pathog*. 2019;15:e1007795.

29. Roy D, Grenier D, Segura M, *et al.* Recruitment of factor H to the *Streptococcus suis* cell surface is multifactorial. *Pathogens*. 2016;5. DOI:10.3390/pathogens5030047.

30. Graveline R, Segura M, Radzioch D, *et al.* TLR2-dependent recognition of *Streptococcus suis* is modulated by the presence of capsular polysaccharide which modifies macrophage responsiveness. *Int Immunol*. 2007;19:375–389.

31. Auger JP, Christodoulides M, Segura M, *et al.* Interactions of *Streptococcus suis* serotype 2 with human meningeal cells and astrocytes. *BMC Res Notes*. 2015;8:607.

32. Schwerk C, Papandreou T, Schuhmann D, *et al.* Polar invasion and translocation of neisseria meningitidis and *Streptococcus suis* in a novel human model of the blood-cerebrospinal fluid barrier. *PLoS One*. 2012;7:e30069.

33. Vadeboncoeur N, Segura M, Al-Numani D, *et al.* Pro-inflammatory cytokine and chemokine release by human brain microvascular endothelial cells stimulated by *Streptococcus suis* serotype 2. *FEMS Immunol Med Microbiol*. 2003;35:49–58.

34. Jacobs AA, Loeffen PL, Van Den Berg AJ, *et al.* Identification, purification, and characterization of a thiol-activated hemolysin (suilysin) of *Streptococcus suis*. *Infect Immun*. 1994;62:1742–1748.

35. Leung C, N V D, Lukoyanova N, *et al.* Stepwise visualization of membrane pore formation by suilysin, a bacterial cholesterol-dependent cytolysin. *Elife*. 2014;3:e04247.

36. Norton PM, Rolph C, Ward PN, *et al.* Epithelial invasion and cell lysis by virulent strains of *Streptococcus suis* is enhanced by the presence of suilysin. *FEMS Immunol Med Microbiol*. 1999;26:25–35.

37. Seitz M, Baums CG, Neis C, *et al.* Subcytolytic effects of suilysin on interaction of *Streptococcus suis* with epithelial cells. *Vet Microbiol*. 2013;167:584–591.

38. Lun S, Perez-Casal J, Connor W, *et al.* Role of suilysin in pathogenesis of *Streptococcus suis* capsular serotype 2. *Microb Pathog*. 2003;34:27–37.

39. Chen S, Xie W, Wu K, *et al.* Suilysin stimulates the release of heparin binding protein from neutrophils and increases vascular permeability in mice. *Front Microbiol*. 2016;7:1338.

40. Zhang S, Zheng Y, Chen S, *et al.* Suilysin-induced platelet-neutrophil complexes formation is triggered by pore formation-dependent calcium influx. *Sci Rep*. 2016;6:36787.

41. Jobin MC, Fortin J, Willson PJ, *et al.* Acquisition of plasmin activity and induction of arachidonic acid release by *Streptococcus suis* in contact with human brain microvascular endothelial cells. *FEMS Microbiol Lett*. 2005;252:105–111.

42. Vecht U, Wisselink HJ, Jellema ML, *et al.* Identification of two proteins associated with virulence of *Streptococcus suis* type 2. *Infect Immun*. 1991;59:3156–3162.

43. Smith HE, Vecht U, Wisselink HJ, *et al.* Mutants of *Streptococcus suis* types 1 and 2 impaired in expression of muramidase-released protein and extracellular protein induce disease in newborn germfree pigs. *Infect Immun*. 1996;64:4409–4412.

44. Baums CG, Valentin-Weigand P. Surface-associated and secreted factors of *Streptococcus suis* in epidemiology, pathogenesis and vaccine development. *Anim Heal Res Rev*. 2009;10:65.

45. Smith HE, Vecht U, Gielkens AL, *et al.* Cloning and nucleotide sequence of the gene encoding the 136-kilodalton surface protein (muramidase-released protein) of *Streptococcus suis* type 2. *Infect Immun*. 1992;60:2361–2367.

46. Zhang W, Liu G, Tang F, *et al.* Pre-absorbed immunoproteomics: a novel method for the detection of *Streptococcus suis* surface proteins. *PLoS One*. 2011;6:e21234.

47. Rui L, Weiyi L, Yu M, *et al.* The serine/threonine protein kinase of *Streptococcus suis* serotype 2 affects the ability of the pathogen to penetrate the blood-brain barrier. *Cell Microbiol*. 2018;20:e12862.

48. Pian Y, Wang P, Liu P, *et al.* Proteomics identification of novel fibrinogen-binding proteins of *Streptococcus suis* contributing to antiphagocytosis. *Front Cell Infect Microbiol*. 2015;5:19.

49. Pian Y, Li X, Zheng Y, *et al.* Binding of human fibrinogen to MRP enhances *Streptococcus suis* survival in host blood in a alphaXbeta2 Integrin-dependent manner. *Sci Rep*. 2016;6:26966.

50. Li Q, Fu Y, Ma C, *et al*. The non-conserved region of MRP is involved in the virulence of *Streptococcus suis* serotype 2. *Virulence*. 2017;8:1274–1289.

51. Wang J, Kong D, Zhang S, *et al.* Interaction of fibrinogen and muramidase-released protein promotes the development of *Streptococcus suis* meningitis. *Front Microbiol*. 2015;6:1001.

52. Li Q, Ma C, Fu Y, *et al*. Factor H specifically capture novel Factor H-binding proteins of *Streptococcus suis* and contribute to the virulence of the bacteria. *Microbiol Res*. 2017;196:17–25.

53. Li X, Liu P, Gan S, *et al.* Mechanisms of host-pathogen protein complex formation and bacterial immune evasion of *Streptococcus suis* protein Fhb. *J Biol Chem*. 2016;291:17122–17132.

54. Kouki A, Haataja S, Loimaranta V, *et al.* Identification of a novel streptococcal adhesin P (SadP) protein recognizing galactosyl-α1-4-galactose-containing glycoconjugates: convergent evolution of bacterial pathogens to binding of the same host receptor. *J Biol Chem*. 2011;286:38854–38864.

55. Pian Y, Gan S, Wang S, *et al.* Fhb, a novel factor H-binding surface protein, contributes to the antiphagocytic ability and virulence of *Streptococcus suis*. *Infect Immun*. 2012;80:2402–2413.

56. Madar Johansson M, Bélurier E, Papageorgiou AC, *et al.* The binding mechanism of the virulence factor *Streptococcus suis* adhesin P subtype to Globotetraosylceramide is associated with systemic disease. *J Biol Chem*. 2020;295:14305–14324.

57. Ferrando ML, Willemse N, Zaccaria E, *et al.* Streptococcal adhesin P (SadP) contributes to *Streptococcus suis* adhesion to the human intestinal epithelium. *PLoS One*. 2017;12:e0175639.

58. Kong D, Chen Z, Wang J, *et al.* Interaction of factor H-binding protein of *Streptococcus suis* with globotriaosylceramide promotes the development of meningitis. *Virulence*. 2017;8:1290–1302.

59. À D-R, Roig-Borrellas A, García-Melero A, *et al.* α-Enolase, a multifunctional protein: its role on pathophysiological situations. *J Biomed Biotechnol*. 2012;2012:156795. doi: 10.1155/2012/156795

60. Esgleas M, Li Y, Hancock MA, *et al.* Isolation and characterization of alpha-enolase, a novel fibronectin-binding protein from *Streptococcus suis*. *Microbiology*. 2008;154:2668–2679.

61. Li Q, Liu H, Du D, *et al.* Identification of novel laminin- and fibronectin-binding proteins by far-western blot: capturing the adhesins of *Streptococcus suis* type 2. *Front Cell Infect Microbiol*. 2015;5:82.

62. Feng Y, Pan X, Sun W, *et al. Streptococcus suis* enolase functions as a protective antigen displayed on the bacterial cell surface. *J Infect Dis*. 2009;200:1583–1592.

63. Chen B, Zhang A, Xu Z, *et al.* Large-scale identification of bacteria-host crosstalk by affinity chromatography: capturing the interactions of *Streptococcus suis* proteins with host cells. *J Proteome Res*. 2011;10:5163–5174.

64. Jiang H, Wu T, Liu J, *et al.* Caveolae/rafts protect human cerebral microvascular endothelial cells from *Streptococcus suis* serotype 2 α-enolase-mediated injury. *Vet Microbiol*. 2021;254:108981.

65. Liu H, Lei S, Jia L, *et al. Streptococcus suis* serotype 2 enolase interaction with host brain microvascular endothelial cells and RPSA-induced apoptosis lead to loss of BBB integrity. *Vet Res*. 2021;52. DOI:10.1186/s13567-020-00887-6.

66. Xu J, Fu S, Liu M, *et al.* The two-component system NisK/NisR contributes to the virulence of *Streptococcus suis* serotype 2. *Microbiol Res*. 2014;169:541–546.

67. Zheng JX, Li Y, Zhang H, *et al.* Identification and characterization of a novel hemolysis-related gene in *Streptococcus suis* serotype 2. *PLoS One*. 2013;8:e74674.

68. Du P, Zheng H, Zhou J, *et al.* Detection of multiple parallel transmission outbreak of *Streptococcus suis* human infection by use of genome epidemiology, China, 2005. *Emerg Infect Dis*. 2017;23:204–211.

69. Charland N, Harel J, Kobisch M, *et al. Streptococcus suis* serotype 2 mutants deficient in capsular expression. *Microbiology*. 1998;144(Pt 2):325–332.

70. Okura M, Auger J-P, Shibahara T, *et al.* Capsular polysaccharide switching in *Streptococcus suis* modulates host cell interactions and virulence. *Sci Rep*. 2021;11. DOI:10.1038/s41598-021-85882-3.

71. Lin L, Xu L, Lv W, *et al.* An NLRP3 inflammasome-triggered cytokine storm contributes to streptococcal toxic shock-like syndrome (STSLS). *PLoS Pathog*. 2019;15:e1007795.

72. He Z, Pian Y, Ren Z, *et al.* Increased production of suilysin contributes to invasive infection of the *Streptococcus suis* strain 05ZYH33. *Mol Med Rep*. 2014;10:2819–2826.

73. Takeuchi D, Akeda Y, Nakayama T, *et al.* The contribution of suilysin to the pathogenesis of *Streptococcus suis* meningitis. *J Infect Dis*. 2014;209:1509–1519.

74. Allen AG, Bolitho S, Lindsay H, *et al.* Generation and characterization of a defined mutant of *Streptococcus suis* lacking suilysin. *Infect Immun*. 2001;69:2732–2735.

75. Sun Y, Li N, Zhang J, *et al.* Enolase of *Streptococcus suis* serotype 2 enhances blood-brain barrier permeability by inducing IL-8 release. *Inflammation*. 2016;39:718–726.

76. Weinert LA, Chaudhuri RR, Wang J, *et al.* Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nat Commun*. 2015;6:6740.

77. Auger JP, Chuzeville S, Roy D, *et al.* The bias of experimental design, including strain background, in the determination of critical *Streptococcus suis* serotype 2 virulence factors. *PLoS One*. 2017;12:e0181920.

78. Baida GE, Kuzmin NP. Cloning and primary structure of a new hemolysin gene from Bacillus cereus. *Biochim Biophys Acta*. 1995;1264:151–154.

79. Chen Y-C, Chang M-C, Chuang Y-C, *et al.* Characterization and virulence of hemolysin III from vibrio vulnificus. *Curr Microbiol*. 2004;49:175–179.

80. Baida GE, Kuzmin NP. Mechanism of action of hemolysin III from Bacillus cereus. *Biochim Biophys Acta (BBA)-Biomembranes*. 1996;1284:122–124.

81. Bieberich E. Sphingolipids and lipid rafts: novel concepts and methods of analysis. *Chem Phys Lipids*. 2018;216:114–131.

