# Re-visiting clonal-complex classifications: a novel machine learning approach for investigating population dynamics of antimicrobial resistance in *Campylobacter jejuni*

Dessislava Veltcheva

St. Peter's College, University of Oxford
Biology Department
A thesis submitted for the degree of Doctor of Philosophy

Supervisors: Professor Michael Bonsall and Professor Martin Maiden

Michaelmas 2023

Funding Bodies:

# Declaration

The work herein is my own except where explicitly stated in relevant chapters and in Contributions. No part of this work has been submitted for any other degree or professional qualification.

Where the work has been published, this has been demonstrated through the integrated thesis form that has been filled in for each chapter. Permission has been sought from co-authors of published papers, to use the work that I produced as part of my thesis.

Dessislava Veltcheva

# Abstract

*Campylobacter jejuni* (*C.jejuni*) is a widespread and highly diverse gram-negative zoonotic bacterium, which is a leading cause of bacterial food poisoning worldwide. Robust systems of variant classification, using Multi-locus Sequence Typing (MLST), remain a vital part of understanding the epidemiologic features of *C. jejuni*. This work re-evaluates clonal-complex classification and identifies inconsistencies in specific variants, such as clonal-complex 353, among modern whole-genome sequenced isolates. In addition, this thesis introduces a novel approach to MLST classification called "NeighbourGroups", a generalisable machine learning approach that reproducibly, robustly, and rapidly classifies 7 housekeeping MLST loci onto core genome level groupings. This work has determined that for *C.jejuni*, we need 20 variant groups; we undertook a comparative analysis of each isolate between the two classification: clonal-complex and NeighbourGroups. This work identified sub-variant grouping among CC21 and CC353 and revealed novel dynamics of antimicrobial resistant among *C.jejuni*. Furthermore, potential new housekeeping genes have been identified through this study indicating that this same approach could be applied to other pathogens to identify one locus that can demonstrate cgMLST level NeighbourGroups. This approach could be then applied across pathogens to further identify potential AMR targeted locus that we have not currently sought for. Moreover, once we have locus per pathogen we could even identify if we do have a housekeeping gene across pathogens that could be used for speciation. Furthermore, current clonal-complex schemes have been re-evaluated to show if we want to keep using it which one can stay the same, which one should be

changed, and which ones could be retired as a clonal-complex due to lack of isolate counts.

In summary, this thesis has both re-evaluated the gold standard of molecular epidemiology, identified issues that shown it cannot be resolved in clonal-complex level and created a solution through the establishment of NeighbourGroups which will both support the newly whole genome sequenced isolates and past historical isolates that we only have MLST information. It is hoped that the methods described will be applied beyond *C.jejuni* and contribute positively to epidemiological research.

# Statement of Authorship:

This is an integrated thesis where chapters 2 and 3(pre-print) are published work involving other co-authors. Each chapter includes a section describing co-authors' contributions, and I am the primary author of the published work. Chapters 4 and 5 are currently being edited to be submitted to journals in the future, but I am the first author for both as well. Chapter 6 and 7 are written for paper format but it has not been prepared for submission to a specific journal at this stage.

Chapter 2: Veltcheva, D., Colles, F.M., Varga, M., Maiden, M.C. and Bonsall, M.B., 2022. Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018]. *Microbial Genomics*, *8*(9), p.000875.

Chapter 3: Veltcheva, D., Richer, S., Sheppard, S., Varga, M., Colles, F., Bonsall, M. and Maiden, M., 2023. NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters.

Chapters 4 and 5 are written in a format to be sent to Microbial Genomics https://www.microbiologyresearch.org/content/journal/mgen

Chapters 6 and 7 are written for publication format but not yet consolidated on which journal they will be submitted.

Chapters 1(introduction) and 8(Discussion) are two chapters written in a non-publication format for this thesis.

During this PhD study, I have co-authored a paper on another paper for Covid-19's Exit strategies:
Rawson, T., Brewer, T., Veltcheva, D., Huntingford, C. and Bonsall, M.B., 2020. How and when to end the COVID-19 lockdown: an optimization approach. *Frontiers in Public Health*, *8*, p.262.

# Acknowledgements

I want to thank the people who helped me in my PhD journey.

First, I would like to thank my supervisor, Professor Michael Bonsall, who allowed me to start this journey in 2019 and has been open for me to go in the direction I wanted to develop. Thank you very much for your time and guidance through the COVID-19 pandemic.
I would also like to thank Professor Martin Maiden, who has believed in my NeighbourGroups work and helped me believe in my ability. Thank you very much for your time. I valued our epidemiological conversations, and learning from the founding father of MLST was truly an honour. I want to thank the Maiden Lab for providing a lovely group meeting space, and I enjoyed being part of Maiden Lab. Thank you to Fran, Keith and James, who have been curating the databases and providing the PubMLST database.

Next, I thank my funding bodies, the Foods Standards Agency and Interdisciplinary Bioscience DTP. I am very thankful to the Foods Standards Agency for letting me work on a project on *Campylobacter. I* hope that NeighbourGroups and trend analysis will be helpful in understanding the dynamics of *Campylobacter* further. I would like to thank the DTP, who have been incredible from my first year and have always been welcoming. Thank you very much for your funding and for letting me enhance my skills through training and conferences.

I would like to thank the people who have been there for me, guiding me in my PhD journey. I thank my incredible mentors, Dr. Anusha Panjwani and Dr. Margaret Verga. Anusha, I would like to thank you for all your incredible guidance across my PhD journey! You are a true inspiration, and I am very honoured to have you as my mentor. Margaret, you have always believed in me and have inspired me to go towards data visualisation! Thank you so much for always being so kind and for giving me the opportunity to develop in Data Science further.

I would like to thank a couple of friends who were there on this 5-year journey. Thank you, Seb, for being my MERG buddy! I am very happy that we got to develop as young researchers together. I would like to thank the upper years in the Zoology (Biology) department who have helped me navigate my way. Shout out to Natalia, who has helped me learn the beauty of AMR! A big thank you to Sunali, who has taught me the beauty of statistics! DIvjot for her passion in AMR but for her beautiful inside-out spirit. Big thank you to the friendship that blossomed with Alex and Kristopher for starting PhD cohort together and finishing together. I would also like to thank Delali for constant support, even not in the same city! A big thank you to Meg, who has been a sunshine! Juliana, for our too-long recordings. Vikki for her wittiness and her laughter. There were more people that are not on this list thank you so much for your supports!

Next, I would like to thank my family. First, I would love to thank my mom and dad, who have always been there for me even when we have 9 9-hour time zone difference. Thank you for believing in me and for all your support. I would also like to thank my grandma, who has always been my cheerleader благодаря ти много бабо! I would like to thank my brother and Naoko for their kindness and their support.

Finally, I would like to thank Stephen! I am so happy that we were able to work together on the NeighbourGroups paper to join our common interest and build a classification. Thank you so much for being there for me. I appreciate you a lot, and I am looking forward to both of us becoming Doctors!

# Acronyms

| Terminology | Definition |
|---|---|
| AMR | Antimicrobial Resistance |
| *aspA* | Aspartase A |
| BIGSdb | Bacterial Isolate Genome Sequence Database |
| Bp | Base pair |
| *C.coli* | *Campylobacter coli* |
| *C.jejuni* | *Campylobacter jejuni* |
| CC | Clonal-complex |
| cfu | Colony forming unit |
| cgMLST | Core genome Multi-locus sequence typing |
| FSA | Foods Standards Agency |
| *glnA* | Glutamine synthetase |
| *gltA* | Citrate synthase |
| *glyA* | Serine hydroxy methyltransferase |
| *gyrA* | Gyrase A |
| HGT | Horizontal genetic transfer |
| ML | Machine Learning |
| MLST | Multi-locus sequence typing |
| NLP | Natural Language Processing |
| PCR | Polymerase Chain Reaction |
| *Pgm* | Phosphor glucomutase |
| PubMLST | Public database for molecular typing and microbial genome diversity |
| rMLST | Ribosomal Multi-locus Sequence Typing |
| ST | Sequence type |
| *TetO* | Tetracycline operator |
| *Tkt* | Transketolase |
| UKHSA | The UK Health Security Agency |
| *uncA* | ATP synthase alpha subunit |

# Table of Contents

# List of Figures

# List of Tables

# Introduction:

This introduction will cover the essential topics, including Introduction to *Campylobacter jejuni,* bacterial classification system, antimicrobial resistance, and current epidemiological studies. Then, we will illustrate the existing knowledge gap in the field and demonstrate how different chapters will focus on filling these gaps. Furthermore, a small NLP project was conducted to illustrate the trends in literature across time.

## ***Campylobacter*** **Genus**

### History of *Campylobacter* Discovery

*Campylobacter* was first discovered in 1913 by McFaydean and Stockman, who tried to understand the microorganisms causing abortion in sheep and cattle (Lorenzo *et al.*, 2018). The organism did not have an official name until the 1919 when Smith and Taylor could isolate the same organism from bovine fetal fluid, which led to the name *Vibrio fetus* (Theobald Smith and Taylor, 1918). In 1973, Chatelain and Veron proposed a name change to the genus *Campylobacter*, leading to the change from *V.fetus* to *Campylobacter fetus* (VÉRON and CHATELAIN, 1973). In the 20th century, *C. fetus* was identified as a causative agent of human bloodstream infections. With the invention of bacterial isolation in the 1980s, the relevance of *Campylobacter* to human health was discovered (Guerrant *et al.*, 1978).

## *Campylobacter* species

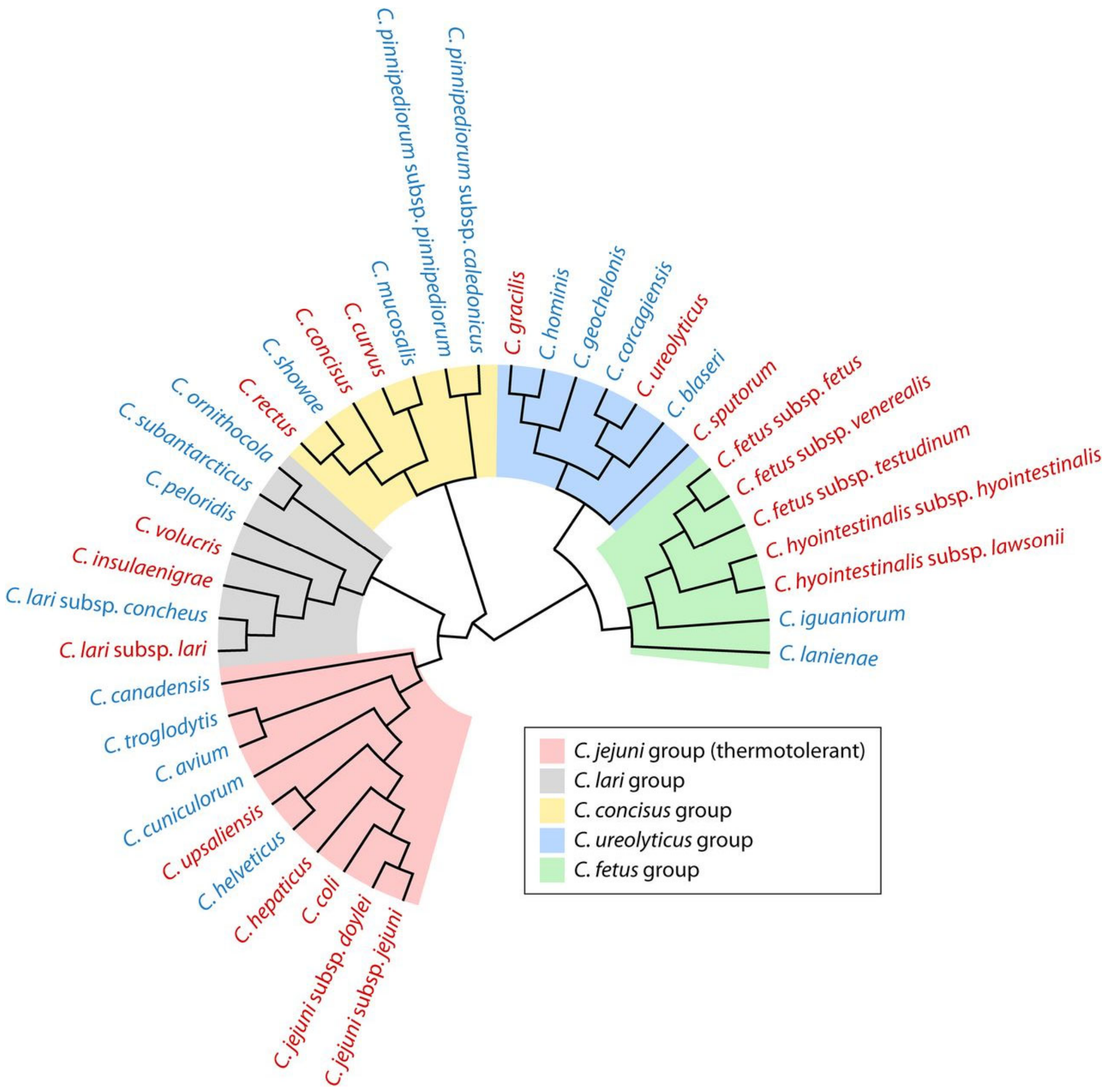


*Figure 1 Campylobacter's species phylogenetic relationships shown in 5 sub-groups (Costa and Iraola, 2019)*

There are 32 officially described *Campylobacter* genera and nine further subspecies (Costa and Iraola, 2019). These species can be divided into five sub-groups, but some species lack sufficient genomic information to understand the emergence and expansion of these variants (Costa and Iraola, 2019) (Figure 1). *C. jejuni* and *C.coli* are known to have undergone gene exchange, with around 15% of their nucleotides diverging (Sheppard *et al.*, 2011). Studies demonstrated that 86.5% of the housekeeping genes are shared between *Campylobacter coli* and *Campylobacter jejuni* (Dieckmann *et al.*, 2021). *C.coli* clusters are divided into three clades. Where clade 1 has genome-wide introgression with *C.jejuni*, whereas clades 2 and 3 only have less than 2% of its genomes to be ancestral to *C.jejuni* (Skarp-de Haan *et al.*, 2014). Some evidence shows that within housekeeping MLST loci, there is genetic exchange. However, its rate is low, and only one of the loci, *flaA*, have an evidence for imported DNA (Dingle *et al.*, 2005). Clade 1, with CC828 being the highest clonal-complex, is associated with clinical isolates that lead to acute diarrhoea. Clades 2 and 3 are related to environmental sources, such as ecological waters (Emele *et al.*, 2019).

## Bacterial Classification

### MLST, rMLST, cgMLST Schemes

Before the development of whole genome sequencing, Carl Woese and his team invented a genealogical classification method which used a small subunit of 16S ribosomal RNA sequence to distinguish taxonomic relations (Woese, Kandler and Wheelis, 1990). However, to establish a typing scheme that requires a level of resolution that can answer epidemiological questions, it needed a gene-by-gene approach rather than serotyping. However, due to the horizontal genetic transfer rate for *Campylobacter*, a single locus was not enough. Thus, Multi-locus sequence typing (MLST) that can robustly characterize bacterial isolates at the molecular level was invented in 1998 (Maiden *et al.*, 1998b). For *Campylobacter jejuni*, there had been no universally accepted typing schemes, and by creating MLST, it has managed to divide *C. jejuni* and *C.coli* through the classification method (Dingle *et al.*, 2005). MLST uses alleles rather than sequences to allow allele-based comparison between the isolates. The combination of these 7 MLST alleles is assigned to a unique sequence type (ST). Thus, one SNP or a recombination event is not distinguishable (Didelot and Maiden, 2010). Therefore, one of the weaknesses is that we cannot distinguish which changes are recent or are from recombination.

There are three major *Campylobacter* typing schemes. These include MLST, rMLST and cgMLST (Urwin and Maiden, 2003; Jolley *et al.*, 2012a) (Figure 2). MLST aims to portably and robustly assign bacterial isolates to the molecular level variants (Maiden, 2006a). MLST databases have allowed us to directly compare bacterial isolates without physically exchanging isolates, and providing a common language for bacterial typing worldwide (Urwin and Maiden, 2003). The MLST's seven loci are reduced gene fragments ranging between 350 to 600 base pairs (Maiden, 2006b). MLST loci encode protein sequences that are conserved in metabolic function, and their nonsynonymous to synonymous substitution denoted as *dN/dS,* has been investigated to show sufficiency in the target populations (Maiden *et al.*, 1998a).

The seven loci assigned for *Campylobacter* are aspartase A (*aspA*), glutamine synthetase (*glnA*), citrate synthase (*gltA*), serine hydroxymethyltransferase (*glyA*), phosphoglucomutase (*pgm*), transketolase (*tkt*) and ATP synthase α subunit (*uncA)* (Dingle *et al.*, 2001). The minimum distances between these 7 MLST loci are 70kb and thus are thought to be very unlikely to have coinherited the same recombination event (Dingle *et al.*, 2001). These loci evolve slowly as they are under stabilizing selection for conserving metabolic function (Dingle *et al.*, 2002). *Campylobacter jejun*i is described as weakly clonal bacterial populations comprising several clonal-complexes, or so-called "lineages", associated with their common progenitor (Maiden, 2006a). A clonal-complex is assigned central genotypes and its related genotypes grouped into one (Maiden, 2006a). They are bacteria with some clonal element in their population and are thought to be closely related but not identical isolates originating from a common progenitor.

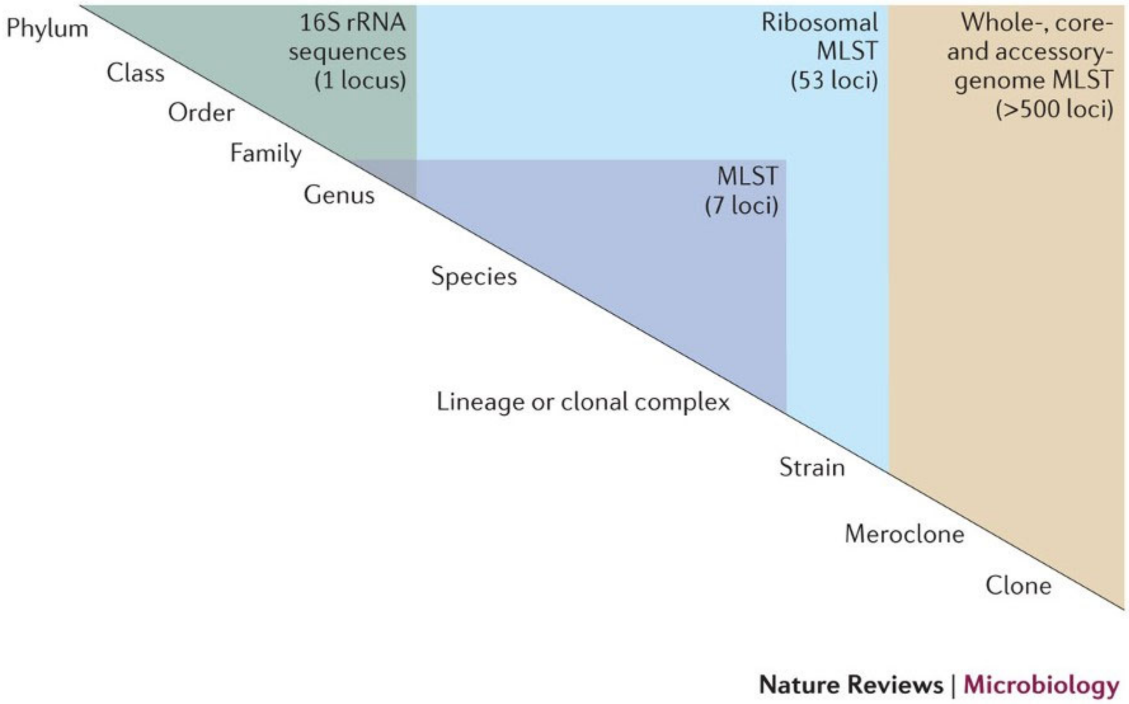


*Figure 2 16S rRNA, MLST, rMLST and core whole genome MLST typing scheme illustrated (Maiden et al., 2013).*

A ribosomal MLST (rMLST) uses variation in 53 genes that encode bacterial ribosomal protein subunits to identify microbial genealogy (Jolley *et al.*, 2012b). A core genome MLST (cgMLST) was created by Maiden Lab to enable the identification of very closely related isolates, which are helpful to identify variants causing outbreaks (Cody *et al.*, 2017). This core genome scheme uses 1,343 loci, a comprehensive set which are present in at least 95% of *Campylobacter* genomes (Cody *et al.*, 2017). The central genotypes were originally chosen informally based on their abundance and whether they were situated in a central position through numerous techniques (Maiden, 2006a).

## Evolution of Sequencing Technologies and Database

Sequencing technologies have evolved rapidly since the first generation of sequencing. Initially Sanger sequencing was developed by Fred Sanger, who developed a method where radiolabelled partial-digestion fragments were used (Sanger *et al* 1977). Next, we have the next-generation sequencing (NGS) emerging between 2004 to 2006, where faster and more affordable methods were developed. This method did not require any bacterial cloning of the DNA fragments and electrophoretic separation of sequencing products, which was very laborious (Dijk *et al.*, 2018). The well-known examples of these approaches are from Illumina and Ion Torrent platforms, where short-read DNA fragments are used for parallel sequencing of where they align to assemble the whole DNA (Hu *et al.*, 2021). One of the drawbacks of next-generation sequencing technologies is their short reads. To improve this drawback, third-generation sequencing (TGS) emerged, which distinguishes itself from NGS by the fact that it enables single molecule sequencing (SMS) in real-time (Van Dijk *et al.*, 2018). A successful example of this method was invented by Pacific Biosciences (PacBio), which called its technology Single-molecule real-time (SMRT). In the 2014, Nanopore sequencing emerged from Oxford Nanopore Technologies with their portable MinION that can sequence direct real-time analysis of long DNA and RNA sequences (Jain *et al.*, 2015). Nanopore sequencing uses electrical currents where the DNA goes through the nanopore and reads the DNA sequencing using base-calling machine learning models.

In *Campylobacter* surveillance, the majority of studies so far have been using NGS technologies, mainly Illumina sequencing. ( Maiden *et al*, 2018; Pascoe *et al.*, 2020). Recent studies are trying to pioneer the usage of Oxford Nanopore technologies for *C. jejuni*. These methods will enhance long-read sequences that are native DNA and will enable on-site screening in the food production leading to quicker food recalls and preventing food poisoning. (Neal-McKinney *et al.*, 2021; Marin *et al.*, 2022).

The MLST database is currently available under the Bacterial Isolate Genome Sequence Database(BigSdb) (Jolley and Maiden, 2010). The database contains metadata, contiguous sequences and allelic data (Jolley and Maiden, 2010). PubMLST is a collection of open-access databases integrating population sequence with provenance and phenotype information. There are over 130 microbial species, 36,469,054 alleles, 1,337,796 isolates and 1,059,797 genomes available in the database that is freely available (accessed on 3 October 2023) (Jolley and Maiden, 2010).

## *Campylobacter* Epidemiology

### Campylobacteriosis and its health burden

*Campylobacter* is a gram-negative corkscrew-shaped bacteria. It's size can range between 0.5 to 5 microns in length with 0.2 to 0.9 microns in width (Cheng and Fischer, 2022). Campylobacteriosis food poisoning is caused by *C.jejuni* and *C.coli* mainly (Facciolà *et al.*, 2017). Around 90% of human Campylobacteriosis is caused by *Campylobacter jejuni,* and the second highest is *Campylobacter coli,* with some other sub-species also causing illness (*Information for Health Professionals* | *Campylobacter* | *CDC*, 2023). Globally, 400 to 500 million people are infected annually with pathogenic *Campylobacter* (Igwaran and Okoh, 2019). In the UK, Campylobacteriosis burdens UK healthcare with around £712.6 million in 2018 with approximately 299,392 cases noted (Daniel *et al.*, 2020). Campylobacteriosis is usually self-limited, but the symptoms can last 5 to 7 days (Fitzgerald, 2015). Symptoms include diarrhoea, cramps, acute abdominal pains, fever, and vomiting (Fitzgerald, 2015). Poultry is known to be the highest-ranked leading pathogen-food combination to cause food poisoning (Hakeem and Lu, 2021). *Campylobacter* is microaerophilic, allowing it to grow in an environment where $CO_2$ and $O_2$ are between 5 to 10 % and temperatures ranging between 37 and 42 degrees celsius, which is suitable for the avian hosts (Williams, Fonseca and Humphrey, 2016a; Lorenzo *et al.*, 2018)

Many Campylobacteriosis cases are sporadic but can also come from point source outbreaks (Montgomery *et al.*, 2018; Kenyon *et al.*, 2020). People can get the infection through three main route causes. Firstly, poor hygiene practices in the kitchen, such as washing the chicken, can cause droplets with *Campylobacter jejuni* to contaminate other food sources and the kitchen surface (Lai *et al.*, 2021). Secondly, not cooking the chicken appropriately will lead to ingestion of live *Campylobacter jejuni,* leading to food poisoning (Facciolà *et al.*, 2017). Thirdly, cross-contamination, such as from packaging or having raw chicken and vegetables cut on the same chopping board will lead to cross-contamination (Cardoso *et al.*, 2021; Santos-Ferreira *et al.*, 2021).

## Source Attribution

A common mis-conception is that some supermarkets have more “cleaner” or “pathogen free” chickens. However this is far from the truth, where around 73% of chickens in the UK were positive for *Campylobacter*, with even the packaging containing around 7% of, leading to potential cross-contamination (Jorgensen *et al.*, 2018). These organisms can reside in the intestines of wild and domesticated animals such as poultry, cattle, sheep, and pigs and wild animals such as geese and birds (Kittl *et al.*, 2013; Food Standards Agency, 2015; Williams, Fonseca and Humphrey, 2016a; Thépault *et al.*, 2017). Currently, poultry is the main reservoir for *C.jejuni* and *C.coli* in the UK (Humphrey *et al*, 2016b; Cody *et al.*, 2019; Goddard *et al.*, 2022). Initially, *Campylobacter* was considered commensal in the chicken microbiota (Elviss *et al.*, 2009). However, recent studies have shown that *Campylobacter* could harm poultry's health and welfare (Sheppard *et al.*, 2013; Humphrey *et al.*, 2014). It can be found in different parts of the poultry gut but is mainly isolated from the large intestine and caeca ( Humphrey *et al*, 2016a). A survey done by the FSA has shown that in the period between 2015 to 2016, 78% of the chickens in the retail sale were positive for *Campylobacter* (above colonies forming unit > 1000 cfu per g on chicken skin (Jorgensen *et al.*, 2019). Thus, there is a high chance that the chicken we buy at the store is already contaminated.

## Guillain–Barré Syndrome and other complications

Although a lot of the cases of Campylobacteriosis are self-limiting, it can also cause risks of post-infectious complications. These include Guillain–Barré Syndrome (GBS), Reactive arthritis (REA), Milner Fisher Syndrome, Ulcerative colitis (UC), and irritable bowel syndrome (IBS) (Facciolà *et al.*, 2017). *C. jejun*i is known to be the most common cause of GBS, leading to 30% of patients being paralysed (Yuki and Hartung, 2012). It is known that 20% of patients will have life-long disability, and 5% will die (Hughes *et al.*, 2007). The cases of GBS are around 0.4 to 4 cases per 100,000 people annually (Black *et al*, 2010).