82. Khosa S, Lagedroste M, Smits SHJ. Protein defense systems against the lantibiotic nisin: function of the immunity protein NisI and the resistance protein NSR. *Front Microbiol*. 2016;7:504.

83. Wu Z, Wang W, Tang M, *et al.* Comparative genomic analysis shows that *Streptococcus suis* meningitis isolate SC070731 contains a unique 105K genomic island. *Gene*. 2014;535:156–164.

84. Zhu Y, Zhang Y, Ma J, *et al.* ICESsuHN105, a novel multiple antibiotic resistant ICE in *Streptococcus suis* serotype 5 strain HN105. *Front Microbiol*. 2019;10:274.

85. Kawada-Matsuo M, Watanabe A, Arii K, *et al. Staphylococcus aureus* virulence affected by an alternative Nisin A resistance mechanism. *Appl Environ Microbiol*. 2020. Apr 1;86(8):e02923-19. doi: 10.1128/AEM.02923-19. PMID: 32086306.

86. Johansson MM, Bélurier E, Papageorgiou AC, *et al.* The binding mechanism of the virulence factor *Streptococcus suis* adhesin P subtype to globotetraosylceramide is associated with systemic disease. *J Biol Chem*. 2020;295:14305–14324.

87. Wu Z, Wu C, Shao J, *et al.* The *Streptococcus suis* transcriptional landscape reveals adaptation mechanisms in pig blood and cerebrospinal fluid. *RNA*. 2014;20:882–898.

88. Gong X, Zhuge Y, Ding C, *et al.* A novel small RNA contributes to restrain cellular chain length and anti-phagocytic ability in *Streptococcus suis* 2. *Microb Pathog*. 2019;137:103730.

Supplementary material

Supplemental data for this article can be accessed from Figshare at https://tinyurl.com/js2nc9

Chapter 10

General Discussion

In this thesis I identified ESBL-positive *Escherichia coli* types capable of persistent colonisation of the human gut, contributing to global spread after travel (Chapter 4). In addition, in Chapter 5 I identified *E. coli* genetic factors associated with the capacity to efficiently colonise humans. The other chapters presented in this thesis offer critical evaluation of the conditions for optimal datasets and methodologies needed to complete this type of research.

Extraintestinal pathogenic *Escherichia coli*

Extraintestinal pathogenic *E. coli* (ExPEC) are defined as *E. coli* types capable of causing extraintestinal disease[1]. Major ExPEC lineages, as defined by sequence type (ST), are ST131, ST73 and ST69. In Chapter 4, I presented the finding that specific ExPEC lineages are strongly associated with an increased carriage duration in international travellers. In addition, the *E. coli* lineages associated with human host colonisation are also considered to be ExPEC (Chapter 5). However, defining when exactly a lineage can be considered capable of causing extraintestinal disease is challenging and varying definitions are used. Some authors consider any lineage of which *E. coli* strains have been shown to cause extraintestinal disease to be ExPEC[2] while others employ a stricter definition and only consider lineages that are often found to cause extraintestinal disease to be ExPEC[3]. Classical ExPEC lineages such as ST131, ST69 and ST73 seem to be well-adapted to colonise the human intestine. In Dutch and Danish surveillance studies, most ExPEC lineages were similarly distributed among extraintestinal infection isolates and asymptomatic faecal carriage isolates[4,5]. In the Dutch study, only ST131 was more prevalent in bloodstream isolates compared to asymptomatic faecal carriage isolates[4]. Additionally, nearly all travellers included in the study presented in Chapter 4 did not report any intestinal or extraintestinal symptoms, in line with other reports of studies about intestinal ExPEC colonisation[6]. Even before the emergence of the most prevalent ExPEC lineage (ST131)[2,7], it was noted that the capability of ExPEC to cause extraintestinal disease might be an "evolutionary by-product" as a result of increased fitness in the human intestine[8]. The ability to cause extraintestinal disease does not increase the transmissibility of *E. coli* in itself, as extraintestinal sites represent a "dead end" for *E. coli* transmission[8]. This raises the question whether the term "ExPEC" correctly reflects the ecology of these *E. coli*. Multiple ExPEC lineages certainly have the virulence genes necessary for extraintestinal infection[9], but the capability of these lineages to colonise the human intestine might play a larger role than previously appreciated. Possibly, the ability of ExPEC to cause extraintestinal disease depends on two factors. First, an increased fitness in the human intestine leads to a higher chance of ExPEC reaching a host environment which is susceptible to

extraintestinal infection. Secondly, ExPEC lineages possess the necessary virulence factors to invade, survive and cause disease once a susceptible host environment is reached. The term ExPEC might mislead readers which are not well acquainted with *E. coli* biology. The term emphasises the virulent properties of these *E. coli* while asymptomatic colonisation of the gut seems to be the primary niche for these *E. coli* lineages.

In Chapter 5, *E. coli* lineages and genes associated with human colonisation were identified. As anticipated, ST131 was the lineage most strongly associated with human colonisation. Additionally, 106 genes were associated with human colonisation. A commonly held view is that ExPEC fitness is largely driven by virulence factors[10–12] (bacterial factors determining capacity to cause disease but not required for bacterial viability[13]). However, of these 106 genes associated with human colonisation, only four – *iucB*, *iucC*, *sat* and *papA* – encode known virulence factors. For *papA* (part of the P fimbriae operon, involved in adherence[14]) and *iucB* and *iucC* (part of the aerobactin operon, involved in iron acquisition[15]), one could argue that these genes do not strictly encode virulence properties, but may also offer an evolutionary advantage for asymptomatic colonisation of the human gut. Most human-associated genes were predicted to encode metabolic enzymes. Notably, a previously unknown cluster of nine genes involved in sialic acid catabolism was found to be strongly associated with human colonisation. These findings further stress that ExPEC have not necessarily evolved towards extraintestinal virulence, but rather towards efficient intestinal colonisation. Future endeavours investigating *E. coli* host adaptation should ensure that metabolic adaptation markers receive adequate attention (also see Future perspectives at the end of this chapter).

Although animal hosts are seemingly not the most important reservoirs of *E. coli* causing human infection[16,17], some *E. coli* lineages found among animal hosts might facilitate transmission of antimicrobial (AMR) genes between host populations. In Chapter 5, isolates belonging to clonal complex 10 (CC10), comprising ST10 and related STs, were isolated from four host species. The *mcr*-1 gene, which might have originated in an animal population[18], is associated with CC10 isolates[19]. This was also observed in the isolate collection studied in Chapter 5 (data not shown). Lineages with similar properties (e.g. CC155) have been previously identified[20]. Although lineages such as CC10 or CC155 might not be the most common cause of *E. coli* disease in humans, these lineages are important to monitor closely. Possibly, these lineages might transmit AMR genes between host populations and therefore make an ideal study case for the application of the One Health framework.

Challenges of bacterial typing

A crucial step in many bacterial genomics analyses is to define discrete "types", which can be compared in studies to infer the genetic relationship between bacteria[21]. Whether it is to trace outbreak progression, to relate phenotypes to genotypes or to infer epidemiological parameters, bacterial samples often need to be discretely grouped. These discrete groups are typically assumed to reflect a shared descent between bacteria. Due to the increased accessibility of whole-genome sequencing (WGS), resolution of typing analyses has greatly increased. However, certain issues with typing have persisted or have been emphasised by WGS analyses.

To divide bacteria into discrete types, cut-off values of similarity have to be used. In this final chapter, I will discuss two genetic similarity cut-offs which have proven challenging to define: the threshold to determine recent *E. coli* transmission and the cut-off for bacterial species delineation.

To assess whether two bacterial isolates share a recent common ancestor, single nucleotide polymorphisms (SNPs) are typically analysed. Usually, a numerical threshold is defined and if two isolates differ in fewer SNPs than the threshold, there is reason to assume that the isolates share a very recent common ancestor. This would indicate possible strain persistence or recent transmission. The SNP threshold accounts for the number of polymorphisms that can be accumulated during a certain period, but also accounts for technical variability that can be expected in the analysis (e.g. due to genomic regions that are difficult to map reads to). In Chapter 4, I defined a cut-off of 25 SNPs to determine which *E. coli* strains were persistent over a period of twelve months. However, I noticed that ST38 isolates from reportedly unrelated travellers differed in fewer than 25 SNPs, possibly suggesting recent transmission. Nonetheless, I could not find any relations between the nine travellers harbouring the highly similar isolates. Based on publicly available WGS data, I determined that these ST38 isolates belonged to two distinct clonal sublineages of ST38. Both sublineages were identified on three or more continents while remaining highly clonal, often differing in 25 SNPs or fewer within a sublineage. Decreasing the SNP cut-off might help in specific cases, but this would lead to true persistence events being missed. Therefore, these findings also highlight that bacterial SNP analyses require more than just comparing the number of SNPs. Relevant analysis parameters need to be considered, such as the expected diversity of the population under study, the sampling density and which background population to compare against, in addition to essential epidemiological metadata. Considering this auxiliary information also requires the development of computational frameworks which can incorporate this information[22]. Better reporting of underlying considerations and methods in SNP analyses would also contribute towards more standardised SNP analyses. As an extension

of standardised reporting guidelines for observational epidemiological studies (STROBE guidelines[23]), specific guidelines for molecular epidemiology in infectious diseases have been compiled (STROME-ID guidelines[24]). Reporting relevant considerations and methods using the STROBE and STROME-ID guidelines would facilitate comparing SNP analyses and possibly aid in establishing best practices for *E. coli* epidemiology. However, a systematic review found that for studies investigating the molecular epidemiology of *Mycobacterium tuberculosis*, adherence to these reporting guidelines was low[25].

Prior to widespread application of WGS analyses, DNA-DNA hybridization (DDH) values have been used to delineate microbial species[26]. With the advent of cost-efficient WGS, many studies have employed WGS to assign novel species. From WGS data, the average nucleotide identity (ANI) can be calculated between two genomes. This value indicates how similar two sequences are at nucleotide level, for the regions that are shared between those two sequences. In an early study using 28 microbial strains, an ANI value of 95% was found to correlate well with the 70% DDH threshold originally used to delineate species[27]. In Chapter 3, I employed this 95% ANI cut-off to assign *Escherichia ruysiae* as a novel species. The *Escherichia* genus can be divided into mostly discrete clusters based on this cut-off (*E. coli* including cryptic clade I, *E. whittamii*[28], *E. fergusonii*[29], *E. albertii*[30], *E. ruysiae* (Chapter 3), *E. marmotae*[31], cryptic clade VI[32] and possibly additional cryptic clades). After a large-scale analysis of >90,000 publicly available prokaryotic genomes, only 0.21% of comparisons of paired isolates showed an ANI value between 83% and 95%, indicating most comparisons fall either above the 95% ANI threshold or well below it[33]. This might suggest that ANI values facilitate a nearly dichotomous assignment to species. However, many genera (including *Streptococcus*, *Staphylococcus*, *Pseudomonas*, *Escherichia*, *Enterobacter*, *Campylobacter* and others) show a long tail of ANI values between 83% and 95% ANI which hinders unambiguous assignment of strains to species, even though ANI values between 83% and 95% remain rare[34]. The criteria for definitive species delineation are still subject to much discussion, although ANI values might seem to offer a practical solution for most bacteria.

Both examples outlined above show that fitting discrete genetic types on complex genomic data is challenging and impossible to perform perfectly. It is important to realise genomic types are a proxy for the genomic identity of a strain, functioning as a human-designed method to interpret similarity between complex and evolving genomes.

Deciphering the genetics of bacterial phenotypes

One of the primary goals of bacterial WGS analysis is to identify the genotypic basis for certain bacterial phenotypes. For example, genes contributing to virulence[35], antimicrobial resistance[36] or carriage duration[37] have been identified this way. However, several considerations are important for this kind of analysis to be successful, which I have outlined below.