## Antimicrobial resistance

### Antimicrobial resistance health burden

Another major issue in the health sector is antimicrobial resistance. The O'Neill Report 2015 stated that drug-resistant infections could cause the deaths of 10 million people by 2050 (O'Neill, 2014). It could cost the global economy $100 trillion if antimicrobial resistance is not addressed in the coming years (O'Neill, 2014). However, with antimicrobials being misused in farm and clinical settings and new antibiotics not being invented, bacteria are becoming more resistant to antimicrobials. (Piddock, 2012; Rello, Bunsow and Perez, 2016). Bacteria's evolution is rapid, and to avoid multi-drug resistant bacteria, we need to slow down the spread of resistance or establish new antimicrobials.

### PubMed "*Campylobacter jejuni*" trends in publications

Furthermore, trends in PubMed publications were investigated with the keyword "*Campylobacter jejuni*" using API calling (Figure 3). PubMed contains MeSH (Medical Subject Headings), categorizing all the PubMed papers. We can see that the MeSH term “drug resistance” is increasing in publications overall. The other high occurrence is from “anti-bacterial agents”, which seems to be decreasing slightly. On the other hand, papers on the topic of “poultry disease” had few publications in the 1980s, but it increased towards 2005. The query was only for *C.jejuni*, but the. The “poultry disease” number overall is declining from the 2005 mark, but it is still being published. Potentially the trend could keep its increasing trend if *C. coli* and *C.hepaticus* are included. Furthermore, “food contamination” also sharply increased from 1995 to 2005 but seems to decrease after 2005 rapidly. The decrease could be because we have more papers focusing on clinical cases and in farm settings rather than focusing specifically on food contamination.

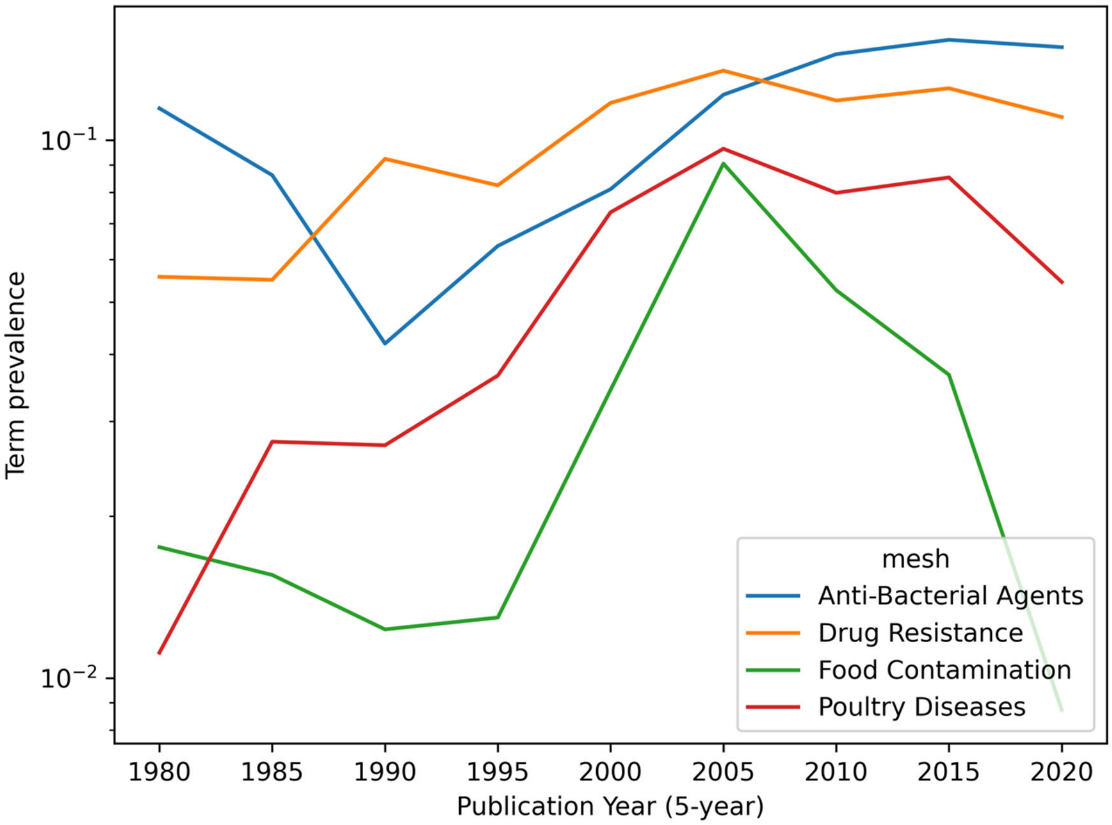
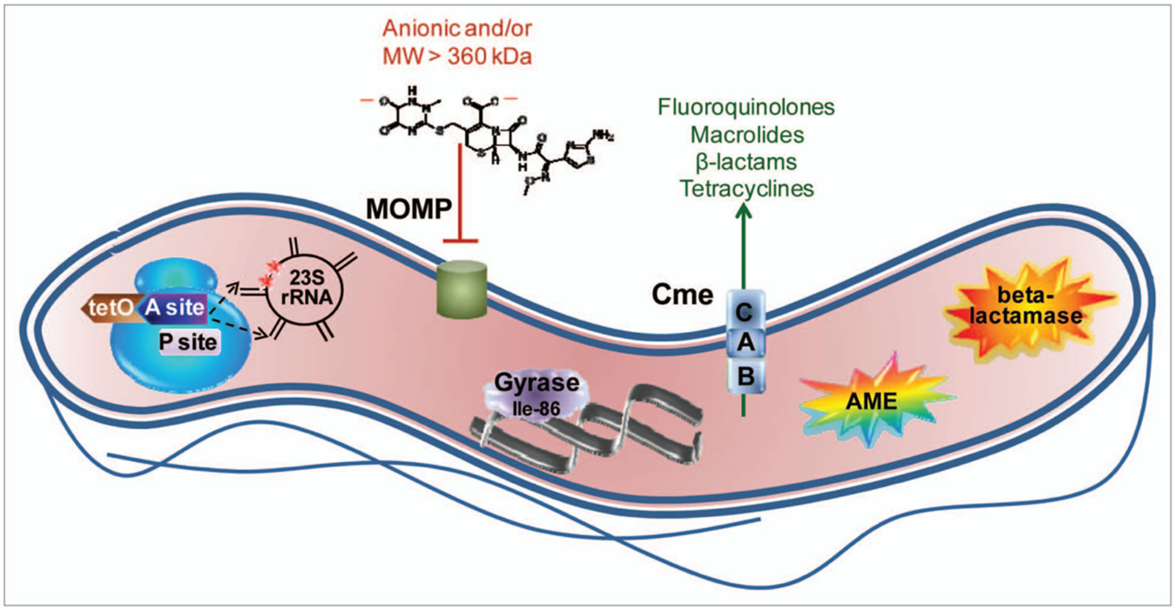


*Figure 3 PubMed MeSH key terms and its abundance across time when "Campylobacter jejuni" is being searched (Figure: created by the author)*

### Antimicrobial resistance mechanism

*C. jejuni* has been recognized by the World Health Organization as one of the twelve priority microorganisms displaying a concerning rise in fluoroquinolone resistance, presenting a significant challenge to public health (Whelan *et al.*, 2019). This fluoroquinolone resistance phenomenon is attributed to a specific point mutation localized within the Quinolone Resistance Determining Region (QRDR) of the *gyrA* gene, a regulator of DNA supercoiling (Han *et al.*, 2012; Lovine *et a*l, 2013). The C257T mutation leads to an amino acid substitution at position T86I within the protein sequence, strongly correlating with heightened fluoroquinolone resistance (Elhadidy *et al.*, 2020; Haldenby *et al.*, 2020). This point mutation in *gyrA* is known to cause high-level resistance to nalidixic acid and fluoroquinolone as phenotype (Espinoza *et al.*, 2020). Furthermore, it is known that putative efflux pump is also associated with fluoroquinolone resistance.



| Antibiotic class | Resistance mechanisms |
|---|---|
| Aminoglycoside | Modification of the antibiotic by aminoglycoside-modifying enzymes (AphA, AadE, Sat)<br>Contribution of efflux is not clear |
| Beta-Lactam | Enzymatic inactivation of the antibiotic by β-lactamase (penicillinase, OXA-61)<br>Decreased membrane permeability of most anionic and MW > 360 kDa antibiotics due to MOMP<br>Efflux through CmeABC and possibly others |
| Fluoroquinolone | Modification of the DNA gyrase target (Thr-86-Ile; also Asp-90-Asn, Ala-70-Thr)<br>Efflux through CmeABC |
| Macrolide | Mutations in 23S rRNA<br>Contribution of mutations in ribosomal proteins L4/L22 is likely minor<br>Efflux through CmeABC and possibly others<br>Decreased membrane permeability due to MOMP |
| Tetracycline | Modification of the target ribosomal A site by TetO binding<br>Efflux through CmeABC and possibly others<br>Contribution of decreased membrane permeability due to MOMP is not clear |

*Figure 4 Antimicrobial resistant mechanisms for Campylobacter (Nicole M Iovine, 2013)*

Intrinsic resistance can be found for novobiocin, bacitracin, vancomycin, and polymyxin/colistin (Nicole M Iovine, 2013). Tetracycline resistance, on the other hand, is usually transferred through a ribosomal protection protein called *tetO,* which is then transmitted in a plasmid-encoded gene (Chopra and Roberts, 2001). Macrolide resistance is caused by a point mutation in the peptidyl encoding region in the domain V of the 23S rRNA gene, leading to a high resistance. For *Campylobacter*, there are three copies of these 23S rRNA (Nicole M Iovine, 2013). Thus, all three must have the point mutation to lead to a macrolide-resistant isolate. The other two low levels of resistance to *Campylobacter* are beta-lactam and aminoglycoside resistance (Figure 4).

## Fitness Benefit of Fluoroquinolone Resistance and One Health

Although Campylobacteriosis is usually self-limiting and antibiotics are unnecessary, if the patients are young, elderly or immune-compromised, macrolides are usually prescribed (Bolinger and Kathariou, 2017). A significant challenge arises from the clinical difficulty distinguishing gastroenteritis caused by *C. jejuni* from illnesses or stemming from other bacterial pathogens. In such cases, fluoroquinolones are often employed as the treatment of choice (Nicole M. Iovine, 2013; Yang *et al.*, 2019). Recent studies indicate that fluoroquinolone-resistant isolates of *C. jejuni* may possess an enhanced virulence potential (Abraham *et al.*, 2020b; Inglis, Taboada and Boras, 2021). Moreover, these resistant variants have demonstrated the ability to form viable biofilms under aerobic conditions, aiding their invasion of epithelial cells (Whelan *et al.*, 2019). A study has indicated that difloxacin or enrofloxacin-treated flocks have shown that 100% of the *Campylobacter* were resistant to both of them, and a high number of them were resistant four weeks after the treatment (Humphrey *et al.*, 2005). It has been shown that fluoroquinolone-resistant resistant *C. jejuni* has enhanced *in vivo* fitness even in the absence of antibiotic selection pressure (Luo *et al.*, 2005). Demonstrating that *C.jejuni*, which is resistant to fluoroquinolones, could have fitness benefit to survival rather than causing fitness cost.

A further study that we must understand is the intricate interplay among enteric pathogens, the animals' environmental context, and their gut microbiome. Furthermore the extension of food chain dynamics and the stages and the complex network "One Health" ecosystem paradigm needs to be addressed (Lin, 2017). Historically, antimicrobials have been integral to animal production systems as both therapeutic and growth promoters (Manyi-Loh *et al.*, 2018). Poultry plays a significant role as a primary reservoir for *C. jejuni*. The misuse of antimicrobials in the human population and on farms has contributed to the escalating prevalence of fluoroquinolone resistance, a situation that demands our immediate attention and action.(Nicole M, and Iovine, 2013).

### *Campylobacter*'s antimicrobial resistance: UK

In the UK, a couple of longitudinal studies have focused on identifying the antimicrobial resistance of *Campylobacter jejuni* (Cody *et al.*, 2012; Vliet *et al.*, 2022). A 6-year longitudinal study was conducted between 2003 and 2009, which has shown that nine clonal-complexes are associated to be susceptible to fluoroquinolones (CC22, CC45, CC48, CC61, CC257, CC283, CC403, CC658, and CC677). Seven were related to high fluoroquinolone resistance (CC49, CC206, CC354, CC446, CC460, CC464, and CC607) (Cody *et al.*, 2012). In the UK around 38% of cases in the UK clinical isolates were resistant to ciprofloxacin in 2008 (Cody *et al.*, 2010).

Foods Standards Agencies has conducted a project FS102121 (previously FS241044): a microbiological survey of *Campylobacter* contamination in chilled chicken at a retail sale in the UK starting from July 2014 to 2020 (Food Standards Agency, 2015). In the year 2 study between 2015 and 2016, *C. jejuni* resistance to ciprofloxacin increased to 54.2%, with tetracycline resistance of 67.7% (Jorgensen *et al.*, 2017). For *C. coli*, its ciprofloxacin resistance was 48.2%, and tetracycline resistance was 56.9% out of the 548 isolates surveyed. In these six years, the resistance level did not change much, with 52.9% for *C. jejuni* and 43.7% for *C. coli*. However, tetracycline resistance has increased to over 60% (Table 1) The most recent study from the FSA has aggregated 20 years' worth of data points from various projects to identify antimicrobial resistance of chicken isolates from 2001 to 2020(Jorgensen, Rodgers, *et al.*, 2019). This survey, combined with the FSA original survey, has shown the resistance patterns in the poultry (Table 1). It has demonstrated that for both *C. jejuni* and *C. coli*, the level of resistance for fluoroquinolones and resistance from back in 2000 has significantly increased. It has shown that in *C. jejuni*, ciprofloxacin resistance has risen from 13% (2001-2005) to 47% (2011-2018), then it has gone up to 52.9% (2018-2020). For *C. coli,* the ciprofloxacin resistance has increased from 18% (2001-2005) to 48% (2014-2018) to 46.3% (2018-2020). For *C. coli,* the ciprofloxacin resistance has increased mainly from 27% (2001-2005) to 66% (2014-2018) to 61.4% (2018 – 2020). Tetracycline resistance in *C. coli* started at 23% (2001-2005), then increased to 55% (2011-2018), and gone up further to 66.3% (2018-2020).

*Table 1 A microbiological survey of campylobacter contamination in fresh whole UK-produced chilled chickens at retail sale from various surveys*

| **Project Name: Study period** | **Ciprofloxacin *C.jejuni* (%)** | **Ciprofloxacin *C.coli* (%)** | **Tetracycline *C.jejuni* (%)** | **Tetracycline *C.coli* (%)** | **Citations** |
|---|---|---|---|---|---|
| UKHSA:2001-2005 | 13.0 | 18.0 | 27 | 23.0 | (Jorgensen, Rodgers, *et al.*, 2019) |
| UKHSA:2011-2018 | 47.0 | 48.0 (2014-2018) | 66.0 (2014-2018) | 55.0 | (Jorgensen, Rodgers, *et al.*, 2019) |
| FSA 1: 2014-2015 (n = 283) | 49.0 | 55.0 | 63.0 | 68.0 | (Jorgensen *et al.*, 2015) |
| FSA 2: 2015-2016 | 54.2 | 48.1 | 53.2 | 48.1 | (Jorgensen *et al.*, 2017) |
| FSA 3: 2016-2017 (n = 585) | 41.0 | 52.0 | 54.0 | 62.0 | (Jorgensen *et al.*, 2018) |
| FSA 4: 2017-2018 (n = 1,044) | 52.0 | 48.0 | 52.0 | 60.0 | (Frieda *et al.*, 2019) |
| FSA 5: 2018-2019 (n = 1009) | 51.7 | 42.1 | 61.1 | 62.9 | (Jorgensen, Charlett, *et al.*, 2021) |
| FSA 6: 2019-2020 (n = 1008) | 52.9 | 43.7 | 61.5 | 66.3 | (Jorgensen, Kesby, *et al.*, 2021) |
| UKHSA 2018-2020 | 52.9 | 46.3 | 61.4 | 66.3 | (Jorgensen, Rodgers, *et al.*, 2019) |

### *Campylobacter*’s antimicrobial resistance: In the worldwide context

Here, we will bring some case studies from outside of the UK (Finland, Australia and China) to demonstrate the Campylobacter’s antibiotic resistance level.

Finland has steadfastly adhered to stringent controls governing antimicrobial usage in its farm animal populations. A study between 2000 and 2012 yielded compelling results, with 80% of broiler and bovine isolates retaining susceptibility to antimicrobial agents(Olkkola *et al.*, 2016). This finding attests to the efficacy of Finland's rigorous regulatory framework (Olkkola *et al.*, 2016).

On the other hand, despite the exclusion of fluoroquinolones in the usage of its livestock, Australia has encountered a perplexing phenomenon. Rigorous biosecurity measures aimed at preventing inter-flock contagion have not sufficed to curb the emergence of fluoroquinolone-resistant isolates within Australian poultry populations (Abraham *et al.*, 2020b). The hypothesis was that these resistant isolates could have arrived from external sources, possibly via pest species, wild birds, or human agency. The presence of fluoroquinolone-resistant isolates in a country where these antimicrobials are not sanctioned raises profound concerns for biosecurity(Abraham *et al.*, 2020a).

In China, there has recently been a shift in Campylobacteriosis being caused by *C.coli* than from *C. jejuni* (Zhang *et al.*, 2022). It also demonstrates that poultry is the most dominant product attributed to clinical Campylobacteriosis, with *C. coli* being the dominant species causing Campylobacteriosis, which is not seen in other European studies.

Moreover, in the farm setting, multiple antimicrobials are currently used, including macrolides, florfenicol, fluoroquinolones, tetracycline and others, causing the potential shift to *C.coli* prevalence (Zhang *et al.*, 2022). Furthermore, out of the isolates studied, gyrA T86I amino acid substitution was found in 94.23% of the sample (Zhang *et al.*, 2022). For *tet(O)*, the presence in 2021 is at 78.85%, an alarming resistance rate for *C.coli* (Zhang *et al.*, 2022). Even for *C. jejuni,* it has been shown that all isolates had at least resistance to one antimicrobial. Resistant to ciprofloxacin was found in 94.5% of *C. jejuni* and 93.5% for tetracycline (Zhang *et al.*, 2020). For Erythromycin(macrolides) was 9% resistance for *C.jejuni,* but for *C.coli,* it was 44%. Recently, it has been known that erythromycin-resistant *Campylobacter* has been found in poultry, which is a big problem in clinical settings due to being prescribed to patients currently (Zhang *et al.*, 2016).

## Machine Learning/Mathematical Modelling application to *Campylobacter*

In recent years, more mathematical models and machine learning approaches have been applied to *Campylobacter*. A mathematical model was implemented to identify dynamics within the flock of broiler chickens, where stochastic models were used to determine how more robust *Campylobacter* variants that are more robust and fit would outcompete the weaker variants *(Rawson, et al*, 2019). Another model used Bayesian models to explore the seasonal variations factors that influence the *Campylobacter jejuni* and coli presence in broiler chickens (Rawson *et al.*, 2020). This study identifies that individual birds have quite discrepancies in the responses to *C.jejuni* and *C.coli*, and identifying “super-shedder” birds in the flock and removing these super-shedder could lower the rates of shedding within the flock (Rawson *et al.*, 2020). There is also a machine learning study that focused on predicting source attribution of Campylobacteriosis (Arning *et al.*, 2021). This study has raised the attribution score from 64% (iSource population genetic approach) to 71% for MLST and identified that clonal-complex 21 is the most common disease-causing lineage (Arning *et al.*, 2021).

# Gap of knowledge:

## Lack of AMR trends within clonal-complex in longitudinal studies

The literature on *Campylobacter jejuni* on AMR focuses on species-level resistance. Furthermore, studies have investigated longitudinal studies. Moreover, the only handful FSA project focused on the chilled chicken and did not focus to identify variant changes on the clonal-complex level. Thus, in this thesis, in chapters 2 and 6, we will identify fluoroquinolones and tetracycline resistance and its trend across time using clonal-complex and NeighbourGroups through multiple variants. Logistic regression, clustering, and heatmap analysis will be extensively used to identify pattern recognition

## Lack of Clonal-complex revisited studies

A clonal-complex was developed for *Campylobacter* in the 2001 using only 194 isolates. Since then, MLST has been used worldwide as a crucial epidemiological variant classification. Nevertheless, no studies have revisited the classification, 20 years on to validate that it is still robust. Chapters 5 and 7 examined these questions to answer these knowledge gaps.

### Necessity for a novel classification that combines cgMLST and MLST data using Machine Learning

Given the availability of whole genome sequencing technologies and reduced sequencing costs, many recent isolates have cgMLST data. However, we also have historic isolates that do not contain cgMLST alleles. Thus, creating a classification that can be able to predict cgMLST-level variant groupings using MLST loci alone will be beneficial to the field. Therefore, NeighbourGroups were established in chapter 3 and applied to all the consecutive chapters afterwards.

### Gap in bioinformatics capability and lack of transparent pipelines for users:

We currently have over 130 bacterial species in the PubMLST classified with established MLST schemes. Housekeeping genes have been investigated through each species, but no method can predict cgMLST-level variant groups. In Chapter 4, we apply NeighbourGroups to *Campylobacter jejuni;* and the same pipeline can be used for establishing NeighbourGroups for all pathogens. The users do not require any HPC cluster to establish NeighbourGroups; and step-by-step guides are available on GitHub which is discussed in Chapter 3. These pipelines will be useful to fill the gap in bioinformatics knowledge to develop species-level NeighbourGroups for all pathogens.

# Chapter Outlines:

## Chapter 1: Introduction

Introduces the main topics ranging from *Campylobacter jejuni's* epidemiology, antimicrobial resistance mechanisms, and its classification systems to classify the variants. The introduction also covers gaps of knowledge in the field using small NLP studies that have investigated PubMed research to describe current literature pattern.

## Chapter 2: Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018]

https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000875

This chapter, which is a published paper, identified clonal-complexes level patterns for fluoroquinolone resistance in *Campylobacter jejuni.* Through 10,359 isolates from the UK from 1998 to 2018, we have found six distinct patterns of fluoroquinolone resistance using logistic regression. Furthermore, we have uncovered the complexity of *Campylobacter jejuni* as a species through heatmaps and predictive analysis.

## Chapter 3: NeighbourGroups: a machine learning classification tool that assigns microbial Multi-locus genotypes to clusters

https://www.researchsquare.com/article/rs-2666125/v1 [Code available at https://github.com/bgrdessislava/NeighbourGroups/blob/main/README.md].

This chapter, which is a pre-print paper, establishes the new generalizable machine learning classification called "NeighbourGroups" that classifies Multi-locus sequence types with defined precision. It uses cgMLST loci to train a CatBoost algorithm model that can predict the NeighbourGroups using only its MLST loci profile. The method details are illustrated step by step on the GitHub ReadMe file. This chapter is written in a brief communication format for Nature Methods journal. The application of NeighbourGroups, specifically on the *C.jejuni,* is further described in Chapter 4.

### Chapter 4: CatBoost classifier "NeighbourGroups" classification applied to *Campylobacter jejuni*

In this chapter, we have applied NeighbourGroups onto *Campylobacter jejuni*. We have further, compared the isolates classification between the clonal-complex and its NeighbourGroups to uncover more of its population dynamics. This chapter solidifies the importance of the NeighbourGroups approach to bacterial epidemiology, and the same pipeline could be used to establish NeighbourGroups for other bacterial species.

### Chapter 5: *Campylobacter jejuni* clonal-complex 353 re-evaluated through clonal-complex assignment methods: re-classification to multiple NeighbourGroups and proposal of CC581 as a new clonal-complex

This chapter focuses on one of the problematic variants, CC353, and tries to re-classify it using the clonal-complex and NeighbourGroups approach. In this chapter, we also uncover how clonal-complexes are assigned currently and demonstrate why some discrepancies have occurred. Furthermore, it has shown that we need to establish a new clonal-complex 581 onto the official clonal-complex classification.

### Chapter 6: Trends in Fluoroquinolone and Tetracycline Resistance of *Campylobacter jejuni* using NeighbourGroups]

This chapter contains the largest sample sizes (n=14,481) in this thesis to identify trends in fluoroquinolones and tetracycline using NeighbourGroups classification. This chapter combines Chapter 2's logistic regression and heatmaps analysis and Chapters 3 and 4's application of NeighbourGroups onto *Campylobacter jejuni.* It further investigates how some variants are only resistant to one of the two antimicrobial and we look further into bacterial population dynamics.

### Chapter 7: Clonal-complex re-visited: Which clonal-complex will stay the same, will change, and which will be retired

This chapter re-visits the way clonal-complexes are formed by identifying whether, per clonal-complex, we still see the highest allele per 7 MLST are still the same as its central genotype. This study used pie charts to demonstrate the results and split the results into four categories: not aligned, aligned caution and retirement. This chapter is the extension of Chapter 5, where we have tried to investigate clonal-complex classification.

# Chapter 2: Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018]

MICROBIAL GENOMICS

RESEARCH ARTICLE
Veltcheva *et al.*, *Microbial Genomics* 2022;8:000875
DOI 10.1099/mgen.0.000875

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

# Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018]

Dessislava Veltcheva*, Frances M. Colles, Margaret Varga, Martin C. J. Maiden and Michael B. Bonsall

**Abstract**

*Campylobacter jejuni* (*C.jejuni)* is the most common causative agent of bacterial food poisoning worldwide and is known to be genetically highly diverse. *C. jejuni* is increasingly resistant to fluoroquinolone antibiotics, but very few studies have investigated variant-specific patterns of resistance across time. Here we use statistical modelling and clustering techniques to investigate patterns of fluoroquinolone resistance amongst 10,359 UK isolates from human disease sampled over 20 years. We observed six distinct patterns of fluoroquinolone sensitivity/resistance in *C. jejuni* across time, grouping by clonal complex (CC). Some CCs were fully resistant, some shifted from susceptible to resistant following a sigmoidal shape, and some remained susceptible over time. Our findings indicate that the fluoroquinolone resistance patterns of *C. jejuni* are complicated and cannot be analysed as a single species but divided into variant dynamics so that the factors driving resistance can be thoroughly investigated.

## DATA SUMMARY

All data used in this study are open access and can be found in the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) by searching for public project No.110, '1998–2018 UK human isolates (n=10,359)' on the *Campylobacter jejuni/coli* isolates collection public projects dropdown menu on the 'isolate collection' section, or by using the following link: https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1 [1]. In addition, it can also be searched by publication in the PubMLST website.

## INTRODUCTION

Campylobacteriosis is the most widely known cause of human gastroenteritis, affecting 400–500 million individuals worldwide annually [2]. In the UK, healthcare spending related to campylobacteriosis is approximately GBP £50 million per annum [3]. Approximately 90% of campylobacteriosis cases are caused by the bacterium *Campylobacter jejuni*, and the rest are primarily caused by *Campylobacter coli* [4]. Approximately 600,000 people are infected annually in the UK alone and approximately 100 people are hospitalized each year [5, 6]

*Campylobacter* resides in the intestines of wild and domesticated animals, including chickens, cows, and pigs [7], all of which are potential sources of human infection. However, in high-income countries, contaminated chicken meat is the most common source of infection, causing an estimated 60–80% of human disease in the UK [8]. Approximately 73% of chickens sold in supermarkets are contaminated with *Campylobacter* and at least 7% have contamination on the external packaging [9]. Incidence of campylobacteriosis is sporadic, but they have been known to lead to outbreaks in restaurants from food sources such as chicken liver pâté or undercooked chicken meats [10]. Usually campylobacteriosis is self-limiting and symptoms resolve within 1 or 2 weeks,

---

Received 17 May 2022; Accepted 11 July 2022; Published 26 September 2022
**Author affiliations:** [1]Department of Biology, University of Oxford, Oxford, UK.
***Correspondence:** Dessislava Veltcheva, dessislava.veltcheva@spc.ox.ac.uk
**Keywords:** *Campylobacter jejuni*; fluoroquinolones; antimicrobial resistance; statistical modelling; clonal complex; Multi-locus sequence typing ; Generalised Linear Model.
**Abbreviations:** AIC, Akaike information criteria; CC, Clonal complex; *C.coli*, *Campylobacter coli*; *C.jejuni*, *Campylobacter jejuni*; MLST, Multi-locus sequence typing; WHO, World Health Organisation.
1998–2018 UK human isolates (n=10,359): https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1
**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table is available with the online version of this article.