Isolate selection

How isolates are selected before WGS analysis has a dramatic impact on the validity and usefulness of subsequent analyses. In this section I will address two considerations on isolate selection: the number of isolates and the relationship between genotype and phenotype. Generally, including more isolates provides more statistical power and can better associate genotypes with a given phenotype[38]. As the presence or absence of each bacterial gene can be considered a bacterial genotype, oftentimes the number of genotypes far outnumbers the number of isolates in a GWAS[39]. Stringent cut-offs for statistical significance are therefore used (e.g. p-value$<1\times10^{-8}$). If few isolates can be included, it might be more appropriate to reduce the number of genotypes examined (similar to the approach in Chapter 4). Additionally, it might be feasible to include more isolates with a certain phenotype than another phenotype (class imbalance). The study presented in Chapter 4 encountered this as travellers with a very short carriage duration largely outnumbered those with a very long carriage duration. Class imbalance typically affects the sensitivity and specificity of the subsequent analysis[40]. Isolate selection for uneven group sizes might even be leveraged to increase either sensitivity or specificity of an analysis.

The relation between genotype and phenotype also influences the statistical power needed. Antimicrobial resistance is sometimes mediated by only a single or few genes[36], while more complex phenotypes such as host adaptation seem to be mediated by many genes (Chapter 5). Because host adaptation is a complex phenotype, the goal in Chapter 5 was to obtain a diverse set of *E. coli* from different host species to find genetic markers for host adaptation. To reach a sample size large enough for a genome-wide association study, we calculated that at least 200 isolates per host were needed. Thus, we included 1198 *E. coli* isolates from four host species and could associate genotypes to host adaptation for three out of four hosts. Additionally, if several different genotypes can lead to an identical phenotype, these different genotypes also need to be included appropriately. An example of this might be ciprofloxacin resistance in *E. coli* (Chapter 2), which may be caused by *gyrA* mutations, presence of *qnr* genes, or other genotypes, leading individually or combined to the same resistance phenotype. Not including enough samples from a particular lineage with its own genotype-phenotype relation

might lead to missing certain causative genotypes and a decreased sensitivity in new data.

Chapter 9 of this thesis investigated genetic factors contributing to virulence and zoonotic potential in *Streptococcus suis*. *S. suis* virulence is a complex phenotype to study and novel applications of GWAS methods might reveal additional insights on this topic. For example, most zoonotic *S. suis* display a serotype 2 capsule. Consequently, the contribution of serotype 2 capsule genes to *S. suis* zoonotic potential might overshadow other (epi-) genetic factors contributing to zoonosis. This could be overcome by conditioning a GWAS for known genetic determinants, in order to identify minor additional genetic determinants. Recently, such a conditional GWAS method was used to identify minor genetic determinants of azithromycin resistance in *Neisseria gonorrhoeae* in the presence of known azithromycin resistance mutations[41]. The additional resistance mutations that were detected were experimentally shown to significantly contribute to azithromycin resistance in *N. gonorrhoeae*. A similar approach could be used to identify additional genetic determinants of zoonotic potential in the presence of serotype 2 capsule genes in *S. suis*. Other developments such as the application of elastic net to GWAS could also help to reduce false positive associations, which might be expected due to the strong association between serotype 2 and virulence[42].

Data organisation

Databases contain structured sets of data and can often be used to relate one data type to another. For WGS analysis, publicly available databases are crucial and have numerous uses. Databases store isolate genomes used for comparisons[43], genotype-phenotype relationships used to predict phenotypes[44] or schemes used to assign types[45]. Two chapters in this thesis have produced data which can be incorporated for reuse in public databases: Chapter 8 resolved complete genomes of ubiquitously used *Streptococcus suis* strains and Chapter 2 compiled genotype-phenotype relations for *E. coli* ciprofloxacin resistance through a systematic literature review.

Aside from containing as much accurate information as possible, adhering to the FAIR principles increases the usefulness of databases for WGS analyses[46]. The FAIR principles state that data should be Findable, Accessible, Interoperable and Reusable. A great example of a database adhering to FAIR principles is PubMLST[45]. The database has a dedicated website and metadata is available in a searchable resource (Findable), provides an application programming interface (API) for programmatic access[47] as well as a graphical user interface for straightforward access (Accessible), is able to export and read data in a variety of widely-adopted file formats (Interoperable) and the database contains rich metadata that meets the community standard (Reusable). Adopting the FAIR principles allows for much more efficient use of databases such as PubMLST.

Another consideration on databases pertains to their maintenance. Database maintenance is important to add novel data or update obsolete information. Additionally, database infrastructure needs to be on par with the increasing data sizes in microbiology. This requires contributions by a wide variety of experts (subject matter experts, IT infrastructure engineers) and thus requires funding. Although acquiring continuous funding through grants is challenging, several funding agencies have or had dedicated funding calls for databases and software (e.g. Biomedical Resource Grant by the Wellcome Trust or the Open Science programme by the Chan-Zuckerberg Initiative). Additionally, researchers explore novel funding models for sustainable database maintenance[48].

Experimental follow-up

Finally, experimental studies remain an important companion to WGS analysis. Many databases mentioned previously contain aggregated data generated in experimental studies. Additionally, experimental validation forms an important confirmatory activity in studies investigating the genetic basis of phenotypes. A recently revised version of the molecular Koch's postulates, termed the next-generation Koch's postulates, could be a leading guideline for this effort[49]. In Chapter 5, colleagues and I validated the molecular function of a *nan* gene cluster, involved in sialic acid catabolism. The pace at which sample sizes of WGS studies are increasing seem larger than what classic microbiological techniques can support, possibly leading to an "analytical disconnect"[49]. As computational methods move into more complex genotype-phenotype relationships[50], experimental studies should consider the interplay between multiple mutations or the interplay between mutation and strain background. Although this often forms a challenging part of any study seeking the genetic basis of phenotypes, this remains the best way to prove a genotype underlies a certain phenotype.

Future perspectives

Although WGS has provided important insights on *E. coli* biology, pertinent questions remain. Many of these outstanding questions relate to the role of genetics and environment of *E. coli* in its dissemination and potential therapies against *E. coli* carriage or infection.

E. coli genetics

In Chapter 5, colleagues and I identified a cluster of nine genes associated with human colonisation of *E. coli*. Cleaved sialid acid is scavenged from the intestinal environment by *E. coli* and can be used as a carbon source. We have shown that all nine genes together contribute to more efficient growth when sialic acid is available as an energy source. Two out of nine genes encode putative sialylesterases, which might allow *E. coli*

to deacetylate sialic acids more efficiently. For *Streptococcus pneumoniae*, it has been shown that deacetylation of sialic acids contributes to bacterial fitness in *in vitro* growth and *in vivo* infection models[51]. Future studies might investigate this *E. coli* locus further, detailing its transcriptional regulation, differences with the core sialic acid catabolism genes of *E. coli* and its potential role in catabolising differentially acetylated sialic acids. In a broader sense, it would be interesting to investigate other metabolic genes which were associated with human colonisation in Chapter 5. Typically, virulence factors such as adhesins, fimbriae or outer membrane proteins are regarded as important determinants of host specificity. Bacterial metabolism might play a key role in host adaptation as well[52] and could be an exciting avenue for future research.

Additionally, the HECTOR project now focused on human host adaptation in a broad sense, although it can be argued that for *E. coli*, the human host consists of several interconnected environments (e.g. gastrointestinal tract, bloodstream, urinary tract). With current WGS analyses it is possible to dissect the genetic determinants allowing *E. coli* to colonise or exert virulence in a particular environment. A recent study attempted to associate the genotype of septic *E. coli* strains with clinical outcomes and portal of entry (either urinary or gastrointestinal tract). The authors could associate genes encoding P fimbriae with a urinary portal of entry, but could not associate genotypes with any clinical outcome[53]. Efforts like these provide a more accurate subdivision of pathogenic *E. coli* types and as such might improve risk assessment of *E. coli* clones.

Interaction of *E. coli* with host and microbiome

Although the genetics of an *E. coli* strain influence the chance of its successful colonisation or infection, this also strongly depends on the environment itself. The most obvious environmental influence is the interaction of *E. coli* with its host. It has been long known that host genotypes[54] and bacterial genotypes together influence the risk and outcomes of bacterial disease. Recent studies have performed combined analyses of host and bacterial genotypes[55,56]. Combining these data allows investigating interaction effects, which describe whether a phenotype is associated to a specific combination of host and bacterial genotypes. A microbiome study suggested genetic variation in the human *NOD2* gene might be associated with *E. coli* colonisation[57]. However, specific GWASs combining human and *E. coli* genotypes are currently lacking.

A field receiving much attention in recent years is microbiome research. Phenomena such as colonisation resistance due to nutrient competition have been shown to influence the colonisation of intestinal pathogenic *E. coli* types such as *E. coli* O157:H7[58,59]. As it has become clear that ExPEC clones are efficient intestinal colonisers (Chapter 4), it would be interesting to investigate whether competition for certain nutrients – perhaps sialic acids – influences ExPEC colonisation. Additionally, *E. coli* displays a stable population structure

of broadly defined lineages[60,61], although within these broader lineages replacement is common. This phenomenon might be driven by negative frequency-dependent selection, ensuring a stable population structure[62]. Large sequencing efforts, densely sampling *E. coli* populations will help to inform researchers about population dynamics[63]. It should be noted that clinical and human-derived *E. coli* isolates are still often overrepresented, resulting in an unequal coverage of the full diversity of *E. coli*.

Future therapeutic or preventive interventions

Bacterial WGS analysis also informs prudent use of therapeutic or preventive interventions. First, while designing interventions it is important to know which part of the bacterial population needs to be targeted. Resolving bacterial population structure can inform which bacterial types require most urgent attention. Secondly, to ensure the interventions remain effective against pathogenic bacterial types, continuous surveillance is warranted. Finally, WGS analysis can suggest novel molecular targets for designing interventions.

E. coli vaccines are one of the preventive options that could help curb the spread of resistant *E. coli*. The main component that is studied for *E. coli* is the O antigen, encoding part of lipopolysaccharide (LPS)[64]. This antigen is immunogenic, although it displays considerable diversity in the *E. coli* population and over 180 variants are known[64]. Several *E. coli* vaccines with O antigen components are tested in clinical trials, including a four-valent (ExPEC4v, phase 2[65]), a nine-valent (ExPEC9V, phase 3[66]) and a ten-valent version (ExPEC10V, phase 2[65]). WGS analysis of French and English cohorts of *E. coli* bloodstream infections showed that the ExPEC4V vaccine matched the O antigen of 35-48% of cases, while the ExPEC10V vaccine matched the O antigen of 58-72% of all isolates[60,67,68]. Bacterial WGS analysis can optimise the selection of O antigens for these vaccines. Both the French and the English study advised to include the O allele 17 to cover an additional 5-6% of cases[60,67,68]. Other WGS analyses could inform on O antigens displayed by commensal *E. coli* in the human microbiome, to minimise the disruption of the commensal *E. coli* population. Additionally, continuous surveillance of O antigens displayed by pathogenic and commensal *E. coli* is necessary to assess serotype replacement. Serotypes covered by vaccines typically decrease in prevalence but may be replaced by other serotypes which are not covered by the vaccine. This phenomenon is well described for *S. pneumoniae*[69] and necessitates continuous surveillance of vaccine coverage and possibly the introduction of vaccines with a more comprehensive coverage. Finally, WGS analysis can identify conserved bacterial antigens which might serve as vaccine components. A recent study assessed the conservation of possible vaccine antigens of *S. pyogenes* and listed a number of antigens that are conserved across the whole population of *S. pyogenes*[70]. Additionally, development of antimicrobials such as darobactin[71] or other drugs that for example inhibit plasmid transmission[72] provide avenues for combating the

increase in antimicrobial resistance. Also, for the introduction and continued use of these therapies, continued genomic surveillance is warranted.

The combined promises of bacterial vaccines, novel antibiotics and targeted drugs provide possibilities to curb the increase in multi-drug resistant *E. coli* disease. Aside from scientific progress, there is clear role for genomic research and surveillance to inform novel or existing interventions.