**Impact Statement**

Antimicrobial resistance in *C. jejuni* poses a major clinical problem that could ultimately eliminate the efficacy of all known antibiotics in the future. Fluoroquinolone resistance is known to be increasing in *C. jejuni* over time. However, it is a diverse organism and patterns of resistance among its variants, known as clonal complexes, have not yet been well characterized. Our study uses statistical modelling, including generalized linear models and cluster analysis, to identify six distinct patterns of resistance. We estimate that two of the six patterns will reach near complete resistance levels by 2040. This study contributes new insights in the patterns of resistance among the diverse phylogeny of *C. jejuni* . In addition, we propose that antimicrobial resistance should be studied at the level of variants rather than species to describe the underlying traits affecting the evolution of resistance more accurately.

but antibiotics such as macrolides can be prescribed if the condition is prolonged, or for vulnerable individuals, including the young, the immunocompromised and the elderly [11].

Fluoroquinolone resistance in *C. jejuni* is part of the World Health Organization (WHO) high-alert watch list and it is increasing, threatening global antibiotic efficacy [12, 13]. A single-point mutation in the C257T at the *gyrA* gene within the quinolone resistance-determining region (QRDR) leads to Thr86Ile amino acid substitution, which can cause phenotypic resistance to fluoroquinolone in *C. jejuni* [14–17]. This amino acid change can be maintained stably in *Campylobacter* populations without antibiotic selection pressure and *in vivo* chicken colonization studies have shown that fluoroquinolone-resistant *Campylobacter* mutants do not carry a fitness burden [18]. In addition, these resistant isolates have been found to outcompete susceptible isolates, indicating that the mutation confers a selective advantage even in the absence of the antibiotic pressure [18]. It has also been shown that despite discontinuing the use of fluoroquinolones, *Campylobacter* found in farms continue to have high fluoroquinolone resistance even after four years, which suggests that once campylobacters become resistant to fluoroquinolones, it is challenging to reduce their prevalence [12].

*C. jejuni* has high genetic diversity, which makes investigating changing patterns of resistance particularly challenging [19]. Multi-locus sequence typing (MLST) schemes have revolutionized classification of *C. jejuni* globally, with clonal complex (CC) designation based upon seven-locus MLST giving insight into sources of infection [20, 21]. Although an EU growth promoter ban came into place in 2006 for livestock, antibiotics continue to be permitted for therapeutic reasons. While some studies highlight the trend in resistance, the in-depth dynamics and trends of fluoroquinolone resistance within *C. jejuni* CCs following legislative changes in antibiotic use on farms have not been thoroughly studied to date [22, 23].

Previous work has investigated correlation of CC and antimicrobial resistance change across six years from 2003 to 2009 from isolates in Oxfordshire, UK. However, this work did not include other UK regions [24]. Other studies have compared antimicrobial resistance across time but, they only considered *Campylobacter* as a single species, rather than examining variant dynamics [23]. This work aimed to readdress this and investigate patterns of changes in fluoroquinolone resistance changes across *C. jejuni's* clonal complexes.

### Gap statement/Aim

With increasing emphasis being placed on antibiotic stewardship in treating human infections, as well as livestock (including an EU ban on growth-promoting antibiotics in 2006), a decrease in fluoroquinolone resistance at a population level in recent years might be anticipated. Exploring variant-based levels of fluoroquinolone resistance in clinical isolates can provide insight into the sources of antimicrobial resistance emergence. This study aims to assess patterns in fluoroquinolone resistance in UK isolates over time, in relation to clonal complex.

## METHODS

### Metadata acquisition

The data used in our study are publicly available from the PubMLST database [1]. A total of 10,359 *Campylobacter* isolates were obtained from the PubMLST database and used in this study. Our search criteria were: ('Species=*Campylobacter jejuni*' *AND* 'Country=UK' *AND* '1990>=Year<=2020' *AND* 'N50>=20,000' *AND* '1.4Mb<=Genome Size <= 1.8 Mb' *AND* 'Contigs<=50' AND 'source=human_stool'). The samples used can also be found directly on PubMLST public project No.110 - '1998-2018 UK human isolates (n=10,359)': https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1.

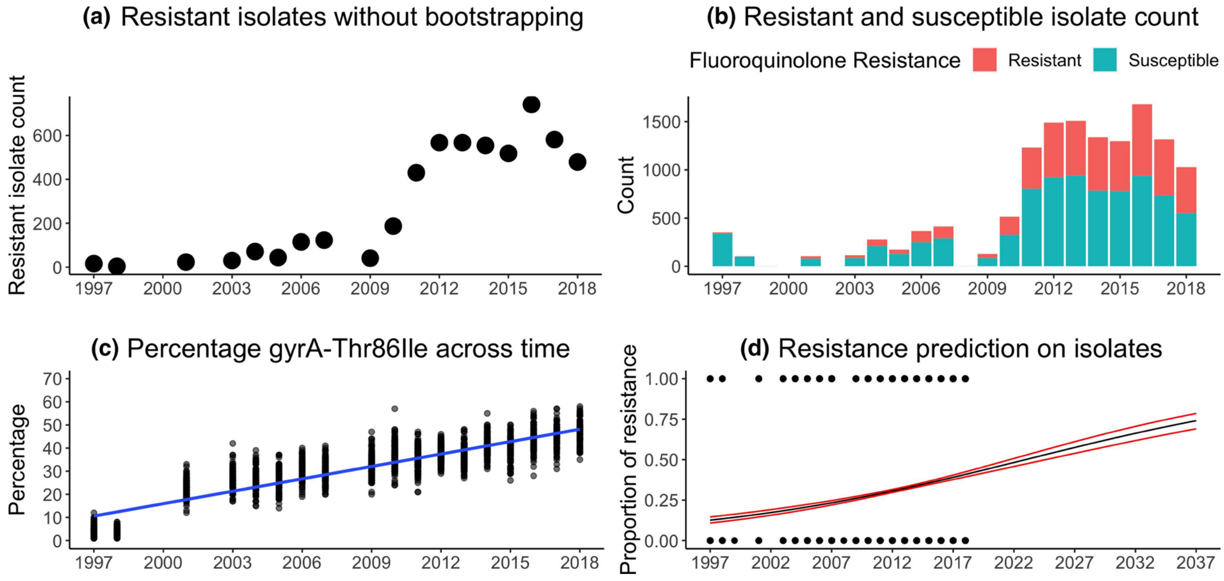


**Fig. 1.** GyrA-Thr86Ile changes across time. (a) Isolates counts shown without bootstrapping. (b) Resistant and susceptible isolates count across time. (c) Percentage of resistant isolates across time with bootstrapping method to give equal sample sizes for each year; 100 isolates randomly picked with 100 iterations. The blue line is a linear regression line ($y$=1.788 × −3560) (d) Prediction of fluoroquinolone resistance isolates by using the generalized model formula in the future years where 1 indicates resistance and 0 indicates susceptible isolates.

### Bioinformatics pipeline and software usage

Using whole-genome sequence data for each of the isolates chosen from the PubMLST database, the *gyrase A* (*gyrA*) (CAMP0950) locus was queried. The *gyrA* nucleotide sequences for all isolates were aligned to one another using MAFFT/7.305 [25]. The nucleotide alignment was translated to an amino acid alignment using Transeq [26]. Due to the absence of phenotypic data for the majority of the samples, the presence or absence of the Thr86Ile amino acid substitution was used to infer resistance or susceptibility, respectively [24, 27]. There are other mutations that are also known to lead to fluoroquinolone resistance, such as Asp90Asn, Thr86Lys, Thr86Ala, Thr86Val and Asp90Tyr, but these usually occur at lower frequency [15, 17]. Therefore, we focused solely on the amino acid substitution of Thr86Ile in this study.

The provenance and phenotypic data from the data acquisition step and the 86th position of GyrA were then aggregated for further analysis. The bootstrap analysis, heatmap clustering and GLM models were run using the R programming language [28], with libraries including ggplot2, stats, d3heatmap, reshape2, heatmap, dplyr, tidyverse, lme4 and Viridis. Code for the methods described is available here: https://github.com/bgrdessislava/Campyjejuni_gyrA.

### Identifying patterns in resistance over time

Due to varying sample sizes across years, bootstrap analysis was used to resample 100 isolates with replacement and compute the percentage resistance. This sampling was repeated 100 times, and a linear model was fitted to visualize the changes in resistance across time. To investigate the levels of resistance further, the amino acid of gyrase86, whether it was threonine or isoleucine, was used to cluster the CCs over time. To investigate the temporal dynamics, we used logistic regression through time to understand changes in the levels of resistance. Generalized linear models using a binary response variable (for susceptible or resistance variant) with year and group as covariates were used to investigate these trends across time further. Further code details can be found in GitHub (see above).

## RESULTS

### *C. jejuni* fluoroquinolone resistance increased over time, 1998–2018

In 1998, 5% of all the isolates were resistant to fluoroquinolones, but by 2018, 45% of all isolates were resistant to fluoroquinolones (Fig. 1a–c). Despite missing temporal data points (in 1999, 2000, 2002 and 2008), the linear regression analysis suggested that each year the percentage of resistant isolates increased by 1.788% (linear regression $y$=1.788 × −3560 with R squared: 0.7905 with F statistic: 6761 on 1 and 1791 DF, $P$-value: <0.0001, Fig. 1c). Predictions from this linear regression model showed continued increase in levels of resistance, with ~75% of all variants predicted to be fully resistant by ~2040 if no effective preventative action is taken (Fig. 1c). A generalized linear model (Fig. 1d) indicated statistical significance

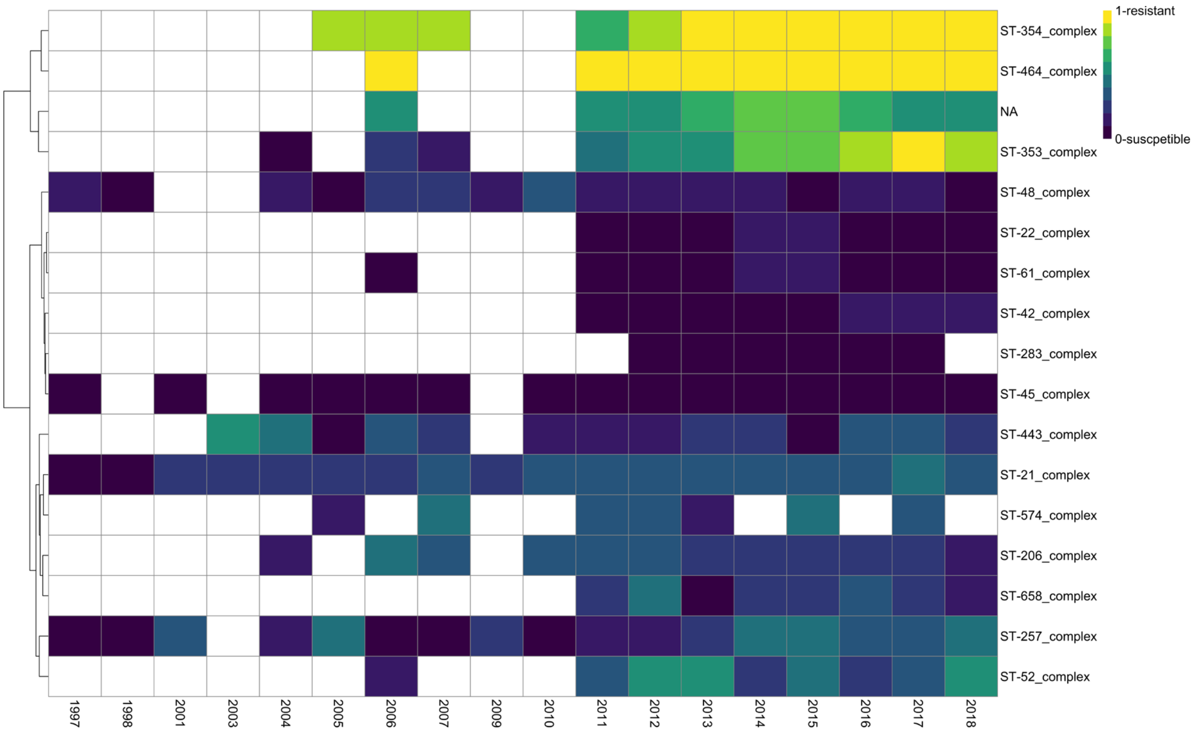


**Fig. 2.** Heatmap showing *C. jejuni* fluoroquinolone resistance by clonal complex across time. Scores were calculated from the resistant isolate count divided by the overall isolate count. Where all isolates were resistant, this gave a value=1 (e.g. 2018 ST-354_complex). White cells indicate there were no data points or omitted due to not passing the quality checks listed below. The top four rows indicate a resistance cluster; rows 5–10 indicate a susceptible cluster and rows 11–17 mixed resistance/susceptible cluster. Only data points containing at least 10 isolates per year, per clonal complex, were used to create this heatmap.

of the interaction between the year and groups covariates. The odds of being resistant increased by a log ratio of 1.077 each year [*P*-value<0.001, confidence intervals of 1.066 (2.5%) and 1.089 (97.5%)].

## Investigating fluoroquinolone resistance patterns in time

Hierarchical clustering analysis, using all 10,359 isolates and their gyrase 86 positions, revealed 3 specific clusters of fluoroquinolone sensitivity/resistance (Fig. 2). Although some CCs were missing in earlier years, this highlighted the potential emergence of new clonal complexes across time that were not prevalent in early years. Furthermore, we investigated whether the growth promoter ban in 2006 had any influence on the resistance, and for this we used the isolates from 1998. The first cluster, comprising CCs 353, 354 and 464, had a high proportion of resistant isolates. In total, 97.5% (345 of 354 isolates) of CC464 in this study were resistant. The second cluster, comprising a further six CCs (CCs 48, 22, 61,42, 283 and 45), predominantly included susceptible isolates. Just two CCs, CC45 and CC283, were fully susceptible over the course of the study. The third cluster was composed of a mix of variants where there was no apparent visual pattern that could be recognized, although resistance was generally greater than in the second cluster, with up to 60% of isolates resistant at a given time.

## Further pattern recognition using trend analysis

Variants could be assigned to one of three main clusters based on their overall resistance (Fig. 2). However, there could be an underlying pattern or trend across time that is not apparent. Therefore, a generalized linear model was used to investigate and identify the resistance patterns across time for each CC (Fig. 3). The three clusters defined in Fig. 2 could be further differentiated into six patterns; two of high resistance, two of mixed resistance/susceptibility, one showing slight reduced resistance and one of susceptibility over time.

By comparing two generalized linear models, one with three clusters (Fig. 2), with an Akaike information criteria (AIC) value of 10,444, and a second model with six groups (Fig. 3) with an AIC value of 10,311, the second model is a better fit for our data. The analysis further showed that the introduction of a growth promoter ban did not have a significant impact on the emergence and dynamics of resistance. Moreover, since 2006, some CCs, such as CC464, CC353 and CC257, increased in frequency.

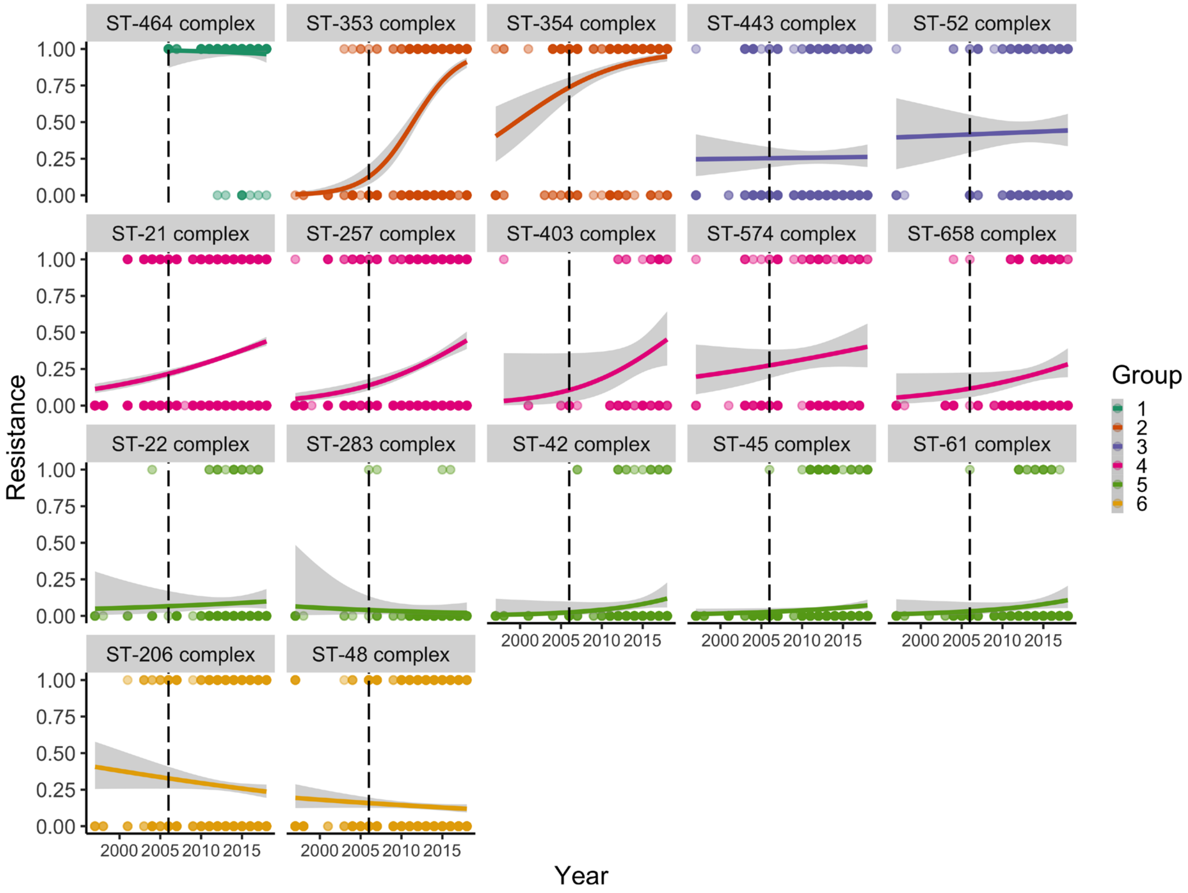


**Fig. 3.** Six patterns in fluoroquinolone resistance changes across clonal complexes. Group 1 illustrates CC464, which predominantly contains resistant isolates. Group 2 contains two clonal complexes, CC353 and CC354, which have had a sharp increase in resistance across time, almost reaching complete resistance in 2018. Group 3 illustrates the outlier cluster, which has no clear trend between resistance and time. Group 4 is a cluster with a weak positive correlation. Group 5 is a cluster that is predominantly susceptible across time. Group 6 is a cluster that has had a slight decrease in resistance. The black dotted line illustrates the 2006 mark where the EU growth promoter ban came into place. Grey shading illustrates the confidence interval. ST indicates clonal complex.

Using the predict function for generalized linear models (Fig. 4), the odds of group 1 (CC464) being resistant increased by a log ratio of 0.903 each year (*P*-value 0.459; 95% confidence interval of 0.663 to 1.16). However, future directions of resistance could not be predicted using the model due to the relatively late appearance of isolates from 2011 onwards, with a high proportion of isolates already showing resistance. For group 2 (CC353, CC354) the odds of being resistant increased by a log ratio of 1.210 each year (*P*-value <0.0001; 95% confidence interval of 1.164 to 1.259). Changes in resistance for group 2 had a sigmoidal shape change from susceptible to resistant, with most isolates predicted to be resistant by 2027.

For group 3 (CC443 and CC52) the odds of being resistant increased by a log ratio of 1.020 each year (*P*-value of 0.429; 95% confidence interval of 0.977 to 1.059). Due to an almost equal number of isolates across time, the prediction of future resistance trends is difficult, and resistance may potentially stay at the same level in the future. For group 4 (CC574, CC403, CC257, CC21, CC658), the odds of being resistant increased by a log ratio of 1.096 each year (*P*-value <0.0001; 95% confidence interval 1.078 to 1.114). Group 4 had a tight confidence interval, illustrating a steady increase in fluoroquinolone resistance, but with isolates not predicted to become fully resistant by 2040.

For group 5 (CC283, CC42, CC22, CC61, CC45) the odds of being resistant increased by a log ratio of 1.084 each year (*P*-value of 0.00369; 95% confidence interval of 1.0299 to 1.15). In terms of prediction, group 5 had a slow increase in resistance but did not follow the steady increase in resistance seen in group 2 or group 4's steady increase in resistance. For group 6 (CC206, CC48) the odds of being resistant increased by a log ratio of 0.982 each year (*P*-value of 0.143; 95% confidence interval of 0.958 to 1.007). Group 6 was the only group that indicated a slight decrease in resistance with a tight confidence interval.

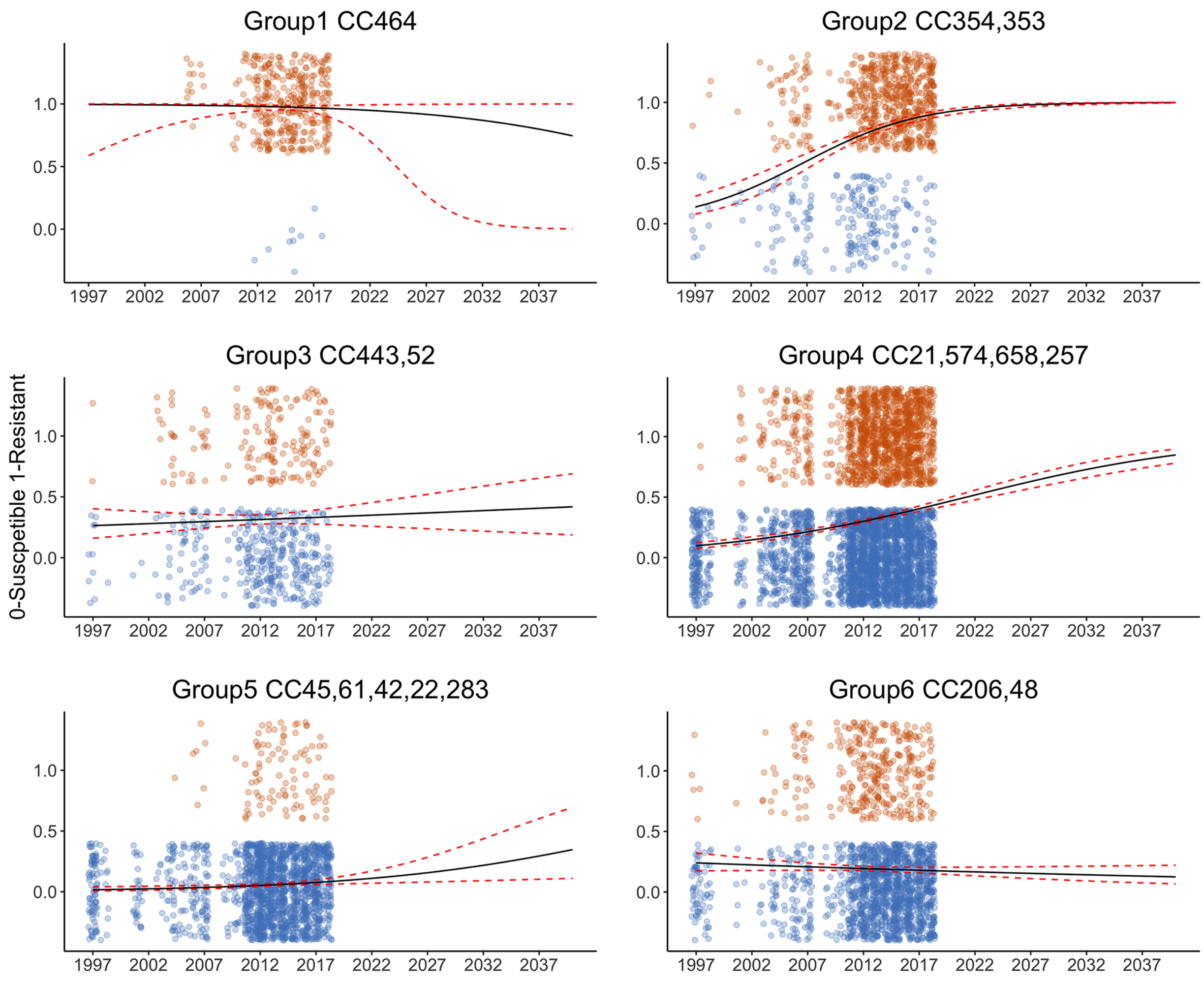


**Fig. 4.** Prediction of fluoroquinolone resistance until 2040. 0 indicates susceptible isolates, 1 indicates the resistant isolates. Jitter function shows the actual number (count) of isolates rather than only having values of 0 and 1, allowing us to see how some groups, such as group 4, have a high number of isolate (counts) compared to group 3, which has a lower number (count). The red lines indicate the confidence intervals. Prediction has been made until 2040.

In summary, the data indicated that for *C.jejuni* the trend for fluoroquinolone resistance was not uniform for the whole organism but rather exhibited six different patterns related to clonal complexes from 1998 to 2018. These six groups also gave different predictions for fluoroquinolone resistance in the future (Fig. 4)

## DISCUSSION

We investigated the changing temporal patterns of antimicrobial resistance across the CCs of *C. jejuni*. Previous studies of changes in fluoroquinolone resistance in *C. jejuni* have primarily focused on changes at a species level [15, 29], and have not been studied in CCs [23, 30]. While some studies have touched on multiple drug resistance, these patterns have not been thoroughly investigated at the CC level and conclusions have mainly been formed at the species level [31]. A study based in Oxford, UK, indicated that in 1995, 7% of the overall *Campylobacter* isolates were resistant to fluoroquinolones, but by 2008 this had risen to 37.5% [32]. Here, we have estimated that if this trend continues, by 2040, up to 75% of overall *C.jejuni* isolates will have fluoroquinolone resistance.

### Host source attribution

This study examined *C. jejuni* from human disease and did not investigate resistance from other *C. jejuni* sources. However, there have been many studies focused on identifying host source attribution in relation to CC [33, 34, Table S1]. A study focusing on host generalist *C. jejuni* CCs, such as CC45 and CC21, discussed how these CCs have adapted to colonize multiple hosts and estimates indicate that every 1.6 years (CC21) and 1.8 years (CC45) there are transmission events between different hosts, making it very challenging to pinpoint the source of clinical infections [35]. On the other hand, CC1034 and CC702 are commonly associated with wild geese, ducks and environmental waters and are rarely found in patients, indicating that overall these sources make a lower contribution to human infections [36]. CC61 is associated with cattle, and the fact that it has low fluoroquinolone resistance suggests that fluoroquinolone pressure in cattle farming might be lower, compared to poultry farming, where the poultry-associated CC464 and CC354 show high levels of resistance (see source attribution reference in Table S1, [37–39]). Combining host association into the model could be the next step to improve future predictions and identify further patterns for managing resistance.

### Generalized linear model to predict future fluoroquinolone resistance

Aspects of studying fluoroquinolone resistance across time have been considered in previous studies, including a six year study of human disease isolates from 2003 to 2009 that investigated the molecular epidemiology of *Campylobacter* in Oxfordshire, UK, and a review paper that focused on finding trends in fluoroquinolone resistance across time [15, 24]. The six year study of clinical isolates indicated that low susceptibility was found in several CCs (22, 45, 48, 61, 257, 283, 403, 658, 677) and high resistance in other CCs (49, 206, 354, 446, 460, 464, 607) [24]. In comparing these results to those of our study, CC353 was not strongly correlated with being resistant, but Cody *et al.*'s study observed seasonality changes in some CC peaking, such as CC353 being highly present in winter months [24]. Another study focusing on trends in antimicrobial resistance between the UK and USA showed that CC353, CC354 and CC464 have the highest levels of fluoroquinolone resistance in the UK; this finding was corroborated by our results [23]

A different study has found high ciprofloxacin resistance in *C. jejuni*, which has high genetic similarity (CC21 allelic type 1) in central Europe, indicating that this variant is potentially spreading clonally [22]. Another study based in Estonia and Latvia has shown that ST353 and ST5 are the most prevalent CCs found in clinical settings in Estonia, which has high ciprofloxacin resistance [40]. It is known that ST353 and ST5 only differ at one of the seven MLST loci, so the prevalence of these sequence types is not surprising, but it is interesting to note that these two sequence types were found to be high in resistance, whereas in our study, we have discovered CC464 and CC354 to be high in resistance. Our results demonstrate that (i) fluoroquinolone resistance is not uniform across *C. jejuni* CCs and (ii) there is a need for more detailed longitudinal surveillance of antimicrobial resistance at a sub-species level to understand more thoroughly how resistance is evolving.

### Restricted use of antibiotics in the EU and other countries

Other countries in the EU report similar trends in rising quinolone or fluoroquinolone resistance amongst *Campylobacter* isolates over time, although levels of resistance vary across time [41–44]. In Germany, nalidixic acid resistance increased from 8.2 to 26.3% between 1990 and 2004 [43]. In Italy, high quinolone resistance was significantly associated with *C. jejuni* isolated from food-producing animals despite a ban on fluoroquinolone usage for treatment [41, 42]. In the USA, the Food and Drug Administration (FDA) investigated how the ban on the fluoroquinolone, enrofloxacin, in poultry impacted on levels of resistance [45]. In the USA, fluoroquinolones were banned in 2005 but fluoroquinolone resistance increased from 13 to 25.3% between 1997 and 2015, which is generally lower than the levels reported in EU countries [46, 47].

These studies indicated more variation within years than between years but concluded that the withdrawal of fluoroquinolones did not contribute to lower resistance levels in *C. jejuni* [45]. In Peru, where there are no current restrictions on antibiotic use, a study compared two periods between 2001–2005 and 2006–2010. Fluoroquinolone resistance increased from 73.1 to 89.9% between 2001 and 2010, including in the Amazon region of Iquitos, where resistance rose from 24.1 to 48.9% [48]. In Australia fluoroquinolones have never been licensed for use in the commercial chicken sector [49]. However, a study found that 14.8% of 108 *C. jejuni* isolates from chickens at slaughter are fluoroquinolone-resistant, despite the absence of fluoroquinolone use for livestock [50]. Amongst these fluoroquinolone-resistant *C. jejuni* isolates, ST7323, ST2083 and ST2343 were highly abundant [50]. ST2343 is part of CC48 and in our study, we found that the overall resistance of this CC decreased, in contrast to trends in the UK to Australia.

Therefore, we found no evidence that the growth promoter ban has had a significant influence on overall fluoroquinolone resistance. Moreover, even in countries, such as Australia, that never licensed antibiotic usage on farms, resistance levels continue to increase.

### Does antibiotic stewardship lead to lower fluoroquinolone resistance levels?

Fluoroquinolone resistance can develop quickly and a study of commercial broiler chickens in the UK compared the level of resistance before, during and after treatment with difloxacin or enrofloxacin clearly shows this effect [51]. This study indicated that once fluoroquinolones were applied, the number of resistant *Campylobacter* isolates increased, and the resistant variants persisted for up to 4 weeks after treatment ended. Many studies have indicated that despite ongoing efforts to improve antibiotic stewardship in both farm and clinical settings, the level of fluoroquinolone resistance has not decreased and instead continues to increase over time [50, 52]. An extensive study analysing the resistome of 39,798 *Campylobacter jejuni* isolates was recently conducted and demonstrated a global trend of increasing tetracycline and fluoroquinolone resistance across time. The chicken-associated

CC354, CC573, CC464 and CC446 had the highest levels of resistance [30]. Another *in vivo* study in the USA determined that fluoroquinolone-resistant *Campylobacter* in chickens outcompeted fluoroquinolone-susceptible variants, illustrating that resistant isolates are biologically more fit in the chicken host [18].

The fact that resistant isolates have a fitness benefit raises the question of what underlying biological mechanisms could be selected for with this mutation in some variants but not others. The resistant clusters, CC464, CC354 and CC353, are known to be more poultry related, whereas CC42 and CC61, which are low in resistance, are known to be cattle related (Table S1). In addition, studies suggest that resistant isolates are associated with higher virulence and survival phenotypes, raising questions about why, although the regulations have changed, we do not see decreases in resistance in some of the variants [23]. A recent study indicated that the acquisition of fluoroquinolone resistance leads to an increase in biofilm formation, which leads to an increased ability to survive in the external environment, and also higher pathogenicity in human infection [53]

Here we examined fluoroquinolone resistance, but the isolates could be resistant to other antibiotics deriving from different mechanisms. A study compared UK and USA isolates and their different antibiotics markers revealed that CC354 and CC464 are high in both fluoroquinolone and tetracycline resistance, while some CCs, such as CC206, are only high in tetracycline resistance, showing different patterns from those in our study [23]. Further investigating the differences between each country's antibiotic regulations and farming and cultural features, including the food chain stages, could provide additional insights into why CCs differ.

## Conclusion

Overall, *C. jejuni* is showing an increase in fluoroquinolone resistance over time in the UK, leading to a prediction of 75% overall resistance by 2040. Our clustering analysis has shown that resistance levels vary by clonal complex, with CC353, CC354 and CC464 being the most resistant, and predicted to become almost 100% resistant to fluoroquinolones by 2040. Our results indicate that consideration of species variant is essential for elucidating the epidemiology and emergence of antimicrobial resistance for this diverse organism.

Funding information
BBSRC DTP (BB/M011224/1) and the FSA (FS101013); Oracle: DDT0006Z AT01.07.

Acknowledgements
The authors acknowledge the support of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. We also acknowledge the Food Standards Agency for their funding, and the BBSRC Doctoral Training Programme for further financial support. The authors thank Sumali Bajaj for advice and support on statistical modelling concepts. D.V. thanks Stephen Richer who has been an enormous support through the PhD process.

Author contributions
Conceptualization: D.V., M.B., M.M., F.C. Methodology: D.V. Software: D.V. Validation: M.V., F.C. Formal analysis: D.V., M.B. Investigation: D.V. Resources: F.C., M.M. Data curation: F.C. Writing – original draft preparation: D.V. Funding: M.M., M.B. Writing – review and editing: M.B., F.C., M.M., M.V. Visualization: D.V. Supervision: M.B., M.M., F.C., M.V. Project administration: D.V., M.B., M.M., F.C.

Conflicts of interest
There is no conflict of interest between the authors.

References
1. **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.
2. **Igwaran A**, **Okoh AI**. Human campylobacteriosis: a public health concern of global importance. *Heliyon* 2019;5:e02814.
3. **Tam CC**, **O'Brien SJ**. Economic Cost of Campylobacter, Norovirus and Rotavirus Disease in the United Kingdom. *PLoS ONE* 2016;11:e0138526.
4. **Sheppard SK**, **Maiden MCJ**. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. *Cold Spring Harb Perspect Biol* 2015;7:a018119.
5. **Goddard MR**, **O'Brien S**, **Williams N**, **Guitian J**, **Grant A**, *et al*. A restatement of the natural science evidence base regarding the source, spread and control of *Campylobacter* species causing human disease. *Proc Biol Sci* 2022;289:20220400.
6. **Zhou S-M**, **Lyons RA**, **Rahman MA**, **Holborow A**, **Brophy S**. Predicting hospital readmission for campylobacteriosis from electronic health records: a machine learning and text mining perspective. *J Pers Med* 2022;12:86.
7. **Heredia N**, **García S**. Animals as sources of food-borne pathogens: A review. *Anim Nutr* 2018;4:250–255.
8. **Yahara K**, **Méric G**, **Taylor AJ**, **de Vries SPW**, **Murray S**, *et al*. Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ Microbiol* 2017;19:361–380.
9. **Jorgensen F*et al***. Survey report. A Microbiological survey of campylobacter contamination in fresh whole UK produced chilled chickens at retail sale (2014-15); 2015
10. **Wensley A**, **Padfield S**, **Hughes GJ**. An outbreak of campylobacteriosis at a hotel in England: the ongoing risk due to consumption of chicken liver dishes. *Epidemiol Infect* 2020;148:e32.
11. **Same RG**, **Tamma PD**. 'Campylobacter infections in children', pediatrics in review. *NIH Public Access* 2018;39:533.
12. **Luangtongkum T**, **Jeon B**, **Han J**, **Plummer P**, **Logue CM**, *et al*. Antibiotic resistance in *Campylobacter*: emergence, transmission and persistence. *Future Microbiol* 2009;4:189–200.
13. **Tacconelli E*et al***. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics; 2017. http://www.cdc.gov/drugresistance/threat-report-2013/ [accessed 5 April 2022].

14. Food Standards Agency. *Enhanced Molecular-Based Surveillance and Source Attribution of Campylobacter Infections in the UK*. 2021.
15. Sproston EL, Wimalarathna HML, Sheppard SK. Trends in fluoroquinolone resistance in Campylobacter. *Microb Genom* 2018;4.
16. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, *et al*. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.
17. Wieczorek K, Osek J. Antimicrobial resistance mechanisms among *Campylobacter*. *Biomed Res Int* 2013;2013:12.
18. Luo N, Pereira S, Sahin O, Lin J, Huang S, *et al*. Enhanced *in vivo* fitness of fluoroquinolone-resistant *Campylobacter jejuni* in the absence of antibiotic selection pressure. *Proc Natl Acad Sci U S A* 2005;102:541–546.
19. Skarp CPA, Hänninen ML, Rautelin HIK. Campylobacteriosis: the role of poultry meat. *Clin Microbiol Infect* 2016;22:103–109.
20. Dingle KE, Colles FM, Ure R, Wagenaar JA, Duim B, *et al*. Molecular characterization of *Campylobacter jejuni* clones: A basis for epidemiologic investigation. *Emerg Infect Dis* 2002;8:949–955.
21. Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, *et al*. Host association of Campylobacter genotypes transcends geographic variation. *Appl Environ Microbiol* 2010;76:5269–5277.
22. Kovač J, Čadež N, Stessl B, Stingl K, Gruntar I, *et al*. High genetic similarity of ciprofloxacin-resistant *Campylobacter jejuni* in central Europe. *Front Microbiol* 2015;6:1169.
23. van Vliet AHM, Thakur S, Prada JM, Mehat JW, La Ragione RM. Genomic screening of antimicrobial resistance markers in UK and US *Campylobacter* Isolates highlights stability of resistance over an 18-year period. *Antimicrob Agents Chemother* 2022;66:e0168721.
24. Cody AJ, McCarthy NM, Wimalarathna HL, Colles FM, Clark L, *et al*. A longitudinal 6-year study of the molecular epidemiology of clinical campylobacter isolates in Oxfordshire, United kingdom. *J Clin Microbiol* 2012;50:3193–3201.
25. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
26. Madeira F, Park YM, Lee J, Buso N, Gur T, *et al*. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* 2019;47:W636–W641.
27. Han J, Wang Y, Sahin O, Shen Z, Guo B, *et al*. A fluoroquinolone resistance associated mutation in gyrA affects DNA supercoiling in *Campylobacter jejuni*. *Front Cell Infect Microbiol* 2012;2:21.
28. Team, R. C. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.r-project.org/ [accessed 19 April 2022].
29. Haldenby S, Bronowski C, Nelson C, Kenny J, Martinez-Rodriguez C, *et al*. Increasing prevalence of a fluoroquinolone resistance mutation amongst *Campylobacter jejuni* isolates from four human infectious intestinal disease studies in the United Kingdom. *PLoS ONE* 2020;15:e0227535.
30. Cobo-Díaz JF, González del Río P, Álvarez-Ordóñez A. Whole resistome analysis in *Campylobacter jejuni* and *C. coli* genomes available in public repositories. *Front Microbiol* 2021;12:1155.
31. Wimalarathna HML, Richardson JF, Lawson AJ, Elson R, Meldrum R, *et al*. Widespread acquisition of antimicrobial resistance among *Campylobacter* isolates from UK retail poultry and evidence for clonal expansion of resistant lineages. *BMC Microbiol* 2013;13:160.
32. Cody AJ, Clarke L, Bowler ICJW, Dingle KE. Ciprofloxacin-resistant campylobacteriosis in the UK. *Lancet* 2010;376:1987.
33. Arning N, Sheppard SK, Bayliss S, Clifton DA, Wilson DJ. Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS Genet* 2021;17:10.
34. Mourkas E, Yahara K, Bayliss SC, Calland JK, Johansson H, *et al*. Host ecology regulates interspecies recombination in bacteria of the genus *Campylobacter* . *Microbiology* 2021. DOI: 10.1101/2021.08.24.457495.
35. Dearlove BL, Cody AJ, Pascoe B, Méric G, Wilson DJ, *et al*. Rapid host switching in generalist *Campylobacter* strains erodes the signal for tracing human infections. *ISME J* 2016;10:721–729.
36. Colles FM, Jones TA, McCarthy ND, Sheppard SK, Cody AJ, *et al*. *Campylobacter* infection of broiler chickens in a free-range environment. *Environ Microbiol* 2008;10:2042–2050.
37. Calland JK, Pascoe B, Bayliss SC, Mourkas E, Berthenet E, *et al*. Quantifying bacterial evolution in the wild: a birthday problem for *Campylobacter* lineages. *PLOS Genet* 2021;17:e1009829.
38. Cody AJ, Colles FM, *et al*. Where does campylobacter come from? A molecular odyssey', advances in experimental medicine and biology. *Europe PMC Funders* 2010;659:47.
39. Pascoe B, Méric G, Yahara K, Wimalarathna H, Murray S, *et al*. Local genes for local bacteria: Evidence of allopatry in the genomes of transatlantic *Campylobacter* populations. *Mol Ecol* 2017;26:4497–4508.
40. Mäesaar M, Meremäe K, Ivanova M, Roasto M. Antimicrobial resistance and multilocus sequence types of *Campylobacter jejuni* isolated from Baltic broiler chicken meat and Estonian human patients. *Poult Sci* 2018;97:3645–3651.
41. European Food Safety Authority and European Centre for Disease Prevention and Control (EFSA and ECDC). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. *EFSA J* 2018;16:e05500.
42. García-Fernández A, Dionisi AM, Arena S, Iglesias-Torrens Y, Carattoli A, *et al*. Human Campylobacteriosis in Italy: emergence of multi-drug resistance to Ciprofloxacin, Tetracycline, and Erythromycin. *Front Microbiol* 2018;9:1906.
43. Luber P, Wagner J, Hahn H, Bartelt E. Antimicrobial resistance in *Campylobacter jejuni* and *Campylobacter coli* strains isolated in 1991 and 2001-2002 from poultry and humans in Berlin, Germany. *Antimicrob Agents Chemother* 2003;47:3825–3830.
44. Tenhagen B-A, Flor M, Alt K, Knüver M-T, Buhler C, *et al*. Association of antimicrobial resistance in *Campylobacter* spp. in Broilers and Turkeys with antimicrobial use. *Antibiotics (Basel)* 2021;10:673.
45. Zawack K, Li M, Booth JG, Love W, Lanzas C, *et al*. Monitoring antimicrobial resistance in the food supply chain and its implications for FDA policy initiatives. *Antimicrob Agents Chemother* 2016;60:5302–5311.
46. Centers for Disease Control. *Antibiotic Resistance Threats in the United States, 2019*. 2019.
47. Rodrigues JA, Cha W, Mosci RE, Mukherjee S, Newton DW, *et al*. Epidemiologic Associations Vary Between Tetracycline and Fluoroquinolone Resistant *Campylobacter jejuni* Infections. *Front Public Health* 2021;9:820.
48. Pollett S, Rocha C, Zerpa R, Patiño L, Valencia A, *et al*. Campylobacter antimicrobial resistance in Peru: a ten-year observational study. *BMC Infect Dis* 2012;12:193.
49. Wallace RL, Bulach DM, Jennison AV, Valcanis M, McLure A, *et al*. Molecular characterization of *Campylobacter* spp. recovered from beef, chicken, lamb and pork products at retail in Australia. *PLOS ONE* 2020;15:e0236889.
50. Abraham S, Sahibzada S, Hewson K, Laird T, Abraham R, *et al*. Emergence of fluoroquinolone-resistant *Campylobacter jejuni* and *Campylobacter coli* among Australian chickens in the absence of fluoroquinolone use. *Appl Environ Microbiol* 2020;86:e02765-19.
51. Humphrey TJ, Jørgensen F, Frost JA, Wadda H, Domingue G, *et al*. Prevalence and subtypes of ciprofloxacin-resistant *Campylobacter* spp. in commercial poultry flocks before, during, and after treatment with fluoroquinolones. *Antimicrob AGENTS Chemother* 2005;49:690–698.
52. Yang Y, Feye KM, Shi Z, Pavlidis HO, Kogut M, *et al*. A historical review on antibiotic resistance of foodborne *Campylobacter Front Microbiol* 2019;10:1509.
53. Whelan MVX, Ardill L, Koide K, Nakajima C, Suzuki Y, *et al*. Acquisition of fluoroquinolone resistance leads to increased biofilm formation and pathogenicity in *Campylobacter jejuni*. *Sci Rep* 2019;9:1–13.

**Statement of Authorship for joint/multi-authored papers for PGR thesis**

To appear at the end of each thesis chapter submitted as an article/paper

# Chapter 2:

| | |
|---|---|
| Title of Paper | Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018] |
| Publication Status | ☒Published ☐ Accepted for Publication<br>☐Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Veltcheva, D., Colles, F.M., Varga, M., Maiden, M.C. and Bonsall, M.B., 2022. Emerging patterns of fluoroquinolone resistance in *Campylobacter jejuni* in the UK [1998–2018]. *Microbial Genomics*, *8*(9). |

## Student Confirmation

| | | | |
|---|---|---|---|
| Student Name: | Dessislava Veltcheva | | |
| Contribution to the Paper | • Conceptualization of the research project<br>• Establishing Methodology<br>• Software and script Creation<br>• Formal analysis and Investigation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project Management<br>(Listed others contribution on the published paper) | | |
| Signature | | Date | 02.10.2023 |

## Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

| | | | |
|---|---|---|---|
| Supervisor name and title: Professor Michael Bonsall | | | |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. | | | |
| Signature | | Date | 2nd October 2023 |

# Chapter 3: NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters

# NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters

**Dessislava Veltcheva** (dessislava.veltcheva@spc.ox.ac.uk)
University of Oxford
**Stephen Richer**
University of Bath
**Samuel Sheppard**
University of Oxford
**Margaret Varga**
University of Oxford
**Frances Colles**
University of Oxford
**Michael Bonsall**
University of Oxford
**Martin Maiden**
University of Oxford

---

**Article**

**Keywords:**

**Posted Date:** March 22nd, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2666125/v1

---

# Abstract

Robust microbial classification systems are essential, but their definition is complicated by the large size and high diversity of microbial populations combined with a widespread horizontal genetic exchange. Multi-locus approaches that index gene variation without explicit phylogenetic classification mitigates these problems, but reproducibly defining high-level groups remains problematic. We describe a generalisable machine learning approach, 'NeighbourGroups', that reproducibly, robustly, and rapidly classifies multi-locus sequence types with defined precision.

# Full Text

Most bacteria exist in very large populations, and the combination of high growth rates, short generation times, extensive horizontal gene transfer (HGT), and strong selection can lead to very high diversity along with variable levels of clonality[1]. For many applications, notably infectious disease epidemiology, robust classification systems that pragmatically and reproducibly differentiate variants at high resolution are essential. Multi-locus sequence typing (MLST) was developed to solve this problem, indexing sequence variation using a limited number, often as few as seven, housekeeping gene fragments without explicitly classifying them phylogenetically[2]. The sequence variation of these fragments is recorded as alleles and combinations of alleles as sequence types (STs) that can be organised into groups or clonal complexes (ccs), sometimes referred to as 'eBurst groups' (sBGs)[3]. As sequence capacity has increased, additional schemes with more loci have been introduced, including ribosomal MLST (rMLST, indexing the 53 ribosomal protein genes) and core genome (cgMLST, indexing all shared genes in a particular population); however, seven-locus classifications remain widely understood and used as a cornerstone for bacterial typing[4,5].

Whilst defining alleles and sequence types are straightforward, as they are effectively summaries of sequence variation, representing higher-level groups, such as clonal complexes, is more problematic. In addition to HGT confusing purely phylogenetic approaches, the existence of intermediate variants can result in all variants merging into a single group. These problems are less intense for schemes with very large numbers of loci, but for seven-locus MLST schemes, pragmatic solutions have been adopted, such as defining clonal complexes with a central genotype[5]. However, while establishing a stable classification system, these approaches can misclassify STs into incorrect clonal complexes, as they rely on assumptions about the representativeness of the data set being analysed, which may or may not be correct. They can also be unstable to the addition of new data.

We have addressed this problem by leveraging the availability of large numbers of whole genome sequences and machine learning techniques. First, cgMLST data are analysed using the minimum spanning tree reconstruction method to establish clusters or 'Neighbour Groups' based on the similarity of their cgMLST profiles. Then, a supervised machine learning algorithm is used to optimally predict the membership of these clusters from fewer loci, such as the MLST loci. The trained algorithm enables a robust probabilistic assignment of a seven-locus genotype to a cluster defined with cgMLST data (Figure

1), which is especially helpful when whole genome sequence data are unavailable, for example, from clinical specimens, as WGS technology is not available or for legacy data. The algorithm is available as a command line tool accessible from https://github.com/bgrdessislava/NeighbourGroups.

An essential parameter for the NeighbourGroups model is the number of classification groups, which is user-defined and can be established empirically. For example, with a *Campylobacter* dataset of >10,000 isolates for which cgMLST data were available[6], we performed a grid search to assess model performance for two to 100 classification groups (Figure 2). Model performance was evaluated with an adjusted Rand score, which determined whether two clusters were similar between the 'testing tree' and 'true tree'. An adjusted Rand score of >0.90 was defined as an excellent prediction, 0.80-0.90 good recovery of groups, 0.65-0.80 moderate recovery, and <0.65 poor recoveries, with low confidence in the reproducibility of the classification. This analysis indicated that, for this dataset, 20 NGroups gave an optimum performance, with an adjusted Rand score of 0.895, showing high agreement between the 20 groups assigned from cgMLST with those assigned from the seven-locus MLST data.

At the time of writing, there were more than 150 MLST schemes available for a wide range of microbial species, primarily bacteria, with hundreds of thousands of isolates typed to the level of seven loci MLST and, in many cases, also with cgMLST. Most of these can be found on the PubMLST website (https://www.pubmlst.org)[7]. From some MLST databases, notably those for *Neisseria* species[8], *Campylobacter jejuni. Campylobacter coli* [6], and *Salmonella enterica* [3], clonal complexes (or eBurst groups), have been defined using a variety of approaches, but for most data collections, such groups have not been rigorously defined or maintained. Given the variability of bacterial population structures, the number of different schemes and the number of isolates available, there is a need for a rational and automated approach to defining groups which can be applied to whole genome and MLST data. This is especially the case for pathogens for which it may not be possible to generate reliable whole genome sequence data from clinical specimens but where this information is beneficial. In addition, the Neighbour Group approach is easily implemented, and its assumptions easily understood, providing a pragmatic complement to other analysis approaches, many of which require whole genome sequences and high-capacity computing [9,10]. A final advantage is that the approach indicates the confidence with which seven locus data can be assigned to whole genome sequence groups.

# Methods

### Isolates acquisition and tree creation

10,359 high-quality *Campylobacter jejuni* U.K genomes were downloaded from PubMLST [https://publmlsts.org ](Jolley, Bray and Maiden, 2018) using the following search query: *'Species=Campylobacter jejuni' AND 'Country=UK' AND '1990>=Year<=2020' AND 'N50>=20,000' AND '1.4Mb<=Genome Size <= 1.8 Mb' AND 'Contigs<=50' AND 'source=human_stool'.* (Only clinical isolate of *C.jejuni* were queried that led to food poisoning in humans for this study.)

*The output of this query is available at:* https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1

### Tree Construction

The NeighbourGroups methodology requires the construction of two phylogenetic trees. The first is a "True Tree" containing all isolates, and the second is a "Training Tree" containing a subset of isolates used for model training. The model's classification accuracy is then assessed using testing isolates from the "True Tree". This step ensures the testing isolates are unseen by the model during training and helps prevent data leakage. Minimum spanning trees were constructed using the coreMLST values (1,343 loci in *C. jejuni*) using GrapeTree (Zhou *et al.*, 2018). The "True Tree" contained 10,359 isolates, and the "Training Tree" had a randomly selected 80% subset. Trees were output in Newick format.

### NeighbourGroup Assignment

The two Newick format trees were initially processed to generate linkage matrices using Python. Briefly, Newick files were read using ete3 (Huerta-Cepas, Serra and Bork, 2016). Following this, trees were converted to cophenetic matrices. Next, cophenetic matrices were converted to condensed distance matrices and, subsequently, linkage matrices using SciPy (Virtanen *et al.*, 2020). Finally, linkage matrices were passed to fcluster to perform hierarchical clustering and extract a predetermined number of groups.

### CatBoost Classifier Training

Following group assignment, the training isolates were passed to the CatBoost supervised learning multi-class classification algorithm. CatBoost is a gradient-boosting algorithm for decision trees and works well with categorical values. The training was performed using the seven housekeeping MLST genes for *Campylobacter jejuni* (*aspA, glnA, gltA, glyA, pgm, tkt* and *uncA*) as features and the NeighbourGroup assignment as the target.

### Testing Classifier Performance

Following training, model performance was assessed using the 20% unseen isolates excluded from the training step. Predicted groups were compared against the "true" groups defined by the "True Tree" using an adjusted Rand Index. The Rand Index is a measure of similarity between two data clustering's.

### Determining Optimal Group Number

The maximum number of classification groups (NGroups) is a user-defined hyper-parameter of the NeighbourGroups methodology. A range of neighbour groups was tested between 2 and 100 to identify an optimal number of NeighbourGroups. For each NGroup number, the classifier was retrained, and the adjusted Rand score was computed.