References

1. Kaper JB, Nataro JP, Mobley HLT. Pathogenic *Escherichia coli*. *Nat Rev Microbiol*. 2004;2(2):123-140. doi:10.1038/nrmicro818
2. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *Clin Microbiol Rev*. 2019;32(3). doi:10.1128/CMR.00135-18
3. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol*. 2021;19(1):37-54. doi:10.1038/s41579-020-0416-x
4. Verschuuren TD, van Hout D, Arredondo-Alonso S, *et al*. Comparative genomics of ESBL-producing *Escherichia coli* (ESBL-Ec) reveals a similar distribution of the 10 most prevalent ESBL-Ec clones and ESBL genes among human community faecal and extra-intestinal infection isolates in the Netherlands (2014–17). *J Antimicrob Chemother*. 2021;76(4):901-908. doi:10.1093/jac/dkaa534
5. Nielsen KL, Stegger M, Kiil K, *et al*. Whole-genome comparison of urinary pathogenic *Escherichia coli* and faecal isolates of UTI patients and healthy controls. *Int J Med Microbiol IJMM*. 2017;307(8):497-507. doi:10.1016/j.ijmm.2017.09.007
6. Kantele A, Lääveri T, Mero S, *et al*. Despite Predominance of Uropathogenic/Extraintestinal Pathotypes Among Travel-acquired Extended-spectrum ß-Lactamase–producing *Escherichia coli*, the Most Commonly Associated Clinical Manifestation Is Travelers' Diarrhea. *Clin Infect Dis*. 2020;70(2):210-218. doi:10.1093/cid/ciz182
7. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, *et al*. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J Antimicrob Chemother*. 2008;61(2):273-281. doi:10.1093/jac/dkm464
8. Le Gall T, Clermont O, Gouriou S, *et al*. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains. *Mol Biol Evol*. 2007;24(11):2373-2384. doi:10.1093/molbev/msm172
9. Shaik S, Ranjan A, Tiwari SK, *et al*. Comparative Genomic Analysis of Globally Dominant ST131 Clone with Other Epidemiologically Successful Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages. *mBio*. 8(5):e01596-17. doi:10.1128/mBio.01596-17
10. Nowrouzian FL, Adlerberth I, Wold AE. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes Infect*. 2006;8(3):834-840. doi:10.1016/j.micinf.2005.10.011
11. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. Pathogenicity-Associated Islands in Extraintestinal Pathogenic *Escherichia coli* Are Fitness Elements Involved in Intestinal Colonization. *J Bacteriol*. 2010;192(19):4885-4893. doi:10.1128/JB.00804-10
12. Russell CW, Fleming BA, Jost CA, *et al*. Context-Dependent Requirements for FimH and Other Canonical Virulence Factors in Gut Colonization by Extraintestinal Pathogenic *Escherichia coli*. *Infect Immun*. 2018;86(3):e00746-17. doi:10.1128/IAI.00746-17
13. Virulence Factors - MeSH - NCBI. Accessed December 29, 2021. https://www.ncbi.nlm.nih.gov/mesh/68037521
14. Båga M, Normark S, Hardy J, *et al*. Nucleotide sequence of the *papA* gene encoding the Pap pilus subunit of human uropathogenic *Escherichia coli*. *J Bacteriol*. 1984;157(1):330-333. doi:10.1128/jb.157.1.330-333.1984
15. de Lorenzo V, Bindereif A, Paw BH, Neilands JB. Aerobactin biosynthesis and transport genes of plasmid ColV-K30 in *Escherichia coli* K-12. *J Bacteriol*. 1986;165(2):570-578. doi:10.1128/jb.165.2.570-578.1986

16. Ludden C, Raven KE, Jamrozy D, *et al.* One Health Genomic Surveillance of *Escherichia coli* Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *mBio*. 2019;10(1):e02693-18. doi:10.1128/mBio.02693-18

17. Day MJ, Hopkins KL, Wareham DW, *et al.* Extended-spectrum ß-lactamase-producing *Escherichia coli* in human-derived and foodchain-derived samples from England, Wales, and Scotland: an epidemiological surveillance and typing study. *Lancet Infect Dis*. 2019;19(12):1325-1335. doi:10.1016/S1473-3099(19)30273-7

18. Shen Z, Wang Y, Shen Y, Shen J, Wu C. Early emergence of mcr-1 in *Escherichia coli* from food-producing animals. *Lancet Infect Dis*. 2016;16(3):293. doi:10.1016/S1473-3099(16)00061-X

19. Matamoros S, van Hattem JM, Arcilla MS, *et al.* Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the mcr-1 gene indicates bacterial diversity but plasmid restriction. *Sci Rep*. 2017;7(1):15364. doi:10.1038/s41598-017-15539-7

20. Skurnik D, Clermont O, Guillard T, *et al.* Emergence of Antimicrobial-Resistant *Escherichia coli* of Animal Origin Spreading in Humans. *Mol Biol Evol*. 2016;33(4):898-914. doi:10.1093/molbev/msv280

21. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin Microbiol Infect*. 2018;24(4):350-354. doi:10.1016/j.cmi.2017.12.016

22. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol*. 2019;36(3):587-603. doi:10.1093/molbev/msy242

23. von Elm E, Altman DG, Egger M, *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007;4(10):e296. doi:10.1371/journal.pmed.0040296

24. Field N, Cohen T, Struelens MJ, *et al.* Strengthening the Reporting of Molecular Epidemiology for Infectious Diseases (STROME-ID): an extension of the STROBE statement. *Lancet Infect Dis*. 2014;14(4):341-352. doi:10.1016/S1473-3099(13)70324-4

25. Cheng B, Behr MA, Howden BP, Cohen T, Lee RS. Reporting practices for genomic epidemiology of tuberculosis: a systematic review of the literature using STROME-ID guidelines as a benchmark. *Lancet Microbe*. 2021;2(3):e115-e129. doi:10.1016/S2666-5247(20)30201-9

26. Wayne LG, Brenner DJ, Colwell RR, *et al.* Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *Int J Syst Evol Microbiol*. 1987;37(4):463-464. doi:10.1099/00207713-37-4-463

27. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*. 2007;57(Pt 1):81-91. doi:10.1099/ijs.0.64483-0

28. Gilroy R, Ravi A, Getino M, *et al.* Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. *PeerJ*. 2021;9:e10941. doi:10.7717/peerj.10941

29. Farmer JJ, Fanning GR, Davis BR, *et al. Escherichia* fergusonii and Enterobacter taylorae, two new species of Enterobacteriaceae isolated from clinical specimens. *J Clin Microbiol*. 1985;21(1):77-81. doi:10.1128/jcm.21.1.77-81.1985

30. Huys G, Cnockaert M, Janda JM, Swings J. *Escherichia albertii* sp. nov., a diarrhoeagenic species isolated from stool specimens of Bangladeshi children. *Int J Syst Evol Microbiol*. 2003;53(Pt 3):807-810. doi:10.1099/ijs.0.02475-0

31. Liu S, Jin D, Lan R, *et al. Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. *Int J Syst Evol Microbiol*. 2015;65(7):2130-2134. doi:10.1099/ijs.0.000228

32. Gangiredla J, Mammel MK, Barnaba TJ, *et al.* Draft Genome Sequences of *Escherichia albertii*, *Escherichia fergusonii*, and Strains Belonging to Six Cryptic Lineages of *Escherichia* spp. *Genome Announc*. 2018;6(18):e00271-18. doi:10.1128/genomeA.00271-18

33. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9(1):5114. doi:10.1038/s41467-018-07641-9

34. Murray CS, Gao Y, Wu M. Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nat Commun*. 2021;12(1):4059. doi:10.1038/s41467-021-24128-2

35. Galardini M, Clermont O, Baron A, *et al.* Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus *Escherichia* revealed by a genome-wide association study. *PLOS Genet*. 2020;16(10):e1009065. doi:10.1371/journal.pgen.1009065

36. Pataki BÁ, Matamoros S, van der Putten BCL, *et al.* Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning. *Sci Rep*. 2020;10(1):15026. doi:10.1038/s41598-020-71693-5

37. Lees JA, Croucher NJ, Goldblatt D, *et al.* Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Cobey S, ed. *eLife*. 2017;6:e26255. doi:10.7554/eLife.26255

38. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genomics*. 2020;6(3):e000337. doi:10.1099/mgen.0.000337

39. Falush D. Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol*. 2016;1(5):1-2. doi:10.1038/nmicrobiol.2016.59

40. Hicks AL, Wheeler N, Sánchez-Busó L, Rakeman JL, Harris SR, Grad YH. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLOS Comput Biol*. 2019;15(9):e1007349. doi:10.1371/journal.pcbi.1007349

41. Ma KC, Mortimer TD, Duckett MA, *et al.* Increased power from conditional bacterial genome-wide association identifies macrolide resistance mutations in *Neisseria gonorrhoeae*. *Nat Commun*. 2020;11(1):5374. doi:10.1038/s41467-020-19250-6

42. Lees JA, Mai TT, Galardini M, *et al.* Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *mBio*. 2020;11(4):e01344-20. doi:10.1128/mBio.01344-20

43. O'Leary NA, Wright MW, Brister JR, *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189

44. Alcock BP, Raphenya AR, Lau TTY, *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2020;48(D1):D517-D525. doi:10.1093/nar/gkz935

45. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124. doi:10.12688/wellcomeopenres.14826.1

46. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi:10.1038/sdata.2016.18

47. Jolley KA, Bray JE, Maiden MCJ. A RESTful application programming interface for the PubMLST molecular typing and genome databases. *Database*. 2017;2017:bax060. doi:10.1093/database/bax060

48. Reiser L, Berardini TZ, Li D, *et al.* Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model. *Database*. 2016;2016:baw018. doi:10.1093/database/baw018

49. Kobras CM, Fenton AK, Sheppard SK. Next-generation microbiology: from comparative genomics to gene function. *Genome Biol*. 2021;22(1):123. doi:10.1186/s13059-021-02344-9

50. Puranen S, Pesonen M, Pensar J, *et al.* SuperDCA for genome-wide epistasis analysis. *Microb Genomics*. 2018;4(6):e000184. doi:10.1099/mgen.0.000184

51. Kahya HF, Andrew PW, Yesilkaya H. Deacetylation of sialic acid by esterases potentiates pneumococcal neuraminidase activity for mucin utilization, colonization and virulence. *PLOS Pathog*. 2017;13(3):e1006263. doi:10.1371/journal.ppat.1006263

52. Alteri CJ, Mobley HLT. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. *Curr Opin Microbiol*. 2012;15(1):3-9. doi:10.1016/j.mib.2011.12.004

53. Denamur E, Condamine B, Esposito-Farèse M, *et al.* Genome wide association study of human bacteremia *Escherichia coli* isolates identifies genetic determinants for the portal of entry but not fatal outcome. *medRxiv*. Published online November 11, 2021. doi:10.1101/2021.11.09.21266136

54. Chapman SJ, Hill AVS. Human genetic susceptibility to infectious disease. *Nat Rev Genet*. 2012;13(3):175-188. doi:10.1038/nrg3114

55. Lees JA, Ferwerda B, Kremer PHC, *et al.* Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun*. 2019;10(1):2176. doi:10.1038/s41467-019-09976-3

56. Wang M, Roux F, Bartoli C, *et al.* Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proc Natl Acad Sci U S A*. 2018;115(24):E5440-E5449. doi:10.1073/pnas.1710980115

57. B onder MJ, Kurilshikov A, Tigchelaar EF, *et al.* The effect of host genetics on the gut microbiome. *Nat Genet*. 2016;48(11):1407-1412. doi:10.1038/ng.3663

58. Bäumler AJ, Sperandio V. Interactions between the microbiota and pathogenic bacteria in the gut. *Nature*. 2016;535(7610):85-93. doi:10.1038/nature18849