### Retraining and Model Deployment

Following the validation of model performance, a final NeighbourGroups model was built using the complete set of isolates. The final trained model can be deployed to classify novel isolates. Each prediction outputs a predicted group and an associated probability. Appropriate probability thresholds may be determined to categorise unknown isolates. The Neighbour Groups methodology described here is highly generalisable and can be applied across a wide range of bacterial species.

# Declarations

### Data Availability

All the data are freely available and can be found on the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018). The datasets used during the current study are available by accessing the publicly available project No.110: 1998-2018 UK human UK isolates (n=10,359). https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1.

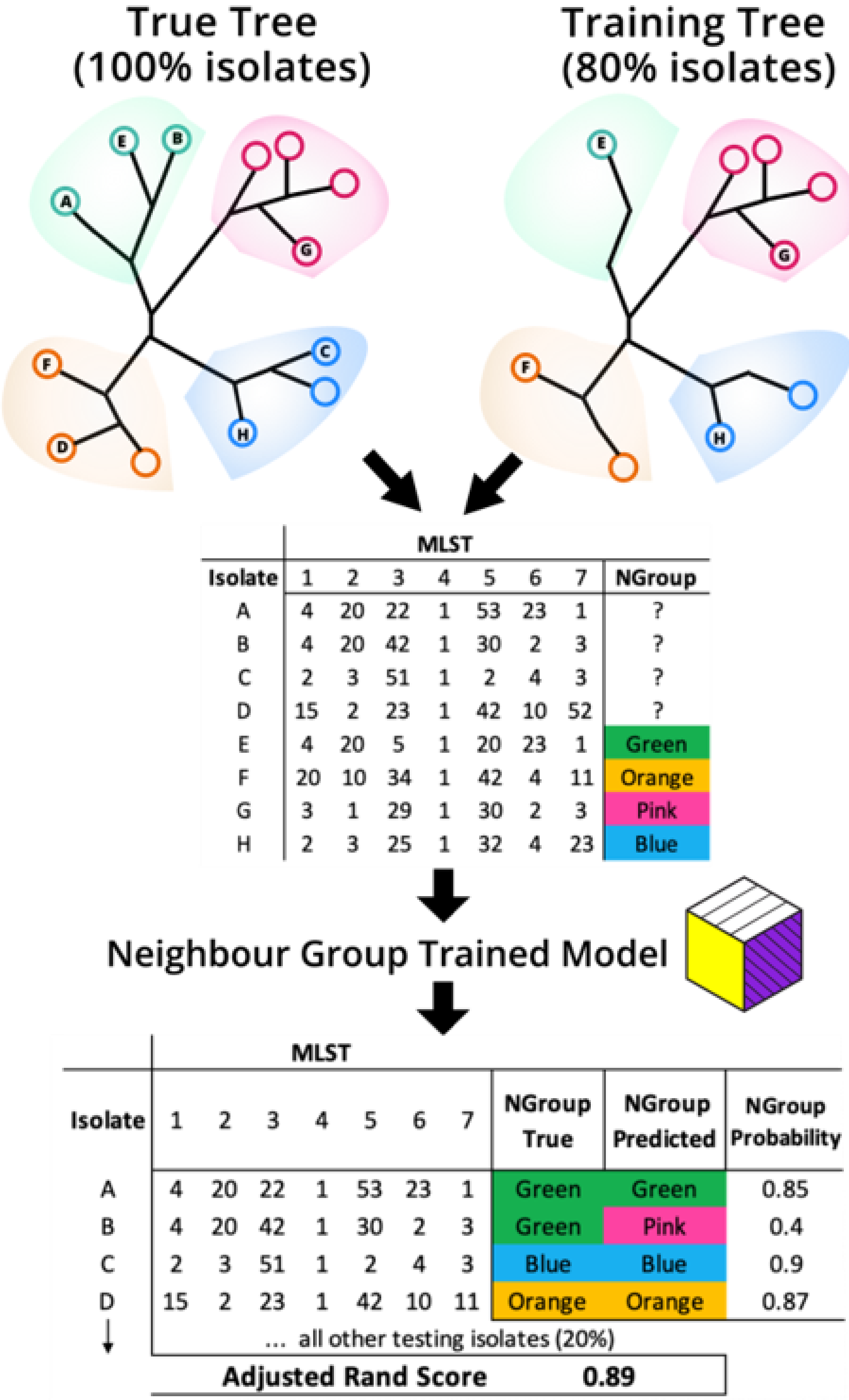
# References

1. VanInsberghe, D., Arevalo, P., Chien, D. & Polz, M. F. How can microbial population genomics inform community ecology? *Philos Trans R Soc Lond B Biol Sci* **375**, 20190253, doi:10.1098/rstb.2019.0253 (2020).
2. Maiden, M. C. J. *et al.* Multi-locus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* **95**, 3140-3145, doi:10.1073/pnas.95.6.3140 (1998).
3. Achtman, M. *et al.* Multilocus Sequence Typing as a Replacement for Serotyping in *Salmonella enterica*. *Plos Pathogens* **8**, e1002776, doi:10.1371/journal.ppat.1002776 (2012).
4. Maiden, M. C. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **11**, 728-736, doi:10.1038/nrmicro3093 (2013).
5. Maiden, M. C. Multi-locus Sequence Typing of Bacteria. *Annu. Rev. Microbiol.* **60**, 561-588 (2006).
6. Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D. & Maiden, M. C. J. Core Genome Multi-locus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates. *J Clin Microbiol* **55**, 2086-2097, doi:10.1128/JCM.00080-17 (2017).
7. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124, doi:10.12688/wellcomeopenres.14826.1 (2018).
8. Bratcher, H. B., Corton, C., Jolley, K. A., Parkhill, J. & Maiden, M. C. A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. *BMC Genomics* **15**, 1138, doi:10.1186/1471-2164-15-1138 (2014).

9. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Computational Biology* **11**, e1004041, doi:10.1371/journal.pcbi.1004041 (2015).

10. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15, doi:10.1093/nar/gku1196 (2015).

# Figures



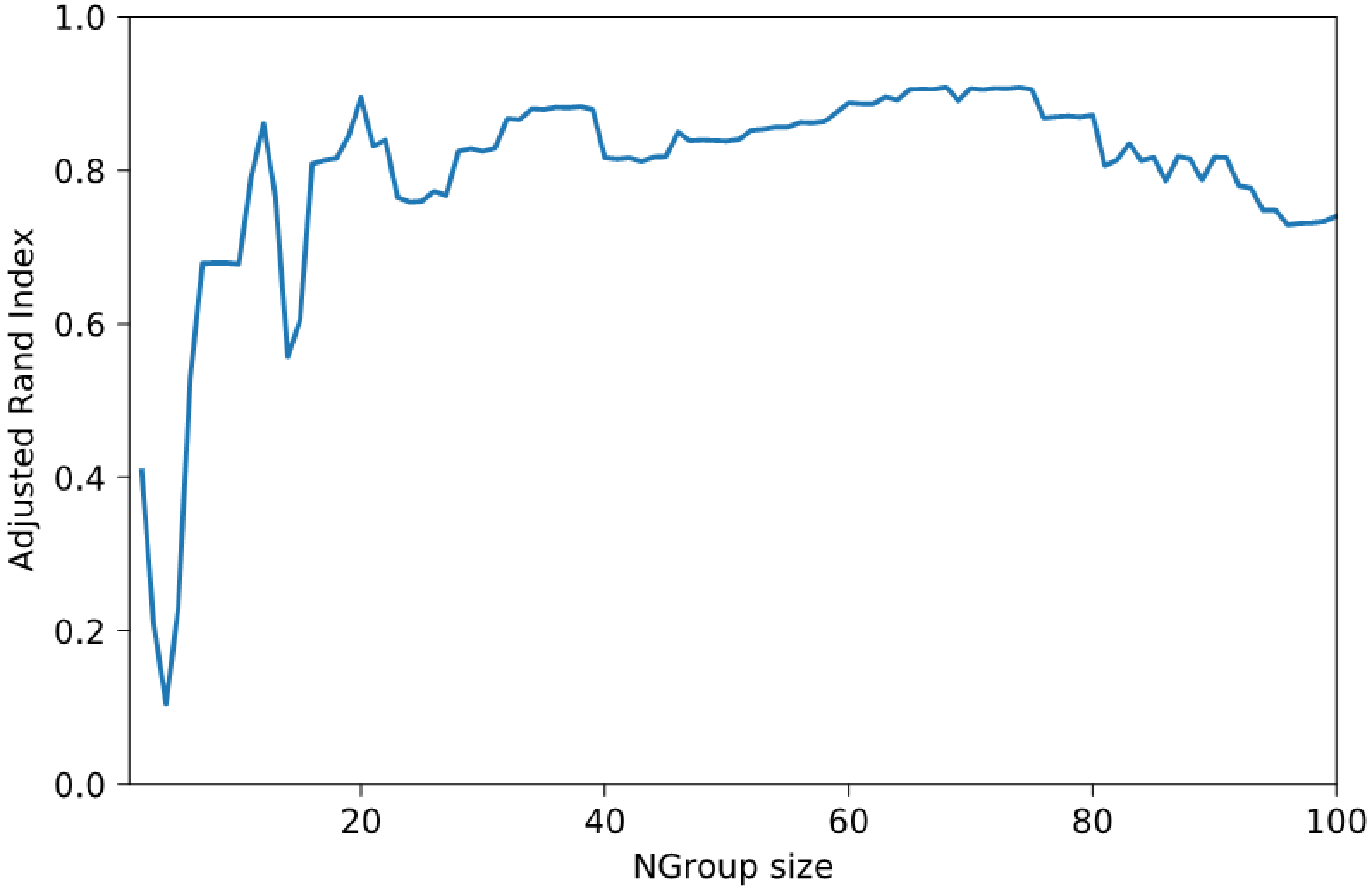
| | MLST | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Isolate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | NGroup |
| A | 4 | 20 | 22 | 1 | 53 | 23 | 1 | ? |
| B | 4 | 20 | 42 | 1 | 30 | 2 | 3 | ? |
| C | 2 | 3 | 51 | 1 | 2 | 4 | 3 | ? |
| D | 15 | 2 | 23 | 1 | 42 | 10 | 52 | ? |
| E | 4 | 20 | 5 | 1 | 20 | 23 | 1 | Green |
| F | 20 | 10 | 34 | 1 | 42 | 4 | 11 | Orange |
| G | 3 | 1 | 29 | 1 | 30 | 2 | 3 | Pink |
| H | 2 | 3 | 25 | 1 | 32 | 4 | 23 | Blue |

| | MLST | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Isolate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | NGroup True | NGroup Predicted | NGroup Probability |
| A | 4 | 20 | 22 | 1 | 53 | 23 | 1 | Green | Green | 0.85 |
| B | 4 | 20 | 42 | 1 | 30 | 2 | 3 | Green | Pink | 0.4 |
| C | 2 | 3 | 51 | 1 | 2 | 4 | 3 | Blue | Blue | 0.9 |
| D | 15 | 2 | 23 | 1 | 42 | 10 | 11 | Orange | Orange | 0.87 |
| ↓ | ... all other testing isolates (20%) | | | | | | | | | |

**Adjusted Rand Score** **0.89**

## Figure 1

Schematic illustration of the NeighbourGroup (NGroup) classification pipeline. Two phylogenetic trees are constructed: (1) the "True Tree", which is reconstructed from all the isolates and (2) the training tree, which uses 80% of the isolates. Isolates A, B, C, and D are test isolates excluded from the training tree, for which the model predicts Neighbour Group (NGroup) membership based on a limited number of loci. Isolates E, F, G, and H are training isolates assigned to Neighbour Groups. The table shows the results obtained from the model, with predicted NGroup membership as established by the "true tree" and predicted from the training tree. Neighbour Group membership probability is estimated and reflects a level of uncertainty in the classification. Adjusted Rand score shows how the model has performed: a score of 0 indicates the performance is no better than chance, and a score of 1 indicates perfect classification.



## Figure 2

Variation in adjusted Rand Index with number of NGroups for a dataset of 10,359 *Campylobacter jejuni*WGSs. The adjusted Rand score is shown for 2 to 100 NGroups. Model performance was poor (low adjusted Rand Index) with fewer NGroups and reached a maximum (0.895) for 20 NGroups, with little added value for higher numbers of NGroups.

**Statement of Authorship for joint/multi-authored papers for PGR thesis**
To appear at the end of each thesis chapter submitted as an article/paper

# Chapter 3:

| | |
|---|---|
| Title of Paper | NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters |
| Publication Status | ☐Published ☐ Accepted for Publication<br>☒Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Veltcheva, D., Richer, S., Sheppard, S., Varga, M., Colles, F., Bonsall, M. and Maiden, M., 2023. NeighbourGroups: a machine learning classification tool that assigns microbial multi-locus genotypes to clusters.<br>Paper Style: Written for Nature Methods: Brief Communication format (Currently Published as Pre-print) |

Student Confirmation

| | |
|---|---|
| Student Name: | Dessislava Veltcheva |
| Contribution to the Paper | • Conceptualization of the research project<br>• Establishing Methodology<br>• Software and script creation<br>• Formal analysis and Investigation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project management |

| | | |
|---|---|---|
| Signature | Date | 02.10.2023 |

Supervisor Confirmation

| |
|---|
| Supervisor name and title: Professor Michael Bonsall |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. |

| | | |
|---|---|---|
| Signature. | Date | 2nd October 2023 |

# Chapter 4: CatBoost classifier "NeighbourGroups" classification applied to *Campylobacter jejuni*

CatBoost classifier "NeighbourGroups" classification applied to *Campylobacter jejuni*

Dessislava Veltcheva, *Stephen Richer, Margaret Varga, Frances Colles, Michael B. Bonsall, Martin C. J. Maiden

***Affiliation*** University of Oxford, *University of Bath

***Corresponding author***: dessislava.veltcheva@spc.ox.ac.uk

***Keywords****: Campylobacter jejuni, multi-locus sequence typing,* clonal-complex, CatBoost algorithm, minimum-spanning tree, classification, taxonomy, machine learning, artificial intelligence

***Repositories***: *https://github.com/bgrdessislava/NeighbourGroups*

## Abstract

*Campylobacter jejuni* is a widespread and highly diverse gram-negative zoonotic bacterium, which is a leading cause of bacterial food poisoning worldwide. Here we applied “NeighbourGroups”, a generalisable machine learning approach that reproducibly, robustly, and rapidly classifies multi-locus sequence types with high precision. We found that 20 NeighbourGroups were optimal to describe the *C. jejuni* diversity. From the availability of whole genome data and the new classification of NeighbourGroups, it is possible to compare the NeighbourGroups to Clonal-complex. The clonal-complex classification is formed by assigning clonal-groups using central genotypes and seven locus MLST data. NeighbourGroups can be classified using the same seven loci, but the classification algorithm was derived from cgMLST data and a newick file. Here we explore how this improved classification system applies on *C. jejuni*. As a result, we demonstrate

that some clonal-complex could be merged into the same NeighbourGroups, such as CC-21 is composed of two NeighbourGroups. Our result indicates that NeighbourGroups can be used to uncover variant-level relations using cgMLST and distinguish discrepancies that were not easy to identify using clonal-complex.

### Data Summary

The data described are available on the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018) and can be accessed with publicly available project No.110: 1998-2018 UK human UK isolates (n=10,359). https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1.

### Gap statement

The clonal-complex classification was established before whole genome data were available. However, now that next generation sequencing techniques have been evolving, we have isolates with cgMLST (1343 loci). Here we established NeighbourGroups for *Campylobacter jejuni*, a classifier built from a large cgMLST dataset that can classify isolates more accurately using the MLST (multi-locus sequence typing) typing scheme.

### Impact Statement

The MLST classification method was recalibrated to classify *Campylobacter jejuni* isolates into groups inferred from cgMLST data (1343 loci). This work compares the NeighbourGroups classification scheme to the clonal-complex classification and represents a generalisable technique appropriate for application to other microbial populations. Through this process we demonstrate which variants needs to change as clonal-complex. This process of re-investigating the golden standard has not been done before.

## Introduction

### *Campylobacter jejuni*

*Campylobacter* is a diverse bacterial genus and a common cause of food poisoning worldwide; *C. jejuni* causes 90% of infections, and *C. coli* causes most of the remaining cases (Jehanne *et al.*, 2020). The application of whole gene sequencing (WGS) has led to the development of cgMLST (1343 loci), which are loci that are known to be present in 95% of *Campylobacter jejuni and Campylobacter coli* (Cody *et al.*, 2017). Multi-locus sequence typing (MLST) is currently used to investigate the epidemiology of *Campylobacter jejuni (C.jejuni)* (Nielsen *et al.*, 2010; Yahara *et al.*, 2017; Arning *et al.*, 2021). This classification has also helped establish clonal groups called clonal-complexes (Maiden, 2006).Clonal-complexes are defined using three heuristic methods: the BURST algorithm, UPGMA cluster analysis and SPLITSTREE (Wareing *et al.*, 2003). These central genotypes are widely found worldwide, such as ST types: 21, 464, 61 etc. New MLST sequence types are assigned to the clonal-complexes when four or more of their seven loci match the central genotypes (Wareing *et al.*, 2003). Until 2021, 34 clonal-complexes for *C. jejuni* and 10 for *C. coli* have been defined. Since the 2021, we have also established one more clonal-complex, 581, which has been added to classify 35 clonal-complexes for *Campylobacter jejuni*.

## NeighbourGroups

Previously, we demonstrated a new machine learning-based classification called "NeighbourGroups" (available here: https://github.com/bgrdessislava/NeighbourGroups) (Veltcheva *et al.*, 2023). This new approach improved the current scheme by defining hierarchical clustered groups on the 1343 cgMLST loci (Cody *et al.*, 2017). First, cgMLST data were analysed using the minimum spanning tree construction method to establish clusters or 'NeighbourGroups' based on the similarity of their cgMLST profiles. Then, a supervised machine learning algorithm was used to optimally predict the membership of these clusters from fewer loci, for example, seven MLST loci. The trained algorithm enables a probabilistic assignment of a seven-locus genotype to a cluster defined with cgMLST data. This method will benefit both historical isolates with MLST locus sequenced but did not have cgMLST sequenced.

Here the NeighbourGroups approach was applied onto *Campylobacter jejuni*. It enables comparisons with clonal-complexes, and NeighbourGroups for an enhanced understanding of *Campylobacter jejuni* diversity and population structure.

## Materials and Methods

### Meta Data Acquisition

All data are publicly available from the PubMLST database (Jolley, Bray and Maiden, 2018). High-quality sequencing and metadata, including location and source of isolation for the 10,359 *Campylobacter jejuni* isolates from the UK during 1990 – 2022, were downloaded from PubMLST {https://publmlsts.org} using the following search query: *'Species=Campylobacter jejuni' AND 'Country=UK' AND '1990>=Year<=2020' AND 'N50>=20,000' AND '1.4Mb< =Genome Size <= 1.8 Mb' AND 'Contigs<=50' AND 'source=human_stool'*

*The output of this query Is available at:* https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=110&submit=1.

### Software usage

Scripts were written using the Python 3 programming language (Van Rossum, G. , Drake, 2009). The minimum spanning model was built with GrapeTree using the PubMLST extension (Zhou *et al.*, 2018) and visualised using iTOL (Letunic and Bork, 2021). Graphs and figures were built using Tableau ('Tableau (version. 9.1)', 2016). Tables were created using Microsoft Excel (version 16.63). NeighbourGroups code is available at https://github.com/bgrdessislava/NeighbourGroups.

## Results

### 1.) *Campylobacter jejuni* can be divided into 20 NeighbourGroups.

As previously described, a minimum-spanning tree was built from the UK isolates using all 1343 cgMLST loci (Veltcheva *et al.*, 2023). A set number of clusters were extracted from the hierarchical clustering using the *fcluster* function (criterion = ‘maxclust’) from the SciPy library (Virtanen *et al.*, 2020). The NeighbourGroups algorithm was then trained to recover these clusters from the seven MLST loci. Classification performance was assessed using the adjusted Rand Index on unseen isolates. The adjusted Rand Index compares the similarity between the two classifications. A score of 0 is equivalent to a random label assignment, and a score of 1 is equivalent to a perfect match. The model performed optimally (adjusted Rand Index = 0.895) when trained on 20 groups (Supplementary: Figure 9). Consequently, 20 groups describe the population structure that can sufficiently represent the diverse classification to support the understanding of *Campylobacter jejuni* population. Furthermore, for each prediction, NeighbourGroups generates a probability score that can be used to assess the level of confidence associated with the prediction.

### 2.) Comparing NeighbourGroups with clonal-complex classification

*C.jejuni* was classified into 20 groups with the NeighbourGroups classification, whereas we have 35 clonal-complex groups. The adjusted Rand Index of NeighbourGroups deriving from the cgMLST minimum spanning Tree (20 groups) was 0.895. While the clonal-complex classification had an adjusted Rand Index of 0.588. Thus, NeighbourGroups classification can more accurately recover the relationships inferred from a minimum spanning tree built from 1343 cgMLST loci, which still only uses the canonical 7 MLST loci to make the predictions.

In both classifications, the group sizes were imbalanced. The three largest groups in both classifications contain around >50% of all isolates. The largest clonal-complex21, includes 28.5% of all isolates compared to 19.9% in the largest NeighbourGroup 15 (Supplementary: Figure 10).

Next, we investigated the representation of clonal-complexes within the NeighbourGroups (Figure 5). 22 of 35 clonal-complexes were consistently assigned to the same NeighbourGroup. For example, all isolates of CC22 were assigned to NGroup 14 (Figure 6). Further, five clonal-complexes, >95% of all isolates, were assigned to the same NeighbourGroup (Figure 6). However, only one classification pair (CC-443 to NGroup2) were in complete mutual agreement. (Supplementary:
Figure 11). Specifically, all isolates of CC443 were assigned to NG2, and NG2 only contained CC443 isolates. The isolates of the remaining eight clonal-complexes (CC21, CC353, CC362, CC433, CC574, CC581 and CC607) were assigned to multiple NeighbourGroups (Figure 6). Notably, the CC21 isolates were approximately evenly split between the NG15 and NG16. These findings demonstrate that CC21 comprises of two distinct variants, commonly denoted as source generalist (Sheppard *et al.*, 2014; Dearlove *et al.*, 2016). Our results indicate that CC21 is a hybrid of two genetically

different populations that currently through clonal-complex is classified into the same group

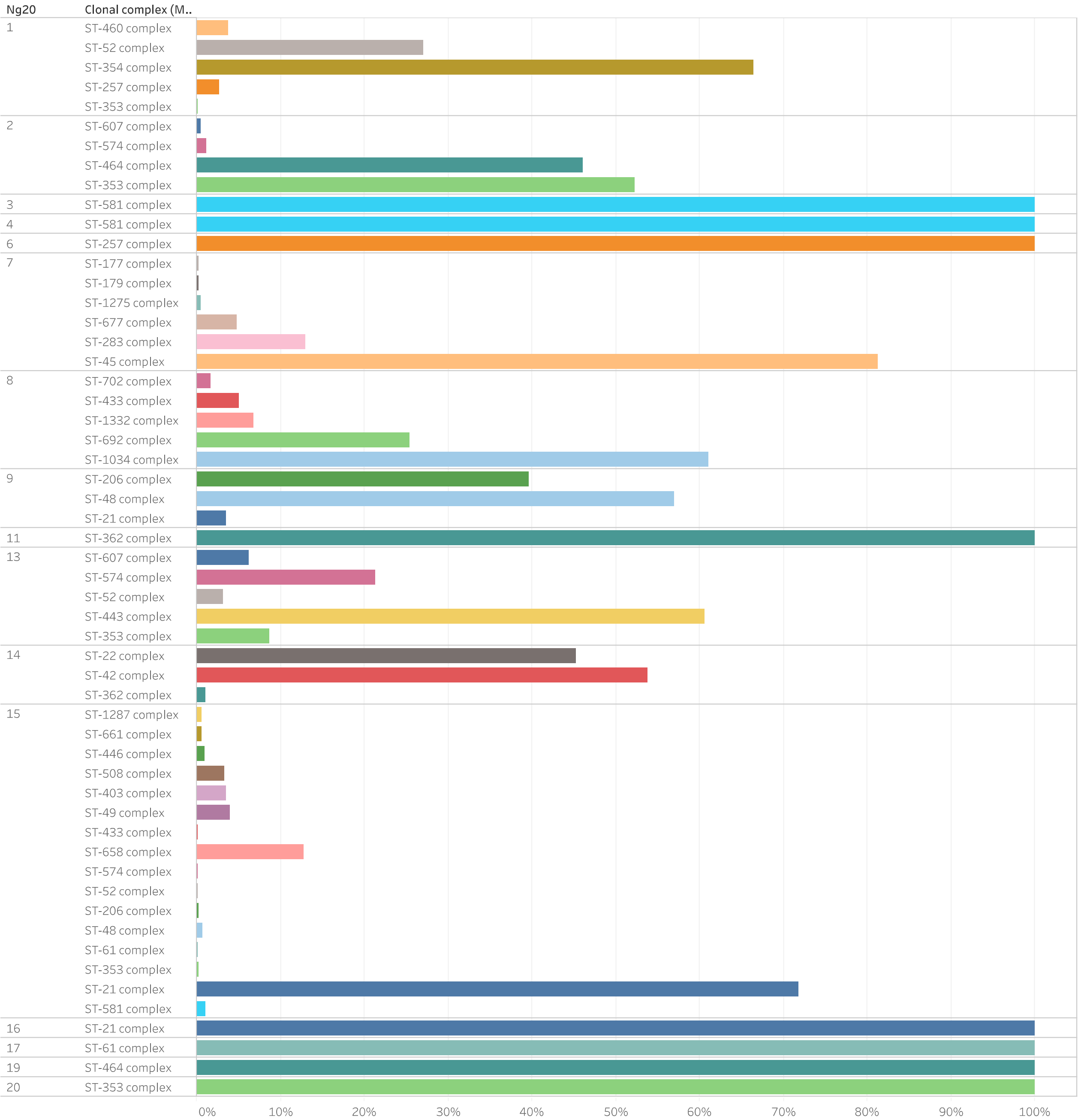


*Figure 5 NeighbourGroups and their clonal-complex isolates composition. The colour indicates the clonal-complex. Each of the NeighbourGroups adds up to 100%.*

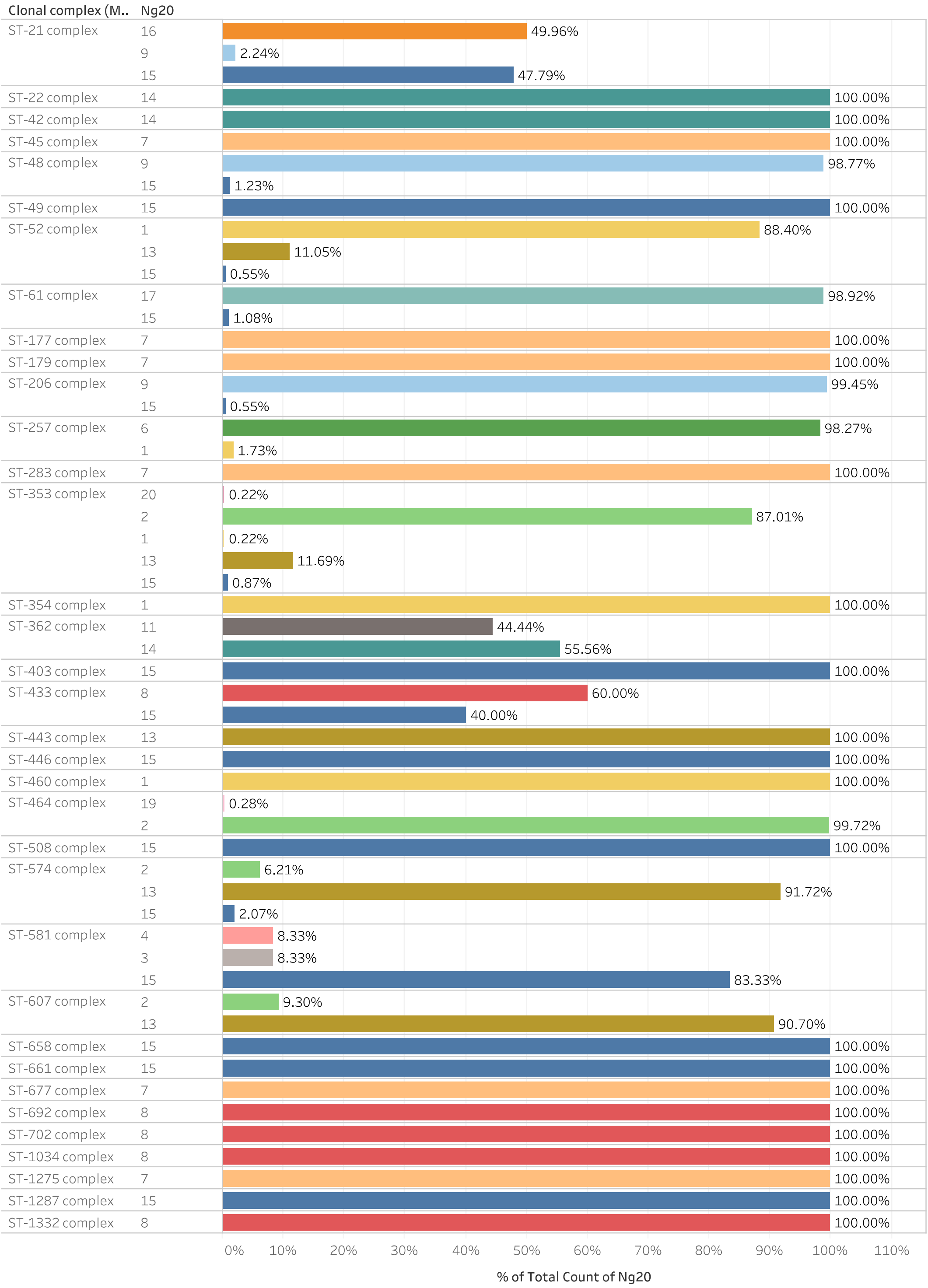


*Figure 6 Bar chart showing clonal-complex and their NeighbourGroups assignments. The colour indicates the NeighbourGroups. Each of the clonal-complex adds up to 100%.*

NG15 was composed of multiple isolates from numerous clonal-complexes with CC21 taking >70% (Figure 5). A total of eight NGroups directly correlated to one clonal-complex, showing that these clonal-complexes with neighbour groups are related to the underlying cgMLST. Furthermore, we can also see some NGroups, such as 1, 8, and 15, were composed of multiple clonal-complexes but often with one dominant clonal-complex (Figure 5). Moreover, we can visualize the connection between the NGroups and clonal-complexes using networks (Supplementary: Figure 12). In this figure, the node sizes represent the sample sizes, the red nodes represent clonal-complexes, and the blue nodes represent NeighbourGroups.