59. Leatham MP, Banerjee S, Autieri SM, Mercado-Lubo R, Conway T, Cohen PS. Precolonized Human Commensal *Escherichia coli* Strains Serve as a Barrier to *E. coli* O157:H7 Growth in the Streptomycin-Treated Mouse Intestine. *Infect Immun*. 2009;77(7):2876-2886. doi:10.1128/IAI.00059-09

60. Kallonen T, Brodrick HJ, Harris SR, *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;27:1437-1449. doi:10.1101/gr.216606.116

61. Royer G, Darty MM, Clermont O, *et al.* Phylogroup stability contrasts with high within sequence type complex dynamics of *Escherichia coli* bloodstream infection isolates over a 12-year period. *Genome Med*. 2021;13(1):77. doi:10.1186/s13073-021-00892-0

62. McNally A, Kallonen T, Connor C, *et al.* Diversification of Colonization Factors in a Multidrug-Resistant *Escherichia coli* Lineage Evolving under Negative Frequency-Dependent Selection. *mBio*. 2019;10(2). doi:10.1128/mBio.00644-19

63. Gladstone RA, McNally A, Pöntinen AK, *et al.* Emergence and dissemination of antimicrobial resistance in *Escherichia coli* causing bloodstream infections in Norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *Lancet Microbe*. 2021;2(7):e331-e341. doi:10.1016/S2666-5247(21)00031-8

64. Poolman JT, Wacker M. Extraintestinal Pathogenic *Escherichia coli*, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field. *J Infect Dis*. 2016;213(1):6-13. doi:10.1093/infdis/jiv429

65. Janssen Research & Development, LLC. *A Randomized, Observer-Blind, First-in-Human Phase 1/2a Study to Evaluate the Safety, Reactogenicity and Immunogenicity of Three Different Doses of VAC52416 (ExPEC10V) in Adults Aged 60 to 85 Years in Stable Health*. clinicaltrials.gov; 2021. Accessed December 1, 2021. https://clinicaltrials.gov/ct2/show/NCT03819049

66. Janssen Research & Development, LLC. *Randomized, Double-Blind, Placebo-Controlled, Multicenter Phase 3 Study to Assess the Efficacy, Safety And Immunogenicity of Vaccination With ExPEC9V in the Prevention of Invasive Extraintestinal Pathogenic Escherichia coli Disease in Adults Aged 60 Years And Older With a History of Urinary Tract Infection in the Past 2 Years*. clinicaltrials.gov; 2021. Accessed December 1, 2021. https://clinicaltrials.gov/ct2/show/NCT04899336

67. Royer G, Clermont O, Condamine B, *et al*. O-antigen targeted vaccines against *E. coli* may be useful in reducing morbidity, mortality and antimicrobial resistance. *Clin Infect Dis*. 2021;74(2):364-366. doi:10.1093/cid/ciab458

68. Lipworth S, Vihta KD, Chau KK, *et al*. Ten Years of Population-Level Genomic *Escherichia coli* and Klebsiella pneumoniae Serotype Surveillance Informs Vaccine Development for Invasive Infections. *Clin Infect Dis*. 2021;73(12):2276-2282. doi:10.1093/cid/ciab006

69. Croucher NJ, Finkelstein JA, Pelton SI, *et al*. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet*. 2013;45(6):656-663. doi:10.1038/ng.2625

70. Davies MR, McIntyre L, Mutreja A, *et al*. Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nat Genet*. 2019;51(6):1035-1043. doi:10.1038/s41588-019-0417-8

71. Imai Y, Meyer KJ, Iinishi A, *et al*. A new antibiotic selectively kills Gram-negative pathogens. *Nature*. 2019;576(7787):459-464. doi:10.1038/s41586-019-1791-1

72. Buckner MMC, Ciusa ML, Meek RW, *et al*. HIV Drugs Inhibit Transfer of Plasmids Carrying Extended-Spectrum ß-Lactamase and Carbapenemase Genes. *mBio*. 2020;11(1):e03355-19. doi:10.1128/mBio.03355-19

Summary
Nederlandse samenvatting

Summary

Escherichia coli is a Gram-negative opportunistic bacterial pathogen in humans and other animals. *E. coli* can develop resistance against many antimicrobials, including the commonly used third-generation cephalosporins. The genes that often confer this resistance to *E. coli* encode extended-spectrum beta-lactamase (ESBL) enzymes. To combat the increase of resistance in resistant *E. coli*, it is important to understand how *E. coli* and its antimicrobial resistance encoding genes spread between reservoirs. With the increased availability of whole-genome sequencing and genomic analysis, novel insights on *E. coli* biology can be gained. Using genomics, this thesis investigated the population structure of *E. coli* across host species and countries, and which genetic elements are associated with host adaptation and dissemination of ESBL-producing *E. coli*. Additionally, I investigated resources or methodologies necessary to support genomic analysis of antimicrobial resistant *E. coli*.

Chapters 2 and 3 introduce two key concepts used in later chapters: antimicrobial resistance and genomics of *Escherichia* spp. In Chapter 2, colleagues and I summarised mechanisms of ciprofloxacin resistance in *E. coli* through a systematic review of literature. Ciprofloxacin is commonly used for the oral treatment of urinary tract infection in primary care and in outpatients. However, ciprofloxacin resistance is now common in *E. coli*, which can be the result of four genetically encoded resistance mechanisms: alteration of molecular targets, decreased cellular accumulation of ciprofloxacin, physical blocking of molecular targets and enzymatic modification of ciprofloxacin. These four mechanisms can be the result of various genetic events, including single base substitutions, frameshift mutations or acquisitions of genes or gene cassettes. Although our results showed that it is possible for *E. coli* to become clinically resistant against ciprofloxacin without target alteration mutations in *gyrA* and *parC*, this was very rare. Large increases in resistance were not observed without *gyrA* or *parC* mutations.

Chapter 3 describes a novel species in the *Escherichia* genus: *Escherichia ruysiae*. The type strain of this species was isolated from a faecal sample of an international traveller. This novel species does not seem to be pathogenic in humans and is rarely identified in human samples during infection. The species comprises cryptic clades III and IV of the *Escherichia* genus. Together with other published or proposed *Escherichia* species *E. coli*, *E. albertii*, *E. fergusonii*, *E. marmotae* and *E. whittamii*, almost the entire diversity of the *Escherichia* genus is currently covered by named species. Taxonomic entities specific to the *Escherichia* genus such as the cryptic clades are not typically included in databases such as NCBI, as opposed to named species. Assigning an officially recognised name to

these *Escherichia* clades facilitates more accessible and comprehensive analyses of the *Escherichia* genus.

The main questions of this thesis are addressed in the subsequent Chapters 4 and 5. Chapter 4 describes a follow-up of the COMBAT study, which studied the acquisition of Enterobacterales harbouring ESBL genes (ESBL-E) in travellers. A total of 633 out of 2001 Dutch international travellers acquired ESBL-E during travel abroad. Seventeen travellers who harboured these ESBL-E in their faeces for more than 12 months after travel (long-term carriers), were further studied here. All long-term carriers harboured ESBL-positive *E. coli*. Thirty-three travellers who harboured ESBL-positive *E. coli* only for one month after travel (short-term carriers) were matched to these long-term carriers based on age, sex and travel destination. Comparing the *E. coli* isolated from long-term and short-term carriers, we found that long-term carriers were far more likely to be colonised by *E. coli* lineages considered to be extraintestinal pathogenic *E. coli* (ExPEC). While ExPEC are often isolated from infection sites outside of the intestine, these findings indicate ExPEC strains are also efficient colonisers of the human intestine.

In Chapter 5, we compiled a dataset of 1,198 whole-genome sequenced *E. coli* isolates from five host species (human, cattle, chicken, pig and wild boar). Through genome-wide association, we identified *E. coli* genes and lineages that were associated with isolation from each of these host species. The genes associated with human colonisation or infection were mostly observed in ExPEC strains, the same *E. coli* lineages that also seemed adapted to human intestine in Chapter 4. Nine of these genes are involved in the metabolism of sialic acid and are termed the human-associated *nan* locus. Knocking out these genes in an ExPEC strain diminished the strains capacity to use sialic acids as an energy source, in *in vitro* growth experiments. Further experiments should be performed to ascertain the functions of two novel putative sialylesterases which are part of the human-associated *nan* locus. Possibly, these sialylesterases enable deacetylation of sialic acids, allowing ExPEC strains to metabolise more variants of sialic acid. We also identified associations of *E. coli* genes with isolation from other hosts. For example, we identified the previously described association of the salmochelin operon (encoded by the *iroBCDEN* genes) with chicken-associated *E. coli* types (avian pathogenic *E. coli*).

Chapters 6 through 9 address challenges in bacterial genomics. These include the benchmarking of phylogenetic methods, establishing good software engineering principles, generating high-quality genomes and defining virulence factors. The two latter chapters relate to *Streptococcus suis*, an emerging zoonotic swine pathogen capable of causing meningitis and sepsis in humans.

In Chapter 6, we simulated datasets of bacterial genomes *in silico*. These genomes evolved according to a predefined phylogeny, termed the ground-truth phylogeny. Genomes underwent various mutations, including base substitutions, insertions, deletions, horizontal gene transfers, gene duplications and gene loss events. From the evolved genomes, phylogenies were reconstructed using seventeen different workflows and all resulting phylogenies were compared to the ground-truth phylogeny. We scored phylogenetic reconstruction workflows based on how similar their produced phylogenies were to the ground-truth phylogeny. We simulated eight datasets, each varying in rates of mutational events (e.g. more or fewer insertions/deletions, more or less horizontal gene transfer, etc.). We found that newly developed k-mer alignment methods performed on par with reference-based read mapping, currently considered the gold standard. We also observed that *de novo* assembly methods used in the workflows strongly influence accuracy of phylogenetic reconstruction, which was not addressed in previous studies. Our results could guide researchers in selecting phylogenetic reconstruction workflows which fit their datasets and needs.

When writing or maintaining software, implementing software tests can greatly help to prevent errors and facilitate code development. In Chapter 7, we defined a set of seven recommendations for researchers in microbial bioinformatics looking to develop software tests. Good software engineering practices are sometimes lacking in the field of microbial bioinformatics as many researchers receive little or no formal training in software engineering. Through this community effort, we established recommendations for researchers planning to implement software tests. These recommendations are based on our experiences with developing software tests during a hackathon organised prior to the ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines (ASM NGS) in 2020.

Chapters 3, 4, 5 and 6 of this thesis used genomes stored in public databases, in addition to original data generated in the research. Most public genomes are draft genomes, which are easier and cheaper to generate than completely resolved genomes. However, genomic regions that are difficult to assemble are not contiguous in draft genomes, meaning contig breaks are present. These unresolved regions might be found in plasmids, phages or other genetic elements and can encode important biological functions. Therefore, it is crucial to generate complete genomes. In Chapter 8, we used Illumina short read sequencing and Oxford Nanopore Technologies long read sequencing on five *S. suis* strains which represent the recent Dutch *S. suis* population and are widely used in experiments. Using short and long reads, we could fully resolve all five genomes and we have made these available through public databases for reuse by anyone. Making these genomes available can support future research endeavours which require complete and accurate information on the genomic makeup of these strains.

A commonly used concept in bacteriology is the concept of virulence factors. These factors comprise components of a bacterium increasing its capacity to cause disease but are considered not essential for its viability. For *S. suis*, many studies have been published describing novel putative virulence factors. These virulence factors are often considered to be important for zoonotic potential as well. Through a systematic review presented in Chapter 9, *S. suis* virulence factors were identified in the literature. A genomic survey of the presence of these putative virulence factors in pig and human-originating *S. suis* genomes showed that 53 virulence factors are encoded by the *S. suis* core genome. This finding indicates that these genes should not be considered typical virulence factors, as these genes seem to be generally required for the viability of *S. suis*. Another 26 virulence factors were identified more often in human than in pig-originating *S. suis*, suggesting some contribution to zoonotic potential. Three genes (*nisK*, *nisR* and *hhly3*) were particularly strongly enriched in human-originating *S. suis*, even after correction for *S. suis* lineage and country of isolation. These results together show that many published virulence factors might not match the definition of a virulence factor. Additionally, certain virulence factors seem to contribute to zoonotic potential.