To understand the application of the NeighbourGroups further, we can view the assignments of the NeighbourGroups to *Campylobacter jejuni* isolate (Figure 7). In the inner band, we can see how NeighbourGroups are separated into their groups. The coloured bar shows CC21, split between two NeighbourGroups, 15 and 16. Finally, for CC353, we can see two major groups in NeighbourGroups 2 and 13, and some are also located in NG15 and NG20; this disentangles the complexity of this clonal-complex across different NGroups.

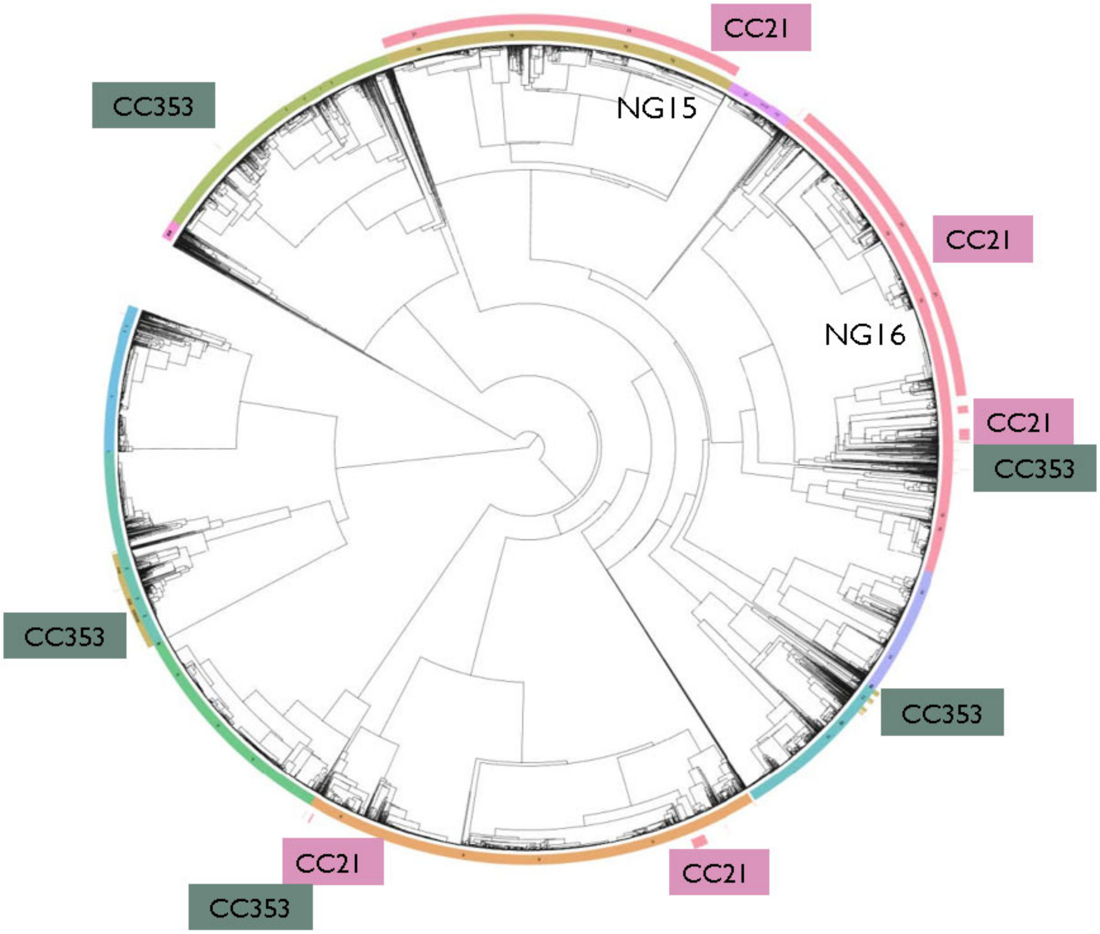


*Figure 7 10359 isolates NeighbourGroups illustrated in the inner ring, and the outer ring includes the CC353(olive colour) and CC21 (red colour), which have been found around the iTOL illustrated cgMLST tree*

### 3.) Single locus prediction of NGroups

As with Clonal-complexes, the NeighbourGroups classifier was built to classify isolates from the *Campylobacter* MLST scheme. This scheme uses internal fragments of the seven house-keeping genes (*gltA*, *uncA*, *tkt*, *aspA*, *glyA* and *pgm*). However, we have also sought to assess how the NeighbourGroups classification algorithm performed when applied to cgMLST scheme and each of their loci.

First, we compared the classification performance when using the complete cgMLST typing of the seven housekeeping genes versus the MLST typing. The adjusted Rand Index of the cgMLST typing was 0.899, only a marginal improvement (+0.004) compared to the MLST gene fragments, indicating that the seven MLST fragments are sufficient for *Campylobacter jejuni* classification. If a single locus was considered, the full-gene representation was more effective for predictive classification ( Table 2). For example, the *tkt* locus alone has a Rand Index of 0.854 compared to a Rand Index of 0.356 for the *tkt* MLST fragment (Table 2).

We then assessed the classification accuracy using combinations of seven MLST gene fragment loci (Figure 8). No rigorous assessment has determined whether seven is the optimal number of loci or whether similar performance can be achieved with fewer loci. Historically, seven loci were the number that worked effectively with *Neisseria meningitidis,* the first organism for which MLST was developed (Maiden *et al.*, 1998). For further study, we compared the mean adjusted Rand score for different sets of MLST loci (Figure 8). As expected, classification performance increases when more loci are used for the prediction (Figure 8). However, the improvements in prediction accuracy diminished with each additional loci until the saturation point was reached. Classification performance was higher when the set of loci included *gltA. However, gltA alone* had poor predictive power, high classification performance could be achieved with fewer MLST loci if *gltA* was included. *GltA* gene provides distinct and complementary information for predicting the *Campylobacter jejuni* classification groups.

However, with the current availability of 1343 cgMLST loci, we sought to identify if any other loci were superior candidates for classification. We utilised the NeighbourGroups algorithm to build a classification system for each locus, the adjusted Rand scores for the top-performing individual loci (Supplementary: Figure 13). In particular, a conserved hypothetical protein Cj1009c (CAMP0932) has an adjusted Rand score of 0.91. This locus alone is more effective at recovering the cgMLST-defined classification than the seven MLST housekeeping genes. In total, 26 loci were sufficient to provide an adjusted Rand Score of greater than 0.85 just with one locus. One of these is the full-gene representation of *tkt.* Thus, we have shown that, combinations of top-scoring loci can further improve predictive performance (Supplementary: Figure 14).

*Table 2 Scores between the whole loci of the 7 MLST vs the reduced base pairs of MLST used for current classification. Green indicates the adjusted rand score higher than 0.8. uncA is also known as atpA.*

| **Loci** | **Gene** | **Score - MLST (complete gene)** | **Score - MLST (Fragmented)** | **Percentage Increase in adjusted Rand score using complete gene** |
|---|---|---|---|---|
| CAMP1541 | tkt | 0.85 | 0.36 | 58.37 |
| CAMP1576 | gltA | 0.73 | 0.35 | 52.44 |
| CAMP0399 | pgm | 0.70 | 0.48 | 31.41 |
| CAMP0093 | atpA - uncA | 0.49 | 0.35 | 27.36 |
| CAMP0367 | glyA | 0.56 | 0.43 | 24.28 |
| CAMP0075 | aspA | 0.45 | 0.37 | 18.16 |
| CAMP0645 | glnA | 0.64 | 0.53 | 16.13 |

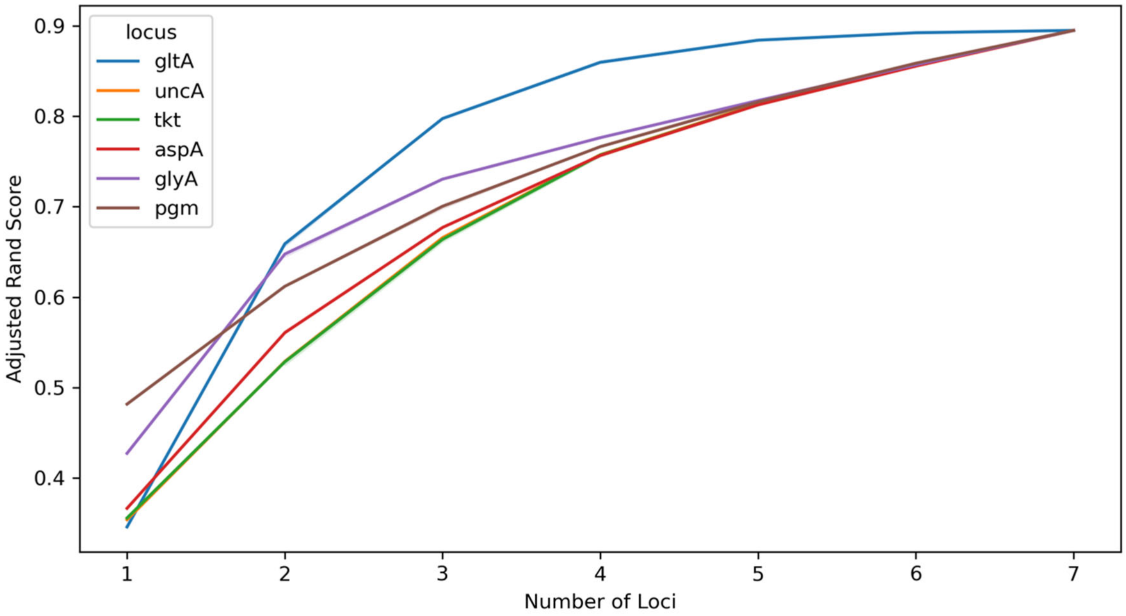


*Figure 8 MLST combination vs adjusted rand score to predict 20-NGroup. 1 and 7 are only one point, but 2 – 6 are different sample sizes. These loci are reduced loci sizes rather than using the whole gene.*

## Discussion

Here the application of NeighbourGroups to a large *Campylobacter jejuni* next-generation sequencing dataset has been shown. Furthermore, a method to improve the model by identifying other loci that could contribute to better model predictions have been identified. We further discuss how the model could be enhanced and determine the next steps for NeighbourGroups. Early classification systems, such as clonal-complexes, were developed before the availability of whole genome sequencing. However, clonal-complexes are used as a first step for the surveillance of epidemiological studies worldwide. Thus having this classification robust is very important. Furthermore, although in this study we used cgMLST data, the reference classification does not need to be these set of loci. Moreover, this model could be developed from any other tree algorithms.

Nevertheless, our results suggest that the seven housekeeping MLST gene fragments effectively defined a representative classification. In this study, we did not investigate branch lengths to dive further into the emergence of specific variants. Although, unless done in a controlled experiment, this will be difficult to be investigated, applying branch lengths can give us temporal dimensions and outbreak dynamics: isolates responsible for outbreaks of campylobacteriosis are expected to have shorter branch lengths, and this variation in branch length could allow us to determine when different variants have emerged across time (Calland *et al.*, 2021).

Another critical step is how many NeighbourGroups are enough per variant/species. To establish the NeighbourGroups, we must train the models using as many isolates as possible. Thus, adding a probability to the outcome helps support our certainty level. However, the probability we produce is not "True Probability" since it can only be computed from our training data set. Expanding the training to a model that includes *Campylobacter coli* will potentially give us insight into some of the clades that are known to have co-evolved from *Campylobacter jejuni*. However, the current low probability scores will be able to provide us with an insight onto the variant that needs further investigation. Furthermore, at this point, we have used 20 groups as the optimal number of clusters using the adjusted Rand index. However, we could also try to cross-validate this approach through different methods.

We have also demonstrated that although these seven MLST loci have been chosen as housekeeping genes, we have another locus within the cgMLST that can predict the

NeighbourGroups with a higher degree of accuracy. These are CAMP0932-conserved hypothetical protein cj1009c, CAMP0471-putative secreted protease, CAMP0308-excinuclease ABC subunit A. Therefore, it will be essential to understand which loci could be used to improve or renew these “housekeeping” genes or even one locus.

Furthermore, using these NGroups, we can re-analyse trend analysis on antimicrobial resistance across time. A previous study showed that some clonal-complexes resist fluoroquinolones over time; some are susceptible, some have increased, and some have a less clear signal (Veltcheva *et al.*, 2022). This classification could potentially lead to the diagnostics of AMR, which could predict antimicrobial resistance by training the model. Through the usage of NGroups, we will not only be able to identify variants, but it will allow us to be backwards compatible with samples from the past.

## Conclusion

We have established a new tool called “NeighbourGroups” that can represent the population structure of *Campylobacter* using machine learning. For *Campylobacter jejuni,* we have shown that the optimal NeighbourGroups number should be 20, having the highest-level adjusted Rand score. Although we have also shown here that if we want to increase this prediction, we could do this by choosing other loci that could improve the prediction of NGroups. Furthermore, by selecting the full length of the seven housekeeping genes, loci will increase the prediction score marginally better.

The comparisons between clonal-complex and NeighbourGroups revealed three groups. The first group (22 out of 35 clonal-complexes) exhibited clonal-complexes that matched 100% onto other NeighbourGroups. The second group shows a clonal-complex to fit more than one Neighbour Group but exhibited one majority of NeighbourGroups above 98%, namely CC48, CC61, CC206, CC257 and CC464 (Figure 6). The third group (CC21, CC52, CC353, CC362, CC433, CC574, CC581, and CC607) shows clonal-complexes with more than two NeighbourGroups assigned, illustrating the necessity of clonal-complexes to be used cautiously for this group (Figure 6). Another aspect of the study is that by reducing the groups from 35 to 20, we now see groups not in the same clonal-complex being categorised together. As a result, cgMLST showing similarity that the clonal-complex was not about to exhibit. NGroups for *C. jejuni* is the first step of applying machine learning to extensive genomic data which can provide predictions onto which groups each isolate belongs. Furthermore with the similar investigation we could predict potential AMR targets, new house keeping genes, new one locus that leads to NGroups or even a new 16S level one locus that could potentially distinguish species.

# Supplementary

| Ngroup | Score | Ngroup | Score |
|---|---|---|---|
| 2 | 0.40827736 | 50 | 0.83807902 |
| 3 | 0.20983507 | 51 | 0.84048461 |
| 4 | 0.10551856 | 52 | 0.85178583 |
| 5 | 0.22952602 | 53 | 0.85353658 |
| 6 | 0.52861425 | 54 | 0.85598275 |
| 7 | 0.67919128 | 55 | 0.85590628 |
| 8 | 0.6798183 | 56 | 0.86210193 |
| 9 | 0.6798183 | 57 | 0.86150623 |
| 10 | 0.67796133 | 58 | 0.86335587 |
| 11 | 0.79159408 | 59 | 0.87556818 |
| 12 | 0.86059301 | 60 | 0.88777918 |
| 13 | 0.76514353 | 61 | 0.88647519 |
| 14 | 0.55715021 | 62 | 0.88625667 |
| 15 | 0.60510142 | 63 | 0.89569922 |
| 16 | 0.80823678 | 64 | 0.89152318 |
| 17 | 0.81322201 | 65 | 0.90544889 |
| 18 | 0.81540964 | 66 | 0.90581615 |
| 19 | 0.84613636 | 67 | 0.90578375 |
| 20 | 0.89504879 | 68 | 0.90843842 |
| 21 | 0.8314469 | 69 | 0.89064303 |
| 22 | 0.83989327 | 70 | 0.90646193 |
| 23 | 0.7649696 | 71 | 0.90511239 |
| 24 | 0.75893961 | 72 | 0.90678384 |
| 25 | 0.76014485 | 73 | 0.90641045 |
| 26 | 0.77264802 | 74 | 0.90814885 |
| 27 | 0.76718815 | 75 | 0.90516156 |
| 28 | 0.82459371 | 76 | 0.86787859 |
| 29 | 0.8283788 | 77 | 0.86951122 |
| 30 | 0.82463245 | 78 | 0.87048435 |
| 31 | 0.82915758 | 79 | 0.86919086 |
| 32 | 0.86777934 | 80 | 0.87186775 |
| 33 | 0.86595633 | 81 | 0.80569554 |
| 34 | 0.88002879 | 82 | 0.81357443 |
| 35 | 0.87915667 | 83 | 0.8349409 |
| 36 | 0.88244544 | 84 | 0.81259902 |
| 37 | 0.88176135 | 85 | 0.816834 |
| 38 | 0.88312833 | 86 | 0.78570376 |
| 39 | 0.87898572 | 87 | 0.81784632 |
| 40 | 0.81647959 | 88 | 0.81473288 |
| 41 | 0.81442053 | 89 | 0.78717943 |
| 42 | 0.81613952 | 90 | 0.81683332 |
| 43 | 0.81148115 | 91 | 0.81640482 |
| 44 | 0.81727864 | 92 | 0.77975535 |
| 45 | 0.81785965 | 93 | 0.77668535 |
| 46 | 0.84959744 | 94 | 0.74780856 |
| 47 | 0.83852943 | 95 | 0.74791828 |
| 48 | 0.83959824 | 96 | 0.7293995 |
| 49 | 0.83894135 | 97 | 0.73109918 |
|  |  | 98 | 0.73168577 |
|  |  | 99 | 0.73344705 |
|  |  | 100 | 0.74023367 |

*Figure 9 Adjusted rand score with different NGroup sizes. Yellow shows the Ngroup 20 value that has been chosen. Blue highlighted 35 indicates the current group sizes for Campylobacter whichhas lower Adjusted Rand score. The 68th group has the highest Adjusted Rand index but not used since we wanted to find lower number of groups than 35.*

We can see that some of the loci already had high prediction scores (Supplementary: Figure 13) and have an even higher adjusted rand score when combined with another locus. Moreover, we see some loci which did not have above 0.85 rand score on their own but showed an excellent match to improve the score, such as atpD (CAMP0095another locus and CAMP0711 Putative perplasmic protein. We have demonstrated that; with the careful choice of cgMLST loci combinations, we can increase the adjusted rand score further.

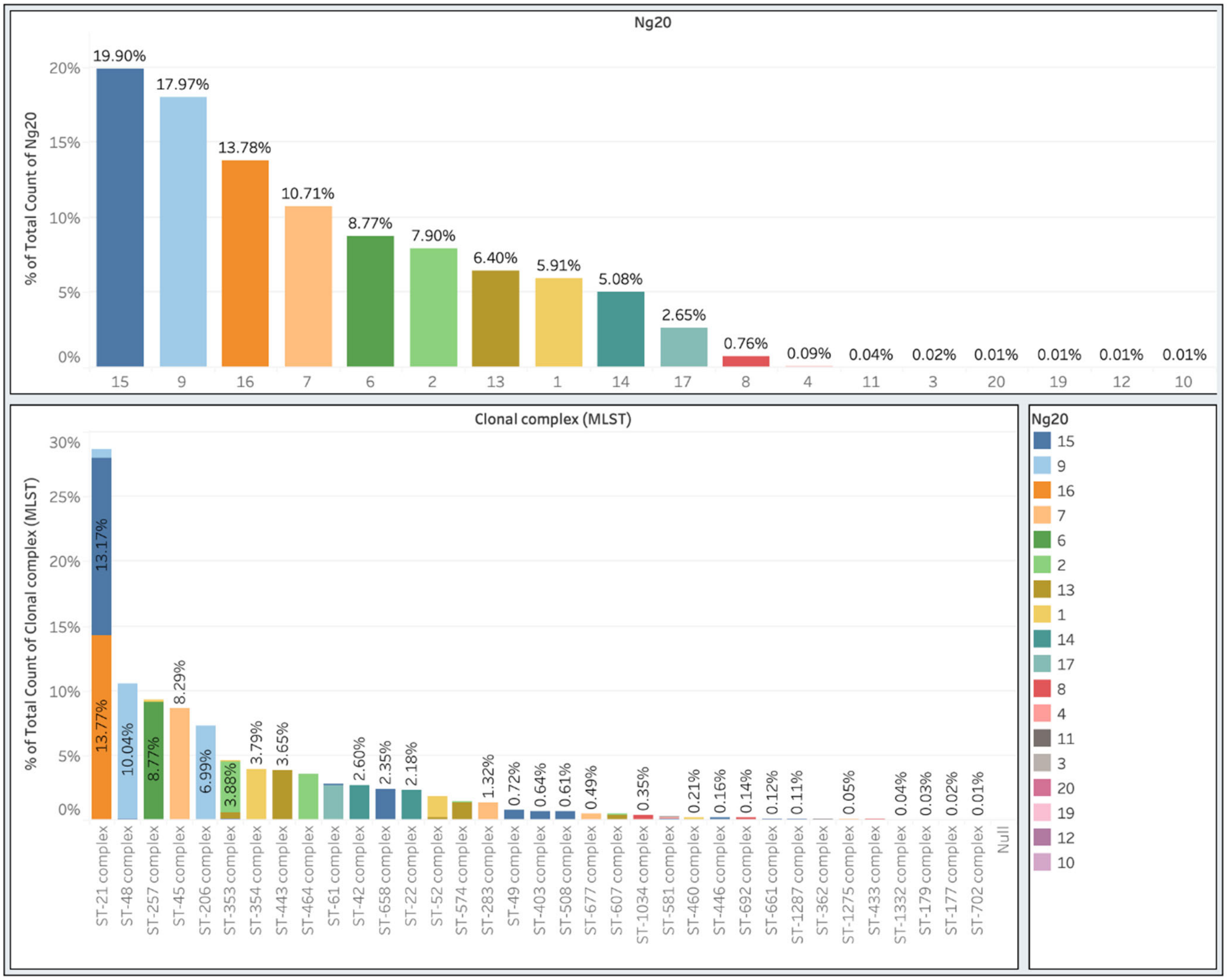


*Figure 10 Comparing 20-NeighbourGroups classification proportions to Clonal-complex.*

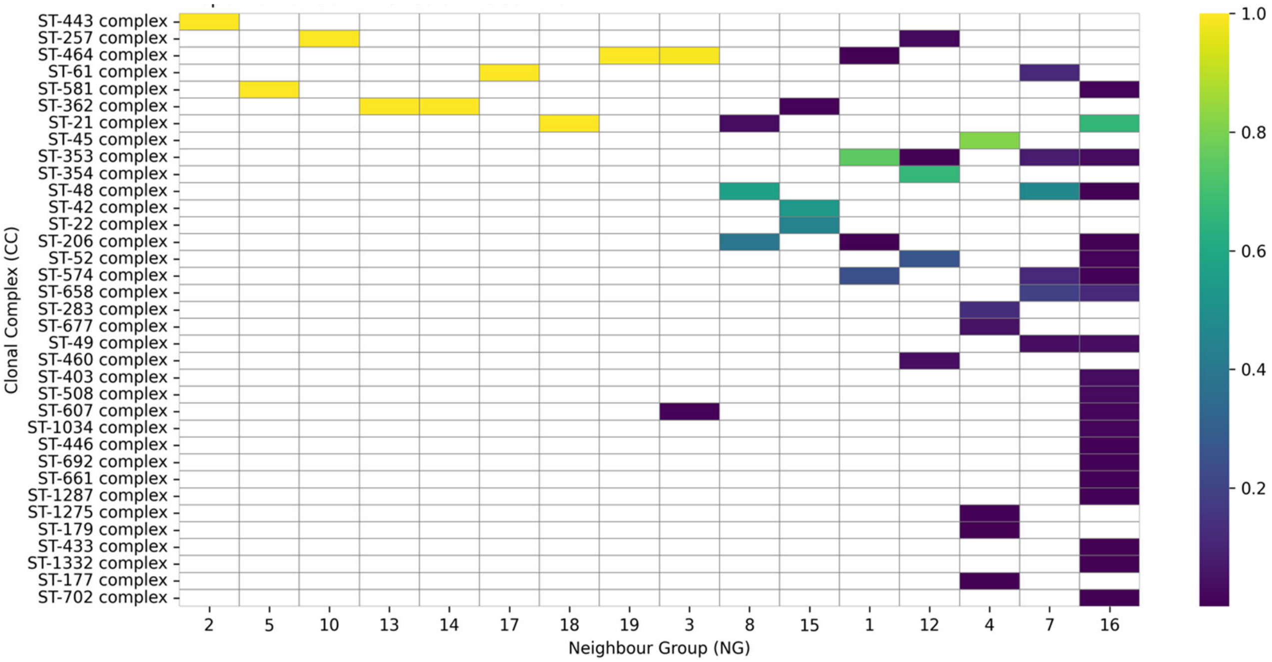


*Figure 11 20-NeighbourGroups to Clonal-complex comparisons using proportions. The result is vertically equal to 100%, illustrating some NeighbourGroups are formed by one clonal-complex [2,5,10,13,14,17,18 and 19], whereas some NeighbourGroups are composed of multiple clonal-complexes.*