Chapter 10 connects the themes of previous chapters and provides directions for future research. Outstanding questions on ExPEC biology and transmission are emphasised. Challenges in bacterial genomic typing and genotype-phenotype analyses are also considered. Finally, this chapter presents future research directions, such as the interplay between *E. coli* and its host and microbiome, potential therapeutic interventions and the importance of bacterial genomics for these endeavours.

Nederlandse samenvatting

Escherichia coli is een Gram-negatieve opportunistische bacteriële ziekteverwekker bij mensen en andere dieren. *E. coli* kan resistentie ontwikkelen tegen veel antibiotica, waaronder de veelgebruikte derde-generatie cefalosporines. De genen die deze resistentie van *E. coli* vaak veroorzaken, coderen voor zogenaamde extended-spectrum bèta-lactamase-enzymen (ESBL). Het is belangrijk om te begrijpen hoe *E. coli* zich verspreidt tussen reservoirs en hoe antimicrobiële resistentie zich verspreidt via *E. coli*. Met de toegenomen beschikbaarheid van genoomsequencing en genomische analyses kunnen nieuwe inzichten in de biologie van *E. coli* worden verkregen. Met behulp van *genomics* heb ik in dit proefschrift de populatiestructuur van *E. coli* in verschillende gastheren en landen onderzocht, en welke genetische elementen ESBL-positieve *E. coli* in staat stellen om zich aan te passen en te verspreiden. Daarnaast heb ik datasets en methoden onderzocht die nodig zijn om genomische analyse van antimicrobieel-resistente *E. coli* te ondersteunen.

Hoofdstukken 2 en 3 introduceren twee sleutelconcepten die in latere hoofdstukken werden gebruikt: antimicrobiële resistentie en *genomics* van *Escherichia* spp. In Hoofdstuk 2 wordt een systematische literatuurstudie beschreven van de mechanismen waardoor *E. coli* resistent kan worden tegen ciprofloxacine. Ciprofloxacine kan dienen als een orale behandelingsoptie voor patiënten met urineweginfectie in de eerste en tweede lijn. Resistentie tegen ciprofloxacine komt echter inmiddels vaak voor bij *E. coli*, wat het gevolg kan zijn van vier resistentiemechanismen: verandering van het moleculaire doelwit van ciprofloxacine, verminderde cellulaire accumulatie van ciprofloxacine, fysieke blokkering van moleculaire doelwit en enzymatische modificatie van ciprofloxacine. Deze vier mechanismen kunnen teweeggebracht worden door verschillende veranderingen in het bacteriële genoom, waaronder substituties, frameshift-mutaties of acquisitie van genen of gencassettes. Hoewel onze resultaten aantonen dat *E. coli* mogelijk resistent wordt tegen ciprofloxacine zonder mutaties in *gyrA* en *parC* (verandering van het moleculaire doelwit), is dit zeer zeldzaam. Grote toenames in resistentie werden niet waargenomen zonder *gyrA*- of *parC*-mutaties.

Hoofdstuk 3 beschrijft een nieuwe soort in het geslacht *Escherichia*: *Escherichia ruysiae*. De typestam van deze soort werd geïsoleerd uit de ontlasting van een internationale reiziger. Deze nieuwe soort lijkt niet pathogeen bij mensen want het wordt zelden geïdentificeerd in diagnostische monsters. De soort omvat *cryptic clades* III en IV van het geslacht *Escherichia*. Samen met andere geaccepteerde of voorgestelde *Escherichia*-soorten *E. coli*, *E. albertii*, *E. fergusonii*, *E. marmotae* en *E. whittamii*, wordt vrijwel de gehele diversiteit van het geslacht *Escherichia* in benoemde soorten omvat. Taxonomische entiteiten die specifiek zijn voor het geslacht *Escherichia*, zoals de *cryptic clades*, worden

doorgaans niet opgenomen in databases zoals NCBI, in tegenstelling tot benoemde soorten. Het vaststellen van wetenschappelijk erkende soorten zorgt dus voor volledigere, toegankelijkere en eenvoudigere analyses van *Escherichia* spp.

De belangrijkste vragen van dit proefschrift worden behandeld in de hoofdstukken 4 en 5. In Hoofdstuk 4 wordt een vervolg van de COMBAT-studie, die de acquisitie van Enterobacterales met ESBL-genen (ESBL-E) in reizigers onderzocht. In totaal acquireerden 633 van de 2001 Nederlandse internationale reizigers ESBL-E tijdens de reis. Alle zeventien reizigers die deze ESBL-E meer dan 12 maanden na hun reis bij zich droegen (langdurende dragers) zijn hier verder bestudeerd. Alle langdurende dragers droegen ESBL-positieve *E. coli*. Drieëndertig reizigers die maximaal één maand na de reis ESBL-positieve *E. coli* hadden (kortdurende dragers) werden gematcht met langdurende dragers op basis van leeftijd, geslacht en reisbestemming. Bij het vergelijken van de *E. coli* geïsoleerd uit de ontlasting van langdurende en kortdurende dragers, zagen we dat langdurende dragers veel meer kans hadden om gekoloniseerd te zijn door *E. coli* types die onder de noemer extra-intestinale pathogene *E. coli* (ExPEC) geschaard worden. Naast dat ExPEC vaak geïsoleerd worden uit infecties van anatomische locaties buiten de darm, laten deze bevindingen zien dat ExPEC ook efficiënt de menselijke darm kunnen koloniseren.

In Hoofdstuk 5 hebben we een dataset van genoomsequenties samengesteld afkomstig van 1.198 *E. coli* isolaten die geïsoleerd zijn uit vijf gastheersoorten (mens, rund, kip, varken en wild zwijn). Door middel van *genome-wide association studies* (GWAS) identificeerden we *E. coli*-genen en -types die geassocieerd zijn met isolatie uit elk van deze gastheersoorten. De genen die geassocieerd zijn met kolonisatie van de humane gastheer werden meestal waargenomen in ExPEC-stammen, dezelfde *E. coli* die ook aangepast leken om de menselijke darm efficiënt te koloniseren beschreven in Hoofdstuk 4. Negen van deze genen dragen bij aan het metabolisme van siaalzuur en dit cluster van genen wordt het humane gastheer-geassocieerde *nan* locus genoemd. Het verwijderen van deze genen uit het genoom van een ExPEC-stam verminderde het vermogen van deze stam om siaalzuren als energiebron te gebruiken in *in vitro* groei-experimenten. Verdere experimenten moeten worden uitgevoerd om de functies van twee nieuwe vermeende sialylesterasen (waarvan de genen deel uitmaken van het nieuwe humane gastheer-geassocieerde *nan* locus) vast te stellen. Mogelijk maken deze sialylesterasen de deacetylatie van siaalzuren mogelijk, waardoor ExPEC meer siaalzuurvarianten kan metaboliseren. We identificeerden ook associaties van *E. coli*-genen met isolatie uit andere gastheersoorten, zoals de eerder beschreven associatie van het salmochelin-operon (gecodeerd door de *iroBCDEN*-genen) met pluimvee-geassocieerde *E. coli*-typen (aviair pathogene *E. coli*).

Hoofdstukken 6 tot en met 9 gaan in op uitdagingen in bacteriële *genomics*. Deze hoodstukken omvatten benchmarking van fylogenetische analysemethoden, het vaststellen van goede software-engineeringprincipes, het genereren van hoogwaardige genoomsequenties en het definiëren van virulentiefactoren. In hoofdstukken 8 en 9 staat *Streptococcus suis* centraal, een opkomende zoönotische varkenspathogeen die meningitis en sepsis bij mensen kan veroorzaken.

In Hoofdstuk 6 hebben we datasets van bacteriële genomen *in silico* gesimuleerd. Deze genomen evolueerden volgens een vooraf gedefinieerde fylogenie, de grondwaarheidsfylogenie genoemd. Genomen ondergingen verschillende mutaties, waaronder substituties, inserties, deleties, horizontale gen-overdracht, genduplicaties en genverlies. Van de geëvolueerde genomen werden fylogenieën gereconstrueerd met behulp van zeventien verschillende workflows en alle resulterende fylogenieën werden vergeleken met de grondwaarheidsfylogenie. We beoordeelden fylogenetische reconstructieworkflows op basis van vergelijkbaarheid van geproduceerde fylogenieën met de grondwaarheidsfylogenie. We simuleerden acht datasets, elk variërend in verhoudingen tussen mutatiegebeurtenissen (bijv. meer of minder inserties/deleties, meer of minder horizontale gen-overdracht, enz.). We ontdekten dat nieuw ontwikkelde k-mer-methodes om DNA sequenties te vergelijken ongeveer even goed presteerden als de huidige gouden standaard (read-mapping methodes die een referentiegenoom gebruiken). We hebben ook waargenomen dat *de novo* assemblagemethoden de nauwkeurigheid van fylogenetische reconstructie sterk kunnen beïnvloeden, wat in eerdere studies niet onderzocht werd. Onze resultaten kunnen onderzoekers helpen bij het selecteren van workflows voor fylogenetische reconstructie die passen bij hun datasets en behoeften.

Bij het schrijven of onderhouden van software kunnen softwaretests helpen om fouten te voorkomen of softwareontwikkeling te vergemakkelijken. In Hoofdstuk 7 hebben we een reeks van zeven aanbevelingen geformuleerd voor onderzoekers in microbiële bio-informatica die softwaretests willen implementeren. In het veld van microbiële bio-informatica ontbreken soms *best practices* op het gebied van software-engineering, aangezien veel onderzoekers weinig of geen formele training in software-engineering hebben gekregen. De aanbevelingen zijn gebaseerd op de ervaringen van een diverse en internationale groep microbiële bio-informatici die deelnamen aan een hackathon voorafgaand aan de ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines (ASM NGS) in 2020.

Hoofdstukken 3, 4, 5 en 6 van dit proefschrift gebruikten genoomsequenties die zijn opgeslagen in openbare databases, naast data die zelf gegenereerd werd voor het onderzoek. De meeste openbare genomen zijn niet helemaal compleet (zogenaamde *draft* genomen), aangezien het genereren van volledige genomen kostbaar is. Genoomregio's die moeilijk te assembleren zijn, zijn niet aaneengesloten in deze *draft*

genomen, wat betekent dat het geassembleerde genoom gefragmenteerd is. De niet-geassembleerde gebieden kunnen worden gevonden in plasmiden, fagen of andere genetische elementen en kunnen coderen voor belangrijke biologische functies. Daarom is het cruciaal om volledige genomen te genereren. In Hoofdstuk 8 gebruikten we Illumina short read sequencing en Oxford Nanopore Technologies long read sequencing op vijf *S. suis*-stammen die de Nederlandse *S. suis*-populatie vertegenwoordigen en die veel worden gebruikt in experimenten. Met behulp van korte en lange reads konden we alle vijf genomen volledig assembleren en we hebben deze beschikbaar gemaakt via openbare databases voor hergebruik vrij van beperkingen. Het beschikbaar maken van deze genomen kan toekomstige onderzoeksinspanningen ondersteunen die volledige en nauwkeurige informatie over de genoomsamenstelling van deze stammen vereisen.