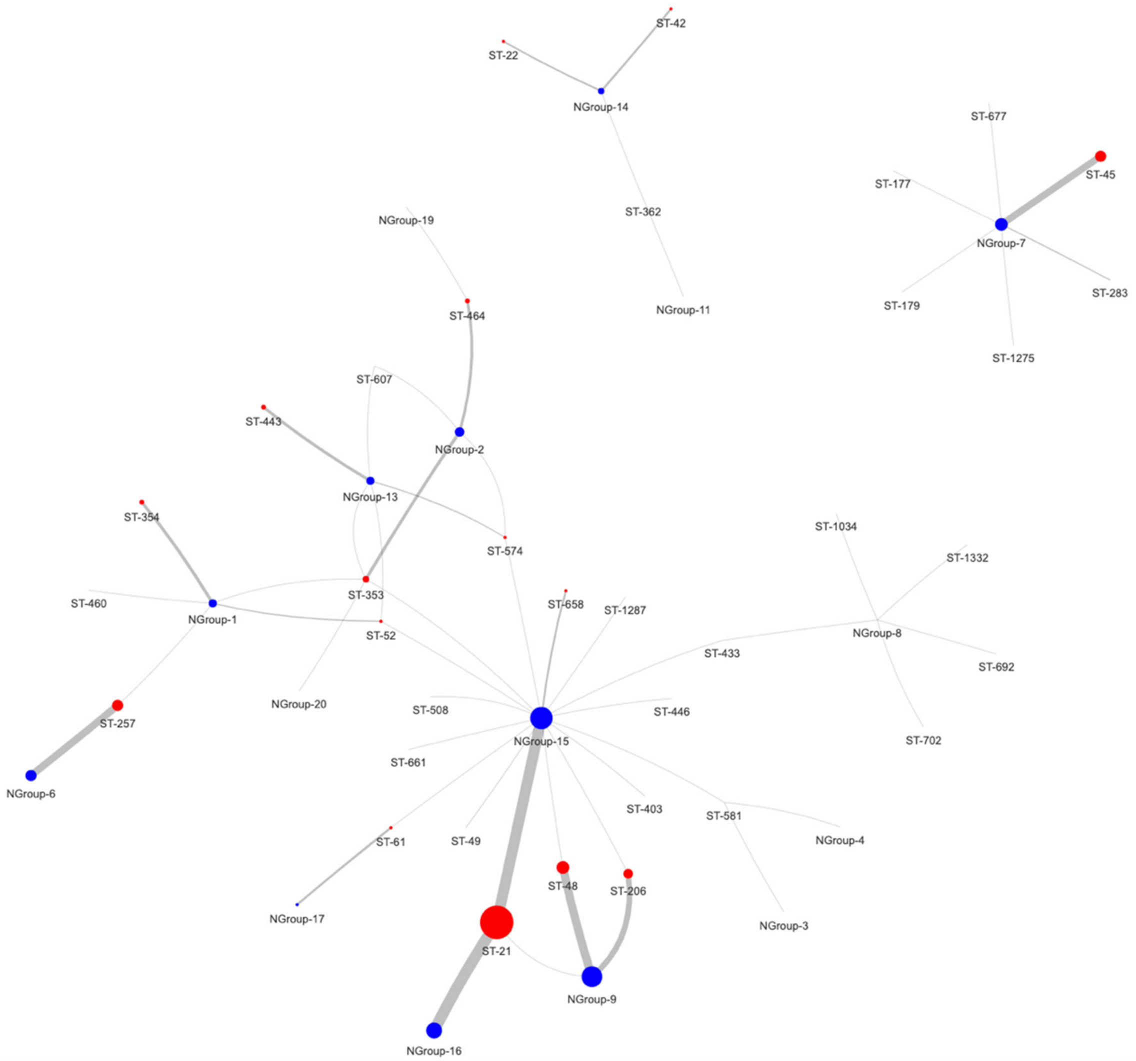
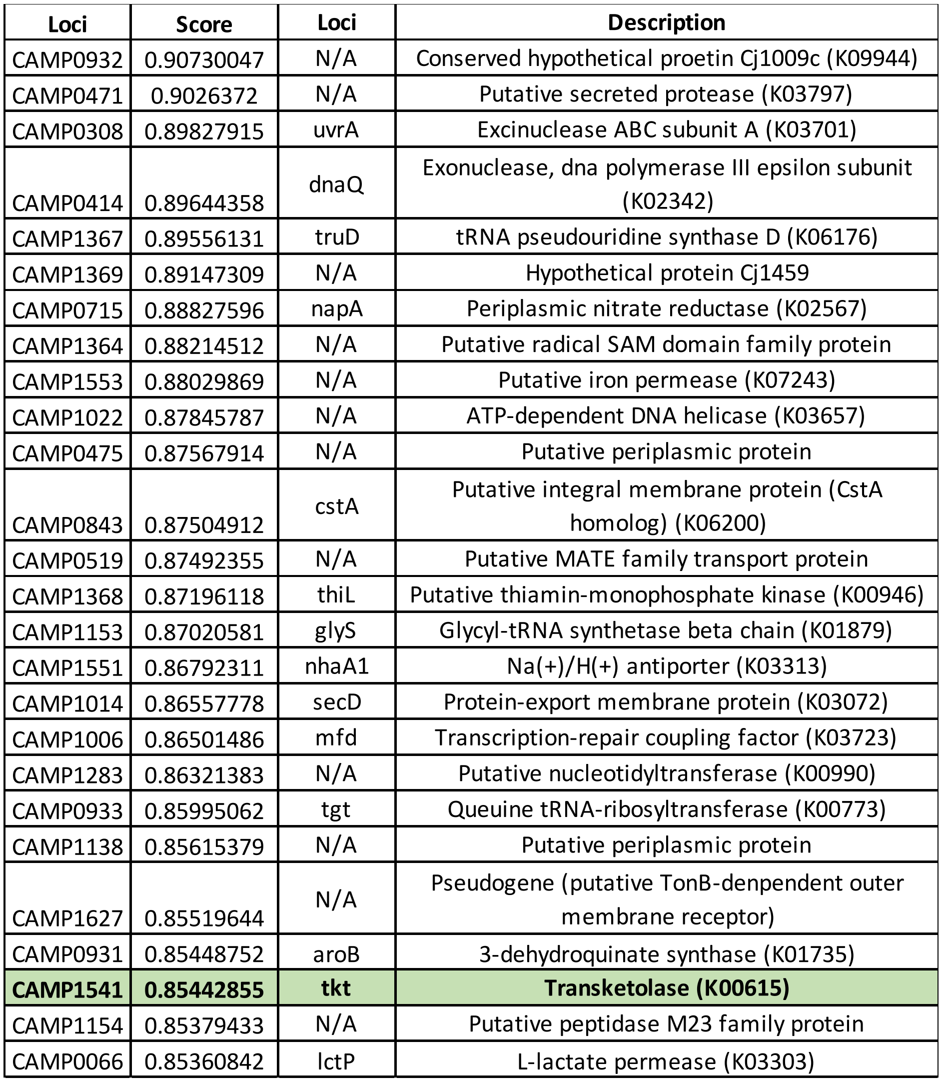
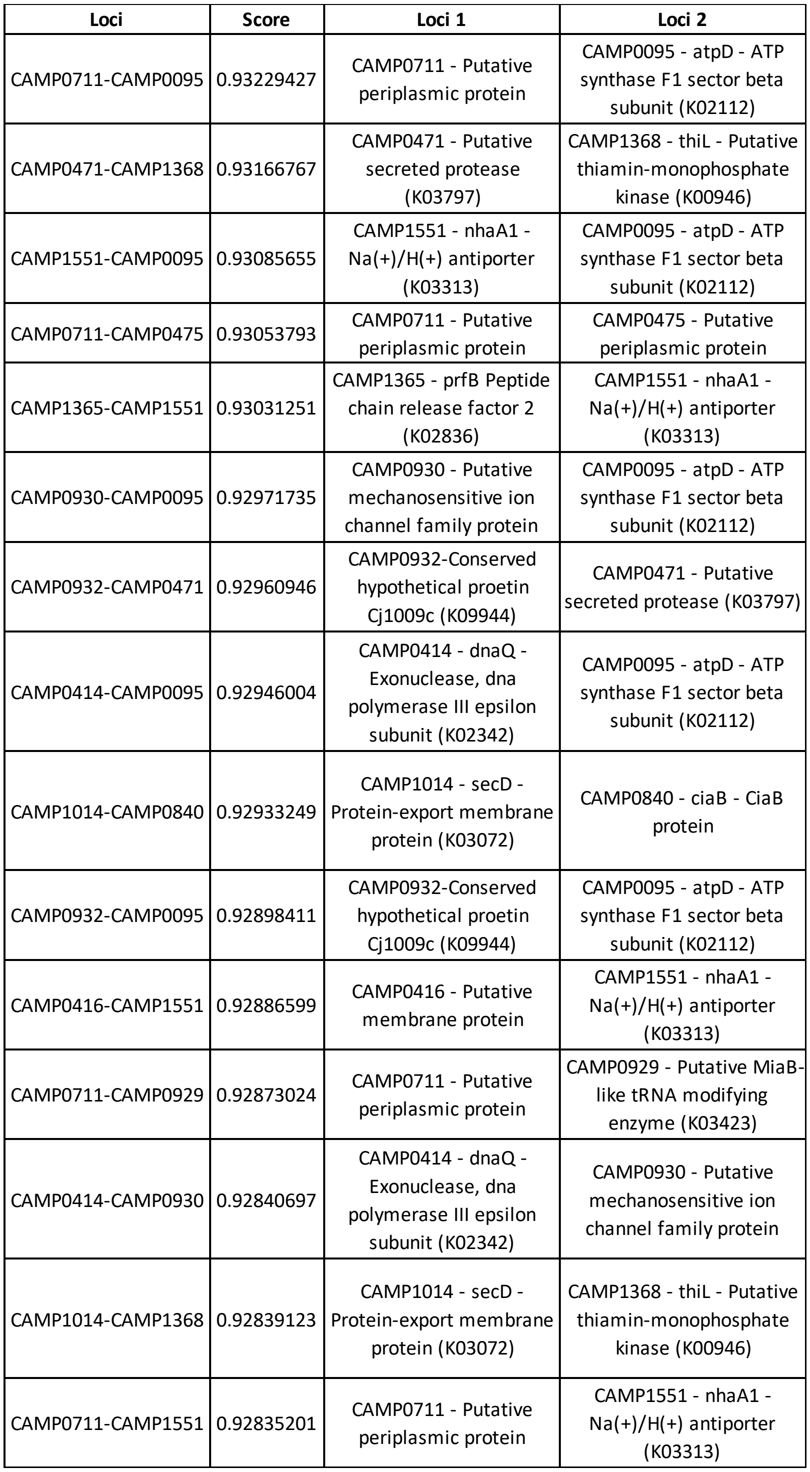


*Figure 12 Network showcasing the relations between clonal-complex and NeighbourGroups. The thickness of the line represents the count of each category. Blue are the NGroups, and red are clonal-complexes. We can see that NGroup 5 has the highest NGroups, and CC21 has the largest isolate counts from the size. We can also see how CC21 is connected to Ngroup 15 and Ngroup 16 with some in NeighbourGroup 9.*



| Loci | Score | Loci | Description |
|---|---|---|---|
| CAMP0932 | 0.90730047 | N/A | Conserved hypothetical proetin Cj1009c (K09944) |
| CAMP0471 | 0.9026372 | N/A | Putative secreted protease (K03797) |
| CAMP0308 | 0.89827915 | uvrA | Excinuclease ABC subunit A (K03701) |
| CAMP0414 | 0.89644358 | dnaQ | Exonuclease, dna polymerase III epsilon subunit (K02342) |
| CAMP1367 | 0.89556131 | truD | tRNA pseudouridine synthase D (K06176) |
| CAMP1369 | 0.89147309 | N/A | Hypothetical protein Cj1459 |
| CAMP0715 | 0.88827596 | napA | Periplasmic nitrate reductase (K02567) |
| CAMP1364 | 0.88214512 | N/A | Putative radical SAM domain family protein |
| CAMP1553 | 0.88029869 | N/A | Putative iron permease (K07243) |
| CAMP1022 | 0.87845787 | N/A | ATP-dependent DNA helicase (K03657) |
| CAMP0475 | 0.87567914 | N/A | Putative periplasmic protein |
| CAMP0843 | 0.87504912 | cstA | Putative integral membrane protein (CstA homolog) (K06200) |
| CAMP0519 | 0.87492355 | N/A | Putative MATE family transport protein |
| CAMP1368 | 0.87196118 | thiL | Putative thiamin-monophosphate kinase (K00946) |
| CAMP1153 | 0.87020581 | glyS | Glycyl-tRNA synthetase beta chain (K01879) |
| CAMP1551 | 0.86792311 | nhaA1 | Na(+)/H(+) antiporter (K03313) |
| CAMP1014 | 0.86557778 | secD | Protein-export membrane protein (K03072) |
| CAMP1006 | 0.86501486 | mfd | Transcription-repair coupling factor (K03723) |
| CAMP1283 | 0.86321383 | N/A | Putative nucleotidyltransferase (K00990) |
| CAMP0933 | 0.85995062 | tgt | Queuine tRNA-ribosyltransferase (K00773) |
| CAMP1138 | 0.85615379 | N/A | Putative periplasmic protein |
| CAMP1627 | 0.85519644 | N/A | Pseudogene (putative TonB-denpendent outer membrane receptor) |
| CAMP0931 | 0.85448752 | aroB | 3-dehydroquinate synthase (K01735) |
| **CAMP1541** | **0.85442855** | **tkt** | **Transketolase (K00615)** |
| CAMP1154 | 0.85379433 | N/A | Putative peptidase M23 family protein |
| CAMP0066 | 0.85360842 | lctP | L-lactate permease (K03303) |

*Figure 13 Loci's higher than 0.85 adjusted rand score to predict 20-NGroup and their loci description. CAMP0932 has often been found in the two combination loci as well. Tkt colour coded in green and shown in the bold banner, is one of the seven MLST loci*



| Loci | Score | Loci 1 | Loci 2 |
|---|---|---|---|
| CAMP0711-CAMP0095 | 0.93229427 | CAMP0711 - Putative periplasmic protein | CAMP0095 - atpD - ATP synthase F1 sector beta subunit (K02112) |
| CAMP0471-CAMP1368 | 0.93166767 | CAMP0471 - Putative secreted protease (K03797) | CAMP1368 - thiL - Putative thiamin-monophosphate kinase (K00946) |
| CAMP1551-CAMP0095 | 0.93085655 | CAMP1551 - nhaA1 - Na(+)/H(+) antiporter (K03313) | CAMP0095 - atpD - ATP synthase F1 sector beta subunit (K02112) |
| CAMP0711-CAMP0475 | 0.93053793 | CAMP0711 - Putative periplasmic protein | CAMP0475 - Putative periplasmic protein |
| CAMP1365-CAMP1551 | 0.93031251 | CAMP1365 - prfB Peptide chain release factor 2 (K02836) | CAMP1551 - nhaA1 - Na(+)/H(+) antiporter (K03313) |
| CAMP0930-CAMP0095 | 0.92971735 | CAMP0930 - Putative mechanosensitive ion channel family protein | CAMP0095 - atpD - ATP synthase F1 sector beta subunit (K02112) |
| CAMP0932-CAMP0471 | 0.92960946 | CAMP0932-Conserved hypothetical proetin Cj1009c (K09944) | CAMP0471 - Putative secreted protease (K03797) |
| CAMP0414-CAMP0095 | 0.92946004 | CAMP0414 - dnaQ - Exonuclease, dna polymerase III epsilon subunit (K02342) | CAMP0095 - atpD - ATP synthase F1 sector beta subunit (K02112) |
| CAMP1014-CAMP0840 | 0.92933249 | CAMP1014 - secD - Protein-export membrane protein (K03072) | CAMP0840 - ciaB - CiaB protein |
| CAMP0932-CAMP0095 | 0.92898411 | CAMP0932-Conserved hypothetical proetin Cj1009c (K09944) | CAMP0095 - atpD - ATP synthase F1 sector beta subunit (K02112) |
| CAMP0416-CAMP1551 | 0.92886599 | CAMP0416 - Putative membrane protein | CAMP1551 - nhaA1 - Na(+)/H(+) antiporter (K03313) |
| CAMP0711-CAMP0929 | 0.92873024 | CAMP0711 - Putative periplasmic protein | CAMP0929 - Putative MiaB-like tRNA modifying enzyme (K03423) |
| CAMP0414-CAMP0930 | 0.92840697 | CAMP0414 - dnaQ - Exonuclease, dna polymerase III epsilon subunit (K02342) | CAMP0930 - Putative mechanosensitive ion channel family protein |
| CAMP1014-CAMP1368 | 0.92839123 | CAMP1014 - secD - Protein-export membrane protein (K03072) | CAMP1368 - thiL - Putative thiamin-monophosphate kinase (K00946) |
| CAMP0711-CAMP1551 | 0.92835201 | CAMP0711 - Putative periplasmic protein | CAMP1551 - nhaA1 - Na(+)/H(+) antiporter (K03313) |

*Figure 14 Top 15 combinations of 2 cgMLST loci with the highest adjusted rand score*

## Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

### Chapter 4:

| | |
|---|---|
| Title of Paper | CatBoost classifier "NeighbourGroups" classification applied to *Campylobacter jejuni* |
| Publication Status | ☐Published ☐ Accepted for Publication<br>☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Authors: Dessislava Veltcheva, *Stephen Richer, Margaret Varga, Frances Colles, Michael B. Bonsall, Martin C. J. Maiden<br>Paper Style: Microbial Genomics from Microbiology Society (Research Article) |

Student Confirmation

| | | | |
|---|---|---|---|
| Student Name: | Dessislava Veltcheva | | |
| Contribution to the Paper | • Conceptualization of the research project<br>• Establishing Methodology<br>• Software and Script creation<br>• Validating results<br>• Formal analysis and Investigation<br>• Data curation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project management. | | |
| Signature Dessy | | Date | 02.10.2023 |

Supervisor Confirmation

| | | | |
|---|---|---|---|
| Supervisor name and title: Professor Michael Bonsall | | | |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. | | | |
| Signature. | | Date | 2nd October 2023 |

# Chapter 5: *Campylobacter jejuni* clonal-complex 353 re-evaluated through clonal-complex assignment methods: re-classification to multiple NeighbourGroups and proposal of CC581 as a new clonal-complex.

# *Campylobacter jejuni* clonal-complex 353 re-evaluated through clonal-complex assignment methods: re-classification to multiple NeighbourGroups and proposal of CC581 as a new clonal-complex.

Dessislava Veltcheva, Margaret Varga, Michael B, Bonsall, Martin C. J. Maiden

***Affiliation:*** University of Oxford

***Corresponding author***: dessislava.veltcheva@spc.ox.ac.uk

***Keywords****: Campylobacter jejuni, multi-locus typing scheme,* clonal-complex, antibiotic resistance, phylogenetic tree, variants, re-classification, microbiology, pathogen

***Repositories****:* CC353_Reclassification_project :*https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=113&submit=1*

## Abstract

*Campylobacter jejuni* is a leading cause of bacterial foodborne gastroenteritis, which is a major health concern. Currently, clonal-complexes are accepted as the classification method for *Campylobacter jejuni*. One clonal-complexe~~s~~, 353 (CC353), is known to be a poultry specialist and has been found in multiple countries. However, a previous study have indicated that CC353 is composed of numerous variants. In this study, we have demonstrated that CC353 comprises of five groups: two primary and two small sub-groups, with one isolate in another group. One of the sub-groups can be qualified as a new clonal-complex, CC581 due to its distinguished characteristics. Furthermore,

applying machine learning classification, “NeighbourGroups” has divided CC353 into four groups. Demonstrating that CC353 do need to be reclassified.

### Data Summary

Data are available for open access in the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018)" The samples can be found in CC353_Reclassification_project”, public project number 113 *https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=113&submit=1*

### Gap statement/Aim

While clonal-complexes have traditionally served as robust variant-level classification method, there is a lack of studies focused on re-evaluating the current clonal-complex classification. This study embarks on a comprehensive investigation of clonal-complex 353, employing diverse methodologies to reassess its variant classification. Through the application of multiple analytical approaches, including the "NeighbourGroups" classification, the primary objective of this research is to resolve the classification conundrum associated with CC353.

### Impact Statement

Incorporating whole genome analysis into the investigative framework has revealed that CC353 in *Campylobacter jejuni* comprises of multiple variants. The study demonstrated the non-monophyletic nature of CC353, making it necessary to reclassify this variant.

Furthermore, it demonstrated that if we want to reclassify in clonal-complex method only, this is not possible for CC353.

## Introduction

*Campylobacter jejuni* (*C. jejuni*), a thermophilic gram-negative bacterium, causes bacterial foodborne illness on a global scale, affecting populations across diverse socioeconomic strata (Endtz, 2020; Clarke *et al.*, 2021). This microorganism commonly finds its habitat within farm animal hosts, including poultry, cattle, and pigs (Sheppard *et al.*, 2010). Within these hosts, the transmission of *C. jejuni* is rapid, leading to colonisation within a couple of days from its initial contact (Awad, Hess and Hess, 2018). *C. jejuni* also exists in other sources, such as wild geese, sewage water, and unpasteurised milk (Lampis *et al.*, 2004; Longenberger *et al.*, 2013; Mourkas *et al.*, 2019).

Human infection with *C. jejuni* is typically characterised by clinical manifestations such as bloody diarrhoea, abdominal pain, and fever. However, most of these cases follow a self-limiting course, with documented outbreaks occurring relatively infrequently (Misawa *et al.*, 2002; Couturier, Hale and Couturier, 2012). The most common route of human contamination involves inadequate cooking practices and cross-contamination from food items that contain the bacterium (Colles *et al.*, 2003). In 2010, *Campylobacter* contributed to a 166 million cases of diarrheal illness globally (Devleesschauwer *et al.*, 2017). Among these cases, 31,700 were diagnosed with Guillain-Barré syndrome, a severe neurological disorder (Devleesschauwer *et al.*, 2017). Notably, over half of these cases were concentrated in Africa, with Southeast Asia as the second most affected region (Havelaar *et al.*, 2015).

Multi-locus sequence typing (MLST) was developed in the 1998, utilising *Neisseria meningitidis* to understand epidemiological variants (Maiden *et al.*, 1998). MLST schemes typically employ seven loci known as housekeeping genes. This approach standardised the investigation of *C. jejuni* variants, particularly in source attribution and allowed to have a comprehension of global epidemiological studies (Dingle *et al.*, 2001). The PubMLST database is a curated epidemiological repository, providing valuable data for long-term analyses (currently available: 29,734,629 alleles, 944,215 isolates and 684,022 genomes available on the 18th July 2022) (Jolley, Bray and Maiden, 2018).

Clonal-complexes utilise MLST to investigate bacterial populations with partially clonal population structures (Kate E. Dingle *et al.*, 2002). These partially clonal bacterial populations comprise clonal-complexes, presumed to derive from a common progenitor (Kate E. Dingle *et al.*, 2002; Maiden, 2006b). Clonal-complexes are assigned when four or more loci out of the 7 MLST loci match the central genotypes. This classification enables the identification of clonal relations, aiding in source attribution and comparing variants across the world (Sheppard *et al.*, 2010; Kittl *et al.*, 2013). This approach has subsequently led to the expansion of further variant level classification such as: development of ribosomal multi-locus typing (rMLST), core genome multi-locus typing (cgMLST), and whole-genome MLST schemes (Maiden *et al.*, 2013; Cody *et al.*, 2017).

A previous longitudinal study investigated the fluoroquinolone resistance of *Campylobacter jejuni* using clonal-complexes, displaying six distinct trends in *Campylobacter* resistance across time (Veltcheva *et al.*, 2022). This temporal analysis revealed diverse trajectories, encompassing susceptibility, resistance, and varying degrees of increased resistance to fluoroquinolones among different clonal-complexes. The observation concerning clonal-complex 353 was particularly noteworthy, as it exhibited a distinctive sigmoidal transformation in its fluoroquinolone resistance profile. In 1998, this clonal-complex was predominantly susceptible; however, by 2018, it had undergone a remarkable transition, evolving into a nearly uniformly resistant lineage against fluoroquinolones. This compelling shift underscores the dynamic nature of antimicrobial resistance within *Campylobacter jejuni* populations.

## Materials and Methods

### Meta Data Acquisition

All data are publicly available from the PubMLST database (Jolley, Bray and Maiden, 2018). The samples used for this study are 10,359 *Campylobacter jejuni* isolates (Veltcheva *et al.*, 2022). Within this sample, we have 480, CC353 isolates. Our search criteria for CC353 were: ("Species = *Campylobacter jejuni*" AND "Country = UK" AND "Clonal-complex == 353" AND "N50 >= 20,000" AND "1.4Mb <= Genome Size <= 1.8Mb" AND "Contigs <= 50" AND "source = human_stool"). The samples used can also be found directly on PubMLST public project ID number 111 - "CC353_Reclassification_project ":*https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=113&submit=1*

### Software usage/Bioinformatics pipeline

Using the PubMLST database, the *gyrase A* (*gyrA*, CAMP0950) locus was queried for each isolate. The *gyrA* nucleotide sequences were aligned using MAFFT/7.305 (Katoh and Standley, 2013). Then, using Transq, the amino acid translation was queried (Madeira et al., 2019). Although phenotypic data were not available for all isolates, Thr86Ile amino acid substitution strongly predicts fluoroquinolone resistance, and this have been used as an indication of phenotypic resistance in this study (Cody et al., 2012; Han et al., 2012).

Most of the script used for this study used R programming language (Team, 2020), with libraries including ggplot2, stats, d3heatmap, reshape2, heatmap, dplyr, tidyverse, lme4, and Viridis. Visualisation were performed both in R and Python3 (Van Rossum, G. , Drake, 2009). The minimum-spamming tree were run using GrapeTree through the PubMLST extension (Zhou *et al.*, 2018). The temporal dynamics were visualised using Tableau (‘Tableau (version. 9.1)’, 2016). Tables were created using Microsoft Excel (version 16.63). All the code is available on Github here: https://github.com/bgrdessislava/Campyjejuni_gyrA.

### Exploring the CC353, 5 groups dynamics

Multiple methods have been investigated to identify the differences among five sub-groups of CC353. Firstly, GrapeTree and Tableau were used to explore the relationships among the isolates with the minimum spanning tree algorithm (Zhou *et al.*, 2018; Beard and Aghassibakes, 2021). Secondly, the mechanisms behind MLST types to the clonal-complex have been investigated as to whether this technical method is the cause of the problem. Thirdly, the central genotypes were altered and investigated whether this would resolve the issue. The fourth step investigated the phylogeny of CC353 by using all cgMLST (1343 loci) and observed whether some subgroups are more distantly related. We used FASTTREE and clonalframeML for this step to implement recombination events using an HPC cluster from the University of Oxford (Richards, 2015). Fifthly, the seven housekeeping genes of the group 1 and 2 were compared. A new classification method, “NeighbourGroups”, a model to classify isolates based on the minimum-spamming tree, was implemented to identify CC353 into multiple groups (Veltcheva *et al.*, 2023) (Figure 20). This demonstrated that indeed CC353 should not be classified into 1 group.