Een veelgebruikt concept in de bacteriologie is het concept van virulentiefactoren. Deze factoren omvatten componenten van een bacterie die het vermogen om ziekten te veroorzaken vergroten, maar niet vereist zijn voor zijn levensvatbaarheid van de bacterie. Voor *S. suis* zijn veel studies gepubliceerd die nieuwe vermeende virulentiefactoren beschrijven. Vaak wordt aangenomen dat deze virulentiefactoren ook belangrijk zijn om infecties bij de mens te veroorzaken. Door middel van een systematische review gepresenteerd in Hoofdstuk 9, werden *S. suis* virulentiefactoren geïdentificeerd in de literatuur. Onderzoek naar de aanwezigheid van deze vermeende virulentiefactoren in *S. suis*-genomen van stammen die uit mens en varken geïsoleerd waren, toonde aan dat 53 virulentiefactoren worden gecodeerd door genen die in meer dan 95% van alle *S. suis* genomen voorkomen (kerngenoom). Deze bevinding geeft aan dat deze genen niet als virulentiefactoren in de strikte zin moeten worden beschouwd, aangezien deze genen over het algemeen nodig lijken te zijn voor de levensvatbaarheid van *S. suis*. Nog eens 26 virulentiefactoren werden vaker geïdentificeerd bij *S. suis* afkomstig van mensen dan bij *S. suis* afkomstig van varkens, wat wijst op een mogelijke bijdrage aan het zoönotisch potentieel. Drie genen (*nisK*, *nisR* en *hhly3*) waren bijzonder sterk verrijkt in *S. suis* geïsoleerd uit mens, zelfs na correctie voor *S. suis*-type en land van isolatie. Deze resultaten tonen samen aan dat veel gepubliceerde virulentiefactoren mogelijk niet voldoen aan de definitie van virulentiefactor. Bovendien lijken bepaalde virulentiefactoren bij te dragen aan zoönotisch potentieel.

In Hoofdstuk 10, ten slotte, worden de thema's van voorgaande hoofdstukken verbonden en worden mogelijke richtingen voor toekomstig onderzoek aangegeven. In het bijzonder worden openstaande vragen over de biologie en transmissie van ExPEC benadrukt. Uitdagingen in bacteriële genoom typering en genotype-fenotype-analyses worden ook aangestipt. Tenslotte worden er in dit hoofdstuk toekomstige onderzoeksrichtingen gepresenteerd, zoals het samenspel tussen *E. coli* en zijn gastheer en microbioom, mogelijke therapeutische interventies en het belang van bacteriële *genomics* voor deze ontwikkelingen.

Appendices

Authors and Affiliations
PhD Portfolio
List of publications
Acknowledgement

Authors and Affiliations

Boas C.L. van der Putten, Kees C.H. van der Ark, Thomas J. Roodsant, Constance Schultsz
Amsterdam Institute for Global Health and Development, Department of Global Health, Amsterdam UMC, University of Amsterdam, the Netherlands
Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, the Netherlands

Rik Oldenkamp
Amsterdam Institute for Global Health and Development, Department of Global Health, Amsterdam UMC, University of Amsterdam, the Netherlands

Jarne M. van Hattem, Niek A.H. Huijsmans, Victoria A. Janes, Sébastien Matamoros, Daniel R. Mende, Sara M. Tamminga
Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, the Netherlands

Edwin R. Scholl
Electron Microscopy Center Amsterdam, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

Martin A. Haagmans
Department of Clinical Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands

Martin Bootsma
UMC Utrecht, Utrecht, Netherlands

John Penders
School for Public Health and Primary Care (Caphri), Department of Medical Microbiology, Maastricht University Medical Centre, Maastricht, the Netherlands
School for Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University Medical Centre, Maastricht, the Netherlands

Giovanni Pasquini, Daniel Remondini
Department of Physics and Astronomy (DIFA), University of Bologna, Bologna, Italy

Torsten Semmler, Sumeet K. Tiwari
Genome sequencing and genomic epidemiology, Robert Koch Institute, Berlin, Germany

Lothar H. Wieler
Robert Koch Institute, Berlin, Germany

Christian Berens, Marta Ferrandis-Vila, Thilo M. Fuchs, Christian Menge
Institute of Molecular Pathogenesis, Friedrich-Loeffler-Institut, Jena, Germany

Angelika Fruth
Enteropathogenic Bacteria and Legionella, Robert Koch Institute, Wernigerode, Germany

Charlotte Huber, Vanessa Johanns
Advanced Light and Electron Microscopy, Robert Koch Institute, Berlin, Germany

Christa Ewers
Institute of Hygiene and Infectious Diseases of Animals, Giessen, Germany

Ivonne Stamm
Vet Med Labor GmbH, Division of IDEXX Laboratories, Ludwigsburg, Germany

Astrid Bethe, Stefan Schwarz
Institute of Microbiology and Epizootics, Freie Universität Berlin, Berlin, Germany
Institute of Microbiology and Epizootics, Berlin, Germany

Herbert Schmidt
Institute of Food Science and Biotechnology, Department of Food Microbiology and Hygiene, University of Hohenheim, Stuttgart, Germany

Trung N. Vinh
Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam

Ngo T. Hoa
Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
Microbiology Department and the Micro-Parasitology Unit of the Center for Bio-Medical Research, Pham Ngoc Thach University of Medicine, Ho Chi Minh City, Vietnam

Amanda Fivian-Hughes, Roberto La Ragione, Joy Leng, Jenny M. Ritchie
Department of Pathology and Infectious Diseases, School of Veterinary Medicine, University of Surrey, Guildford, United Kingdom

Julio Álvarez, Lucas Domínguez
VISAVET Health Surveillance Centre and Department of Animal Health, Faculty of Veterinary Medicine, Complutense University of Madrid, Madrid, Spain

María Ugarte-Ruiz
VISAVET Health Surveillance Centre, Complutense University of Madrid, Madrid, Spain

Rafael Mamede, C. Ines Mendes, Pedro Vila-Cerqueira
Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

Brooke M. Talbot
Department of Biological and Biomedical Sciences, Emory University, Atlanta, GA, USA

Jolinda de Korne-Elenbaas
Department of Infectious Diseases, Public Health Laboratory, Public Health Service of Amsterdam, the Netherlands

Luis Pedro Coelho
Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, China
Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, China

Christopher A. Gulvik
Bacterial Special Pathogens Branch, Division of High-Consequence Pathogens and Pathology, Centers for Disease Control and Prevention, Atlanta, GA, USA

Lee S. Katz
Enteric Diseases Laboratory Branch, Division of Foodborne, Waterborne, and Environmental Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
Center for Food Safety, University of Georgia, Griffin, GA, USA

PhD Portfolio

Name PhD student: Boas C.L. van der Putten
PhD period: 1 september 2017 – 31 august 2021
Names of PhD supervisor(s) & co-supervisor(s): Constance Schultsz (promotor) & Arie van der Ende, Daniel Mende (co-promotores)

1. PhD training

	Year	ECTS*
General courses		
Practical Biostatistics	2018	1.4
Scientific Writing in English	2019	1.5
Specific courses		
Infectious diseases (AMC)	2017	1.4
Unix (AMC)	2018	0.6
McGill Summer Institute Genomic Epidemiology of Infectious Diseases	2018	1.4
Seminars, workshops and master classes		
Weekly seminars for Laboratory of Experimental Bacteriology (participant, organiser since February 2020)	2017-2021	3
HECTOR Consortium Meetings	2017-2021	3
AI&II PhD Retreat	2017	0.6
McGill Summer Institute AMR Special Session	2018	0.3
Long-read sequencing symposium	2018	0.3
Symposium to honour the inauguration of dr. Rob Willems as professor of Population Genetics of Antibiotic Resistance at Utrecht University	2018	0.1
Seminar HPC Data Management	2018	0.1
COMPARE Consortium Meeting (with oral presentation)	2019	0.6
Special Interest Group for Bioinformaticians in Medical Microbiology in the Netherlands meeting (with oral presentation)	2019	0.6
Special Interest Group for Bioinformaticians in Medical Microbiology in the Netherlands meeting (with oral presentation)	2019	0.6
Food Safety hackathon at the Quadram Institute for Biosciences	2019	0.9
APROVE Career Event (AMC)	2019	0.1
ASM NGS pre-conference hackathon (organiser)	2020	0.6
Presentations		
Netherlands Centre for One Health AMR meeting	2018	See below
29th European Congress of Clinical Microbiology & Infectious Diseases (ECCMID)	2019	See below

KNVM Scientific Spring meeting	2019	See below
Special Interest Group for Bioinformaticians in Medical Microbiology in the Netherlands meeting	2019	See below
Special Interest Group for Bioinformaticians in Medical Microbiology in the Netherlands meeting	2019	See below
12th International Meeting on Microbial Epidemiological Markers (IMMEM XII)	2019	See below
AMR – Genomes, Big Data and Emerging Technologies	2020	See below
ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines (ASM NGS)	2020	See below
KNVM Scientific Spring meeting	2021	See below
(Inter)national conferences		
Netherlands Centre for One Health Annual Meeting	2018	0.3
Netherlands Centre for One Health AMR meeting (with oral presentation)	2018	0.1
29th European Congress of Clinical Microbiology & Infectious Diseases (ECCMID) (with oral presentation)	2019	1.1
KNVM Scientific Spring meeting (with poster presentation)	2019	0.6
12th International Meeting on Microbial Epidemiological Markers (IMMEM XII) (with oral presentation)	2019	1.1
AMR – Genomes, Big Data and Emerging Technologies (with lightning talk)	2020	0.9
ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines (ASM NGS) (with poster presentation)	2020	1.4
KNVM Scientific Spring meeting (with oral presentation)	2021	0.6
Other		
Student member of NVAO panel for reaccreditation of Biomedical Sciences programmes in the Netherlands	2017	3.1
Student member of NVAO panel for reaccreditation of Life Sciences and Natural Resources programmes at Wageningen University and Research	2018	2.9
Member of Educational Committee of the AMC Graduate School	2018-2020	1.5
Early Career Microbiologist reviewer for the Microbial Genomics journal	2020-2021	1
Guest editor for Special Issue on Databases for the Microbial Genomics journal	2021	1

*1 ECTS has been awarded per 28 hour workload for activity. ECTS per activity is rounded to one decimal.

2. Teaching

	Year	ECTS*
Lecturing		
Medical Microbiology course, BSc Biomedical Sciences	2020	0.5
Tutoring, Mentoring		
Data Carpentry workshop	2019	0.3
Practicals and Medical Microbiology course, BSc Biomedical Sciences	2020	5.5
Supervising		
Internship Niek Huijsmans, BSc Medisch Onderzoek en Laboratoriumwetenschappen at Avans University of Applied Sciences	2020	4
Internship Niels Boek, MSc Biomedical Sciences at University of Amsterdam	2021	4
Other		

*1 ECTS has been awarded per 28 hour workload for activity. ECTS per activity is rounded to one decimal.

3. Parameters of Esteem

	Year
Grants	
Travel grant Food Safety hackathon at Quadram Institute for Biosciences	2019
Travel grant to attend IMMEM XII from the Netherlands Centre for One Health	2019
Awards and Prizes	

4. Publications

Underlined authors have contributed equally.

	Year
Peer reviewed	
van der Putten, B. C. L., Remondini, D., Pasquini, G., Janes, V. A., Matamoros, S., & Schultsz, C. (2019). Quantifying the contribution of four resistance mechanisms to ciprofloxacin MIC in *Escherichia coli*: a systematic review. *Journal of Antimicrobial Chemotherapy*, 74(2), 298–310.	2019
van Montfort, T., van der Sluis, R., Darcis, G., Beaty, D., Groen, K., Pasternak, A. O., … **Others**. (2019). Dendritic cells potently purge latent HIV-1 beyond TCR-stimulation, activating the PI3K-Akt-mTOR pathway. *EBioMedicine*, 42, 97–108.	2019

Pataki, B. Á., Matamoros, S., **van der Putten, B. C. L.**, Remondini, D., Giampieri, E., Aytan-Aktug, D., … Schultsz, C. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning. *Scientific reports*, 10(1), 1–9.	2020
van der Putten, B. C. L., Roodsant, T. J., Haagmans, M. A., Schultsz, C., & van der Ark, K. C. H. (2020). Five Complete Genome Sequences Spanning the Dutch *Streptococcus suis* Serotype 2 and Serotype 9 Populations. *Microbiology resource announcements*, 9(6), e01439-19.	2020
van der Putten, B. C. L., Matamoros, S., Mende, D. R., Scholl, E. R., Schultsz, C., & Others. (2021). *Escherichia ruysiae* sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. *International Journal of Systematic and Evolutionary Microbiology*, 71(2), 004609.	2021
Coolen, J. P. M., Jamin, C., Savelkoul, P. H. M., Rossen, J. W. A., Wertheim, H. F. L., Matamoros, S. P., … **Others**. (2021). Centre-specific bacterial pathogen typing affects infection-control decision making. *Microbial genomics*, 7(8).	2021
Roodsant, T. J., **van der Putten, B. C. L.**, Tamminga, S. M., Schultsz, C., & van der Ark, K. C. H. (2021). Identification of Streptococcus suis putative zoonotic virulence factors: A systematic review and genomic meta-analysis. *Virulence*, 12(1), 2787–2797.	2021

Other

van der Putten, B. C. L., Huijsmans, N. A. H., Mende, D. R., & Schultsz, C. (2021). Benchmarking topological accuracy of bacterial phylogenomic workflows using *in silico* evolution. *bioRxiv*.	2021
van der Putten, B. C. L., van Hattem, J. M., Penders, J., Mende, D. R., Schultsz, C., COMBAT Consortium, & Others. (2021). Extraintestinal pathogenic *Escherichia coli* (ExPEC) are associated with prolonged carriage of extended-spectrum ß-lactamase-producing *E. coli* acquired during travel. *bioRxiv*.	2021
Tiwari, S. K., **van der Putten B. C. L.**, Fuchs, T. M., Vinh, T. N., Bootsma, M., Oldenkamp, R. & Others (2021). Genome-wide association reveals host-specific genomic traits in *Escherichia coli*. Submitted	2021
van der Putten, B. C. L., Mendes, C. I., Talbot, B. M., de Korne-Elenbaas, J., Mamede, R., Vila-Cerqueira, P. & Others (2021). Software testing in microbial bioinformatics: a call to action. Submitted	2021

List of publications

Van der Putten, B. C. L., Remondini, D., Pasquini, G., Janes, V. A., Matamoros, S., & Schultsz, C. (2019). Quantifying the contribution of four resistance mechanisms to ciprofloxacin MIC in *Escherichia coli*: a systematic review. *Journal of Antimicrobial Chemotherapy*, *74*(2), 298–310.

van Montfort, T., van der Sluis, R., Darcis, G., Beaty, D., Groen, K., Pasternak, A. O., … **Van der Putten, B. C. L.** … Others. (2019). Dendritic cells potently purge latent HIV-1 beyond TCR-stimulation, activating the PI3K-Akt-mTOR pathway. *EBioMedicine*, *42*, 97–108.

Pataki, B. Á., Matamoros, S., **Van der Putten, B. C. L.**, Remondini, D., Giampieri, E., Aytan-Aktug, D., … Schultsz, C. (2020). Understanding and predicting ciprofloxacin minimum inhibitory concentration in *Escherichia coli* with machine learning. *Scientific reports*, *10*(1), 1–9.

Van der Putten, B. C. L., Roodsant, T. J., Haagmans, M. A., Schultsz, C., & van der Ark, K. C. H. (2020). Five Complete Genome Sequences Spanning the Dutch *Streptococcus suis* Serotype 2 and Serotype 9 Populations. *Microbiology resource announcements*, *9*(6), e01439-19.

Van der Putten, B. C. L., Matamoros, S., Mende, D. R., Scholl, E. R., Schultsz, C., & Others. (2021). *Escherichia ruysiae* sp. nov., a novel Gram-stain-negative bacterium, isolated from a faecal sample of an international traveller. *International Journal of Systematic and Evolutionary Microbiology*, *71*(2), 004609.

Coolen, J. P. M., Jamin, C., Savelkoul, P. H. M., Rossen, J. W. A., Wertheim, H. F. L., Matamoros, S. P., … Others **On behalf of SIG Bioinformatics in Medical Microbiology NL Consortium**. (2021). Centre-specific bacterial pathogen typing affects infection-control decision making. *Microbial genomics*, *7*(8).

Roodsant, T. J., **Van der Putten, B. C. L.**, Tamminga, S. M., Schultsz, C., & Van Der Ark, K. C. H. (2021). Identification of *Streptococcus suis* putative zoonotic virulence factors: A systematic review and genomic meta-analysis. *Virulence*, *12*(1), 2787–2797.

Acknowledgments

I have many colleagues to thank for the past years at the Department of Medical Microbiology of the Amsterdam UMC. I joined the department for an internship at the Laboratory of Experimental Virology in 2015, after which I found out I actually liked bacteria even more than I did previously.

Starting off, I would like to thank the supervisors I've had over the years. Sébastien, although you were never officially my co-promotor, you showed me the ropes of bacterial genomics. It was also your presentation on *mcr*-1 genomic epidemiology at a LEV meeting that sparked my interest in joining Constance's group for an internship. Thank you for introducing me to the field and I am happy that you are still at the department.

Arie, thank you for acting as my co-promotor. In hindsight, I think I could have learned a lot more from you as a starting PhD student and I should have asked you for more guidance. Your ideas about my thesis have helped improve this greatly and I am glad I can still consult you from time to time on meningococci in my current function.

Daniel, my second co-promotor, I am very glad you joined the department and that you were willing to join my supervisory team. I really enjoyed our corona-walks to Rum Baba and talking about all kinds of stuff. We discussed a lot of good ideas during our walks (also some bad ones) and these really helped during a time we could not regularly visit the office. You are always welcome at IJburg and I'll come visit you more often to continue our walks.

Constance, I cannot thank you enough that you offered me a chance to join your group not once but twice. I had a great time during my internship and an even better one during my PhD. It is a good thing I cannot rid my brain of your mantra "what is your research question?". I still feel you took a leap of faith hiring a fresh-out-of-university student with next to no bioinformatics experience to do a four-year project on bioinformatics. I am glad you did and hope you are too. Although this chapter is closed now, I hope we will still work together for years to come.

Thanks to dr. Eelco Franz, prof. dr. Benno ter Kuile, dr. Alison Mather, prof. dr. Colin Russell, prof. dr. Nina van Sorge, and prof. dr. Heiman Wertheim for agreeing to join my PhD committee. I look forward to having a great discussion with you.

Also thanks to the many people who I got the chance to work with over the years:
- past and present members of Constance's group (Sébastien, Niels, Victoria, Kees, Thomas, Rik, Coral, Jaime, Mary, Yao and Sneha);
- Alje, Jolinda and Sylvia for our GGD projects together (we are not done yet!) and for supervising Niels together;
- AIGHD colleagues (Nwanneka, Dominique, Wiesje, Nina, Artjom, Sabina, Linde, Jacqueline, Sandra, Alyshia, Sophia and many others);
- Rob, Patricia and Edwin who were crucial in describing *E. ruysiae*;
- The HECTOR consortium, with special thanks to Sumeet and Torsten, but definitely also to Trung, Hoa, Julio, Chris B, Christian M, Marta, Maria, Angelika, Martin, Rik, Jennie, Astrid, Stefan, Rob, Joy, Nicole, Mandy and Thilo;
- The COMBAT consortium, especially Jarne, John and Maris;
- Colleagues from the COMPARE consortium, Daniel, Giovanni and Bálint;
- Bas for teaching me about teaching;
- Martijn for all the cables I have borrowed, new computer screens and for your support to LEB during the move;
- Sandra, Payal, Kim and Leonie for answering my many questions about the lab;
- Current colleagues Robin, Astrid, Sandra, Kim, Jasper, Sara, Marieke, Yvonne, Wendy, Agaath, Claudia, Ilse and Wieke for welcoming me in Nina's group and the NRLBM;
- The members of the education committee of the AMC graduate school, especially Twan, Malon, Laura and Sanne; Yvonne for supporting the education committee so well;
- The Quadram Institute for Biosciences and the Netherlands Centre for One Health for travel grants;
- SURFsara for access to computational resources, without which my PhD would have been unimaginably more challenging;
- The Core Facility Genomics of Amsterdam UMC for the much-needed sequencing support;

Also many thanks to the Slack group of microbial bioinformatics (1500 "microbinfies" and counting). Special thanks to supermicrobinfies Inês, Lee and Chris without whom Chapter 7 would never have come to fruition.

Special thanks to Nwanneka. I am very happy and proud to have been your paranymph and lucky to have worked with you on the PhD community of AIGHD.

Looking back, I would also like to thank my past supervisors: Geneviève and Rob, thank you for offering me my very first internship. I still entertain friends and family with carrot facts. The omnipresence of bioinformatics in my scientific career started at the GSL. Thijs, thank you for supervising me at LEV. I learned an enormous amount of laboratory skills from you. These skills are unfortunately withering away while I'm coding, but still help me every day in understanding the wet lab. Jason and Wendy, thank you so much for my time in London. I learned a great deal from both of you. My internship at St. Mary's convinced me I was fit to do a PhD (although I did choose bacteria over viruses, sorry). I only have two PubMed alerts set up: one of them is to keep me informed on the publications from the Barclay lab.

Looking ahead, I also want to thank Nina for offering me a position in her group. I had a great start at the NRLBM and I am very excited for the future. I hope you are too.

A very special thanks to Thomas and Clara who have helped me get through these four years. We all have our own experience going through a PhD and sharing my time at LEB with you has been wonderful. I am very grateful you were both willing to act as paranymph and hope we will stay in touch when our lives lead us to different places than LEB.

Lieve Daddy en Mirjam, lieve Mammie en Rupert, dank jullie wel voor jullie steun de afgelopen vier jaren. Ik heb van ieder van jullie, op ieders eigen manier, zo veel geleerd. Ik ontkom er niet aan om in dit dankwoord wat clichés mee te nemen: ik kan me geen betere ouders wensen.

Lieve Polle, Lars, Tessel, Kiki en Babs, dank jullie wel dat jullie mijn verhalen over bacteriën vier jaar hebben aangehoord. Hier gaan helaas nog heel veel jaren van komen. Jullie zijn altijd welkom in Amsterdam en hoop dat jullie nog vaak komen logeren. Het is ontzettend bijzonder om jullie op te zien groeien en om te zien hoe ieder hun eigen pad inslaat. Ik ben zó trots op ieder van jullie.

Lieve Parkie en Pam, Moonie en Casijn, ik ben heel blij dat jullie dit moment mee kunnen maken. De grootste kracht achter mijn biologische interesse in mijn jeugd ligt ongetwijfeld bij jou, Pam. De natuurwandelingen op de Bosscheweg en je schilderijen van insecten hebben me zeker weten op een pad gezet wat uiteindelijk tot deze PhD leidde. Ook van jullie heb ik bijzonder veel geleerd. Helaas zijn Oma Riekie, Opa Bertus, Bonma en Bonpa er fysiek niet bij, maar dat zijn ze wel in gedachten.

Lieve Bo, Jacques, Mow, Lien en Babcia, ik ben heel blij om zo met open armen ontvangen te zijn door jullie allemaal. Al hebben we misschien nog wel eens aan elkaar moeten wennen, jullie voelen inmiddels echt als familie voor me. Dat we nog maar veel Poolse Kerstmissen mogen meemaken, puzzels mogen leggen, en nog maar vaak naar Lowlands kunnen gaan.

Uiteindelijk, lieve Nico, een dankwoord voor jou. Het is een cliché (verrast dit je nog?), maar met dit proefschrift was ik zonder jou nooit zo ver gekomen als nu. Je hebt het hele werk van kaft tot kaft nauwkeurig gelezen en verbeterd, al heb ik deze alinea nog geheim weten te houden als het goed is. Nog veel belangrijker dan deze praktische hulp is jouw steun en liefde die ik de afgelopen jaren heb gevoeld. Here's to many more, ik houd van je.