## Results

In a previous study, 10,359 clinical isolates *of Campylobacter jejuni* dating back to 1998 and 2018 in the UK were queried for variant-specific patterns in the fluoroquinolone resistance (Veltcheva *et al.*, 2022). CC353 exhibited a sigmoidal transition from near-complete susceptibility in 1998 to nearly absolute resistance by 2018. This pattern could suggest that the fitness advantage associated with this variant or multiple subvariants have been grouped under the CC353 classification through original classification.

In the minimum-spanning phylogenetic tree, CC353 is distributed across five sub-groups (Figure 15), suggesting that they are not monophyletic. Thus, these five sub-groups have been investigated further.

### 1.) CC353 is not monophyletic but It is polyphyletic with multiple groups.

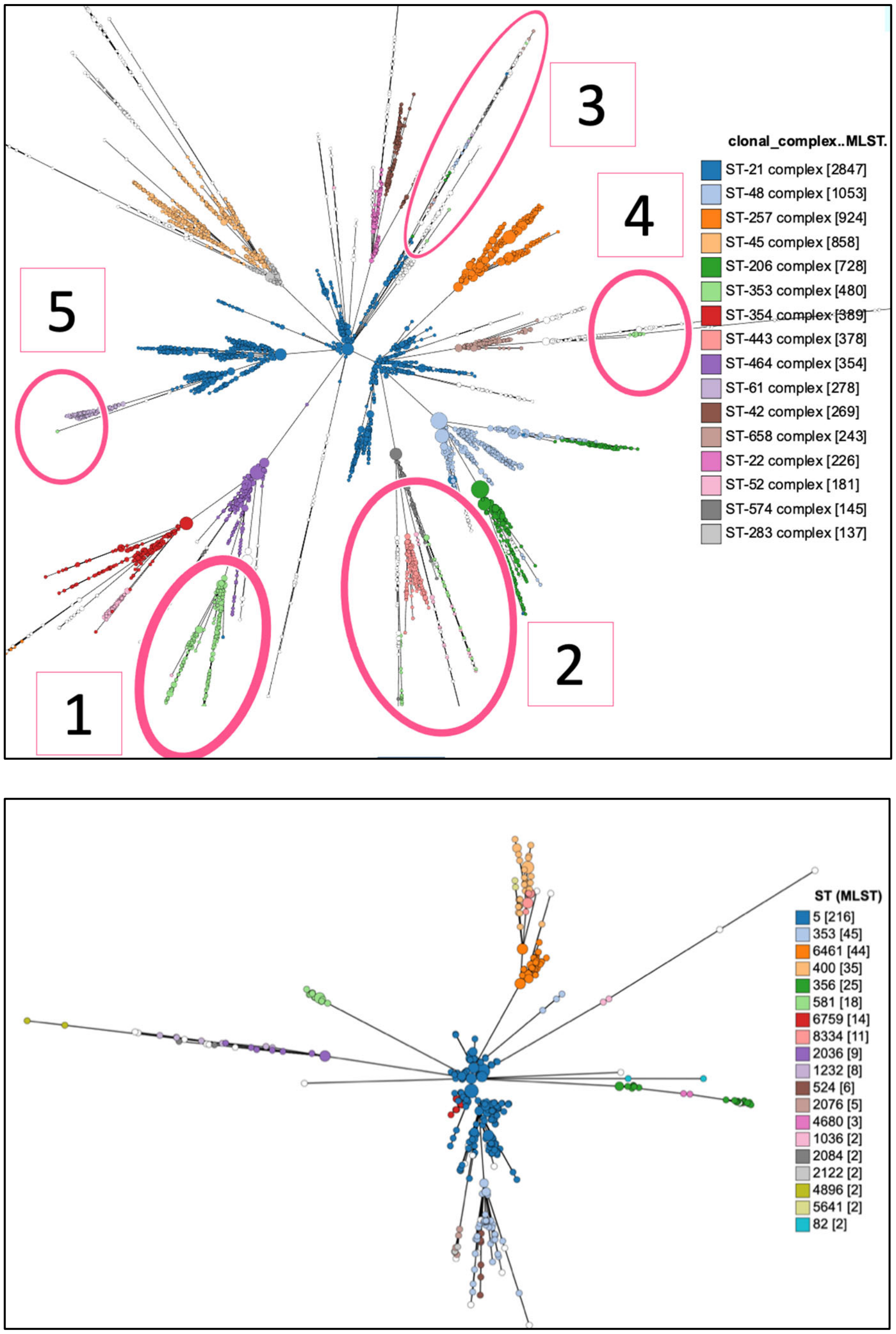


*Figure 15: Image Above minimum-spanning tree of 10,359 UK clinical isolates from 1998-2018 in the UK. Colour-coded by clonal-complexes and circles are the five sub-groups across the tree. Isolates with higher than 100 counts have been colour-coded. Image Bellow: CC353 has been further illustrated using their MLST loci as their colour. We can see ST5 being the most abundant MLST type, and we have also omitted any MLSTs that are less than two each.Any sequence type <2 have been omitted.*

Within clonal-complex 353 (CC353), there are two dominant clusters, Group 1 and a smaller Group 2 (Figure 15). Group 3 comprises only of four isolates stemming from the

CC21 sub-branch, while Group 4 comprises 18 isolates, closely resembling CC658. Group 5, on the other hand, is an outlier, represented by a single isolate that maintains connections with CC61 and CC21. Most clonal-complexes cluster cohesively, as shown in CC257, CC45, CC354, and CC464. In contrast, clonal-complexes such as CC21, CC48, and CC206 tend to span multiple branches. These clonal-complexes have been studied as more diverse (Dingle *et al.*, 2002).

Next a more granular analysis of the composition of these five sub-groups within CC353, encompassing 480 isolates and their respective resistance profiles (Figure 16). Notably, each sub-group exhibits a distinct set of MLST types, reinforcing that they are characterised by different assemblages of MLST profiles (Figure 16). Group 1, comprising the largest number of isolates, contains a diverse array of MLST types, with ST5, ST6461, and ST353 the most prevalent (Figure 16). We can also see that overall ST5 increased in number over time, surpassing ST353, which was predominant at the creation of the classification in 2001(Kate E Dingle *et al.*, 2002). A comparison of resistance patterns shows a mixture of resistant isolates across ST5, ST353, and ST400. Conversely, ST6461 predominantly aligned with resistance, suggesting possible selective pressures to this variant.

Group 2, characterised by a high incidence of ST356, has both susceptible and resistant isolates, with isolates primarily retrieved in Oxfordshire. Although group 3 comprises of four isolates, these isolates are found in different locations: Oxford, Newcastle, and Aberdeen. Group 4 is uniformly susceptible, transcending regional boundaries and exclusively represented by a single MLST type, ST581. These suggest possible misclassification of ST581 within CC353.

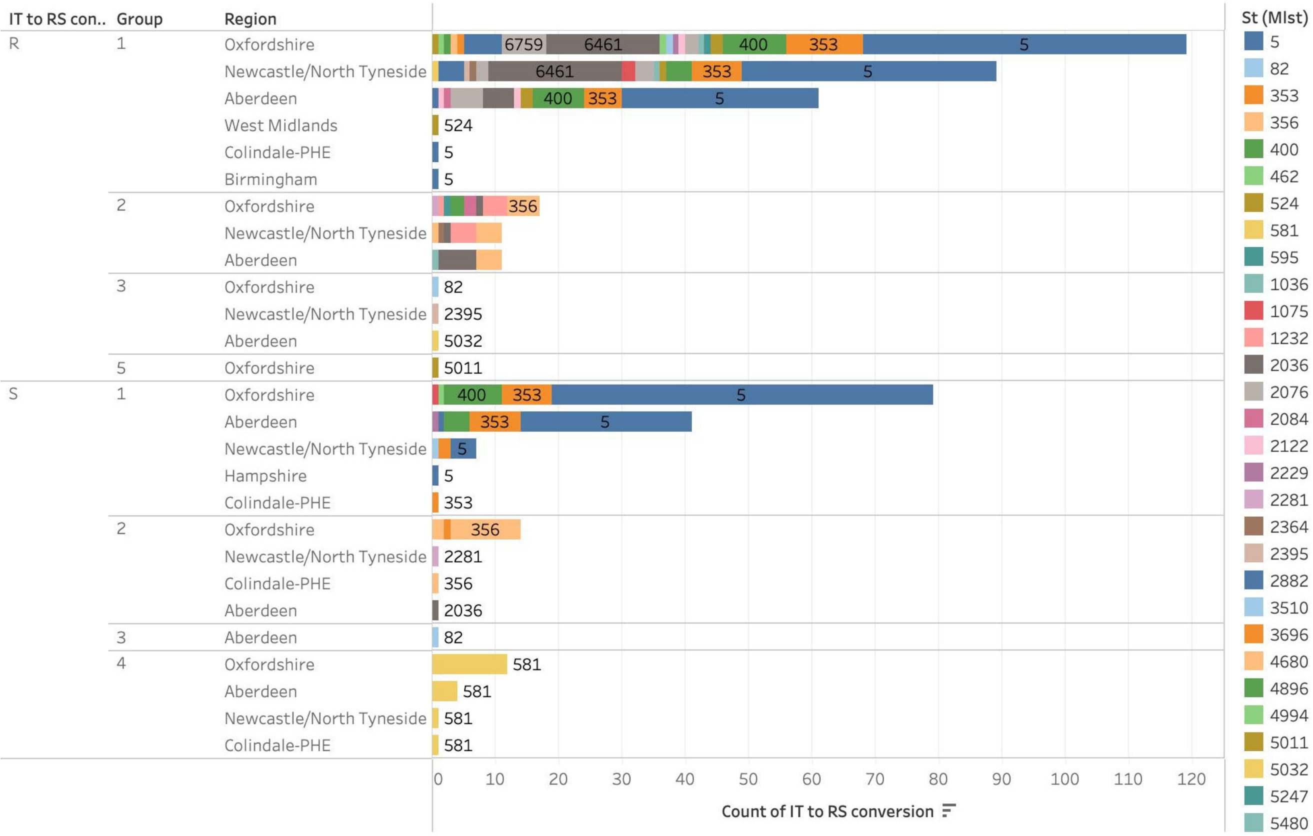


*Figure 16 CC353 - 480 Campylobacter jejuni's analysed for their five sub-groups on temporal, fluoroquinolone resistance and MLST differences; R indicates resistance, and S indicates susceptible. All of ST581 are found in Group 4 as susceptible isolates. Each STs were found in one group only.*

## 2.) The current method of central genotype assignment only contributed to 5% discrepancies in clonal-complex classification.

Next, we will examine the procedure we currently use to assign clonal-complex in detail. In the early 2022, an additional clonal-complex, ST581, was incorporated into this classification framework due to the findings derived from this study (Figure 16 and Figure 18). Consequently, the present repository of central genotypes for *Campylobacter jejuni* and *Campylobacter coli* has expanded to 45 unique central genotypes.

Assigning new MLST types to the existing clonal-complexes entails systematically examining each central genotype's seven MLST alleles. Every new sequence type goes through an assignment to the central genotype which is also the clonal-complex. First, the seven alleles are checked across the 44 central genotypes (Prior to CC581 being added) and whether six of seven housekeeping is the same it gets assigned to that CC. In instances where no congruence is identified at this stage, it checks for five of the same, and if none are found, it further explores the possibility of four shared loci. Without a match with at least four loci among any central genotypes, the novel MLST type is deemed ineligible for assignment to any clonal-complex. Thus, leading to not have clonal-complex assigned. An in-depth analysis was undertaken to evaluate this clonal-complex assignment methodology's implications comprehensively.

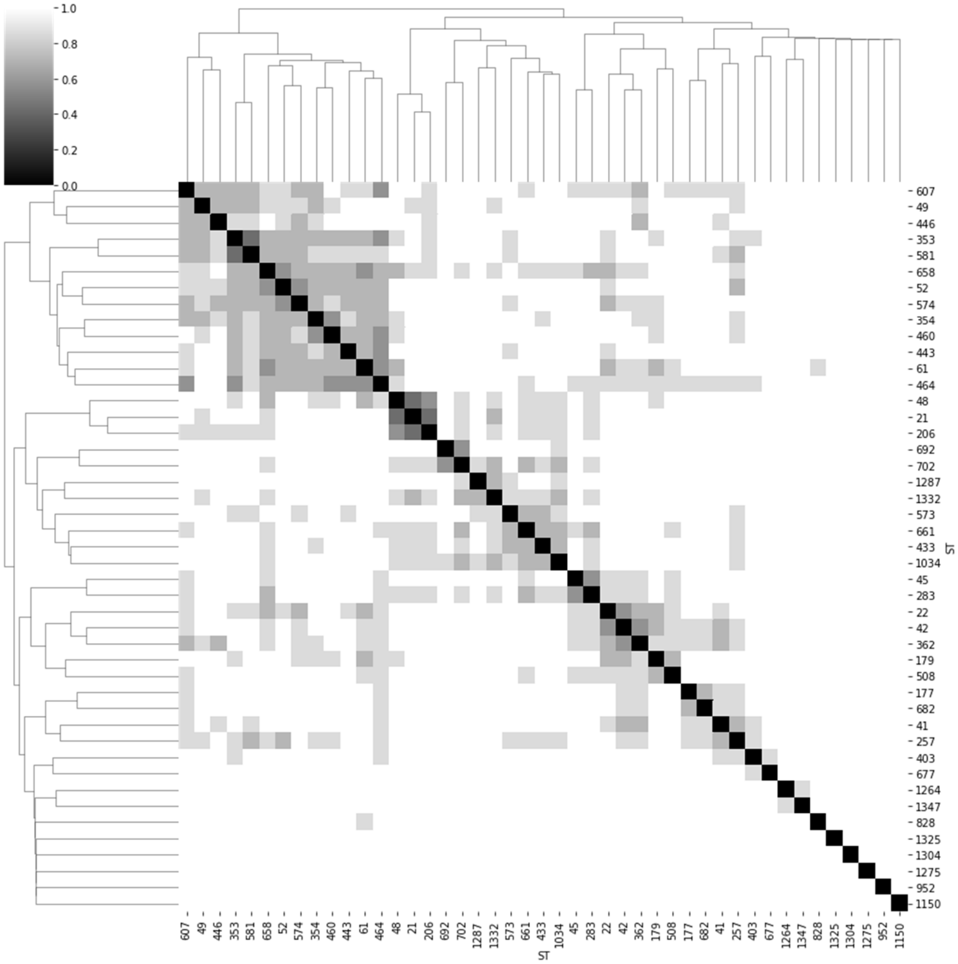


*Figure 17 Matrix of hamming distance compares the central genotypes' similarities by comparing their seven housekeeping MLST loci. The darker colour shows higher similarity. There are three main clusters horizontally. The first cluster contains more host specialist clonal-complexes; the second cluster shows major genotypes that do not have as many similarities with others. The third cluster, especially on the bottom right corner, is the clonal-complexes known to be a generalist.*

First, a matrix system was used to cross-compared analysis among central genotypes. This rigorous evaluation aimed to quantify the extent of genetic similarity, spanning seven distinct loci, among these central genotypes. Hamming distances were used to identify the similarities (Figure 17), resulting in three horizontally clusters. The first cluster contains: CC464, CC658, CC607, CC353, CC581, CC52, CC443, CC354, CC61, CC574, and CC460. Within this cluster, we have a clonal-complex associated with poultry sources: CC464, CC353, and CC354. This similarity in the seven central genotypes could indicate why we see them as poultry specialists.(Di Giannatale *et al.*, 2019; Yu *et al.*, 2020; Saif *et al.*, 2022). Notably, CC61, a recognised cattle specialist, is within this first cluster as well, showing that although we are separating into different clonal-complexes, they could be more closely related than the clonal-complex classification has separated them to be.

The second cluster contains: CC206, CC283, CC702, CC661, CC1034, CC433, CC1332, CC21 and CC48. We can see here that CC21 and CC48 are among the most abundant isolates found globally and are classified in one place. The fact that different clonal-complexes could have duplicated alleles in their housekeeping genes could lead to a lack of confidence in the clonal-complex system. The last cluster in the end is clonal-complexes that do not have matching alleles with other clonal-complexes (CC403, CC828, CC677). This cluster includes CC828, a central genotype that significantly contributed to the classification of *Campylobacter coli*.

Next, a permutation-based methodology assessed the potential for reclassifying MLST types into alternative clonal-complexes. Given the unchanging sequence of central genotypes, randomised reordering of central genotypes was iteratively performed 1000 times. This iterative procedure sought instances where certain MLST types could be reclassified into different clonal-complexes, illuminating alternative classification possibilities (Table 3).

*Table 3 Randomising the order of central genotypes and identifying whether clonal-complex assignment changes across*

| **Clonal-complex changes to** | **Identical CC (No change)** | **1 other** | **2 other** | **3 other** | **4 other** | **Total** |
|---|---|---|---|---|---|---|
| MLST Count | 6783 | 357 | 37 | 3 | 1 | 7181 |
| % | 94.46 | 4.97 | 0.52 | 0.04 | 0.01 | 100.00 |
| | | | | | | |
| Isolate Count(25th June 2023) | 105804 | 4866 | 199 | 6 | 1 | 110876 |
| % | 95.43 | 4.39 | 0.18 | 0.01 | 0.00 | 100 |

Overall, we have 7181 MLST types available (9th May 2022), and we can see that 399 (5.5%) of the MLST could be replaced into another clonal-complex if we randomise the order (Table 3). Although some central genotypes could be changed with more than three central genotypes, these are only around 0.58% of the overall MLSTs. This result confirms how the clonal-complex classification method is only causing around 5.6% of misclassifications. Furthermore 94.46% will not change to another clonal-complex.

Although we can see that around 5.55% of the MLSTs available could be replaced with another clonal-complex, the fundamental question is, do we have the correct central genotypes. In addition, some clonal-complexes are mainly composed of 6 of the same, making the group more similar than if you have a clonal-complex predominantly high in 4 identical alleles.

### 3.) Changing Central genotypes does not solve the CC353 sub-group formation.

Subsequently, it became evident that the issue's root lay not in the technical methodology applied to central genotypes but in classifying ST353 as the central genotype. Thus, the central genotype was modified and examined (Table 4). This transformative endeavour encompassed the implementation of three distinct methodologies: the exclusion of CC353 entirely as central genotype, the substitution of the central genotype with ST5, and the simultaneous exclusion of both CC353 and CC581 from central genotype categorisations.

Following the revision of central genotypes, a notable outcome emerged, wherein the alteration of the central genotype from CC353 to CC5 resulted in the reclassification of 39 out of 46 MLSTs. Nonetheless, it is imperative to acknowledge that this adjustment gave rise to a recurrence of the initial predicament. Groups 1, 2, 3, and 5 were incorporated within the same clonal-complex; therefore, changing to CC5 did not solve the problem (Table 4).

*Table 4 Reclassifying central genotypes and identifying whether we can resolve the problem of 5 groups.*

| MLST | Group | CC Without Central Genotype ST353 | Central genotype changed from CC353 to CC5 | CC without CC581 and CC353 |
|---|---|---|---|---|
| **5** | 1 | 464 | 5 | 464 |
| **82** | 3 | N/A | 5 | N/A |
| **353** | 1 | 581 | 5 | N/A |
| **356** | 2 | N/A | 5 | N/A |
| **400** | 1 | N/A | 5 | N/A |
| **462** | 1 | 581 | 5 | N/A |
| **524** | 1 | 607 | 5 | 607 |
| **595** | 1 | 464 | 5 | 464 |
| **1036** | 1 | 574 | 5 | 574 |
| **1075** | 1 | 460 | 5 | 460 |
| **1232** | 2 | 574 | 5 | 574 |
| **2036** | 2 | 574 | 574 | 574 |
| **2076** | 1 | N/A | N/A | N/A |
| **2084** | 2 | N/A | 5 | N/A |
| **2122** | 1 | N/A | 5 | N/A |
| **2229** | 1 | N/A | N/A | N/A |
| **2281** | 2 | 607 | 607 | 607 |
| **2395** | 3 | 581 | 5 | N/A |
| **2882** | 1 | 581 | 5 | N/A |
| **3510** | 1 | 581 | 5 | N/A |
| **3696** | 2 | 52 | 52 | 52 |
| **4680** | 2 | N/A | N/A | N/A |
| **4896** | 2 | N/A | N/A | N/A |
| **4994** | 1 | N/A | 5 | N/A |
| **5011** | 5 | N/A | 5 | N/A |
| **5032** | 3 | N/A | 5 | N/A |
| **5247** | 2 | 574 | 574 | 574 |
| **5480** | 2 | N/A | N/A | N/A |
| **5641** | 1 | N/A | N/A | N/A |
| **5866** | 2 | 581 | 5 | N/A |
| **6461** | 1 | N/A | 5 | N/A |
| **6759** | 1 | 464 | 5 | 464 |
| **7147** | 1 | N/A | 5 | N/A |
| **7198** | 1 | 464 | 5 | 464 |
| **7203** | 1 | 464 | 5 | 464 |
| **7865** | 2 | 607 | 607 | 607 |
| **8065** | 1 | 581 | 5 | N/A |
| **8197** | 1 | N/A | 5 | N/A |
| **8334** | 1 | 354 | 354 | 354 |
| **8846** | 1 | N/A | 5 | N/A |
| **8885** | 1 | N/A | 5 | N/A |
| **8903** | 1 | N/A | 5 | N/A |
| **9100** | 1 | N/A | 5 | N/A |
| **9385** | 1 | 581 | 5 | N/A |
| **9388** | 1 | 581 | 5 | N/A |
| **9390** | 1 | N/A | N/A | N/A |
|  | Assigned | 24 | 39 | 15 |
|  | Assigned Percent | 52.17 | 84.78 | 32.61 |

An alternative approach involving the removal of CC353 while retaining CC581 yielded a reclassification rate of 52% among the isolates. However, this method left many isolates without any categorical classification. Lastly, excluding CC353 and CC581 resulted in the reclassification of approximately 30% of the MLSTs, leaving a notable portion, particularly ST353, without definitive classification. Hence, it the alteration of central genotypes within CC353 do not comprehensively resolve the problem of CC353 re-classification.

### 4.) The horizontal gene transfer applied tree distinguishes into two groups:” first group [1 and 2 sub-groups], second group [3,4 and 5]

To further dive into the phylogenetic relations of these five groups, cgMLST (1343 loci) were examined for their sequences(Figure 18). The CC353's 480 isolates were run through the maximum likelihood method, and the result Newick file was run through ClonalFrameML to account for horizontal gene transfer. Groups 4 and 5 have diverged earlier than groups 1,2 and 3. Group 3 is phylogenetically closer to Group 4, and Groups 1 and 2 are closely related (Figure 18). Further supporting the argument that group 4, composed of ST581, should be a separate clonal-complex since it is a different branch from any of the other groups.

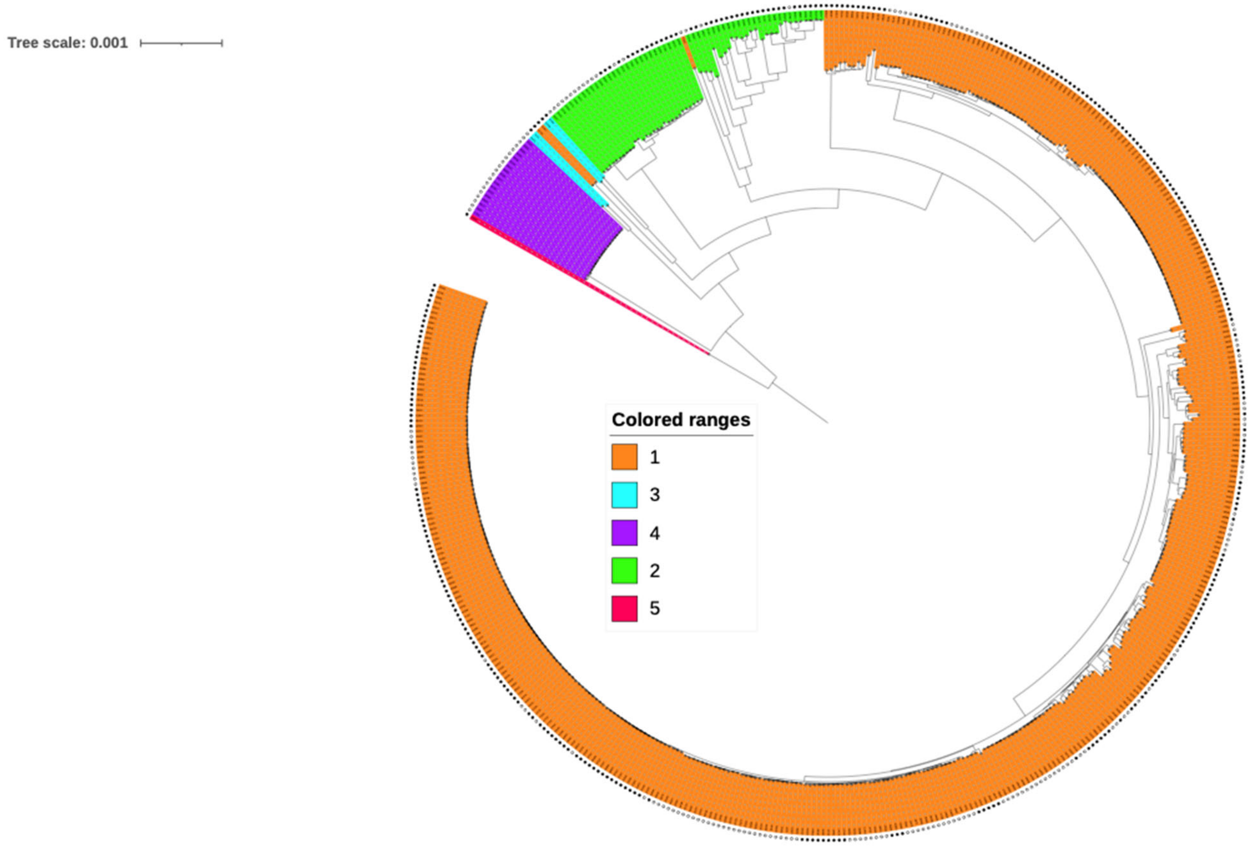


*Figure 18 ClonalframeML analysis infers recombination events within the CC353 isolates. Groups 1 and 2 are much more closely related than others. Groups 4 and 5 seem to differ most from the other group.*

### 5.) MLST group 1 and 2 alleles have three out of seven majority alleles.

Now that we know group 1 and group 2 are more closely a comprehensive comparative analysis of their 7 MLST alleles was conducted (Figure 19). This comparison has shown that three loci, including *gltA, uncA,* and *tkt*, exhibited identical majority alleles across both groups. In contrast, *aspA, pgm, glnA*, and *glyA* displayed divergent alleles usually. Despite both groups belonging to the same clonal-complex, this difference is particularly evident in the case of *pgm*, where a clear differentiation into two distinct groups is readily discernible. Consequently, the results of this analytical investigation led to the finding that CC353's Group 1 and Group 2 represent two specific sub-groups within the broader framework of CC353 and should not be classified as a single clonal-complex.

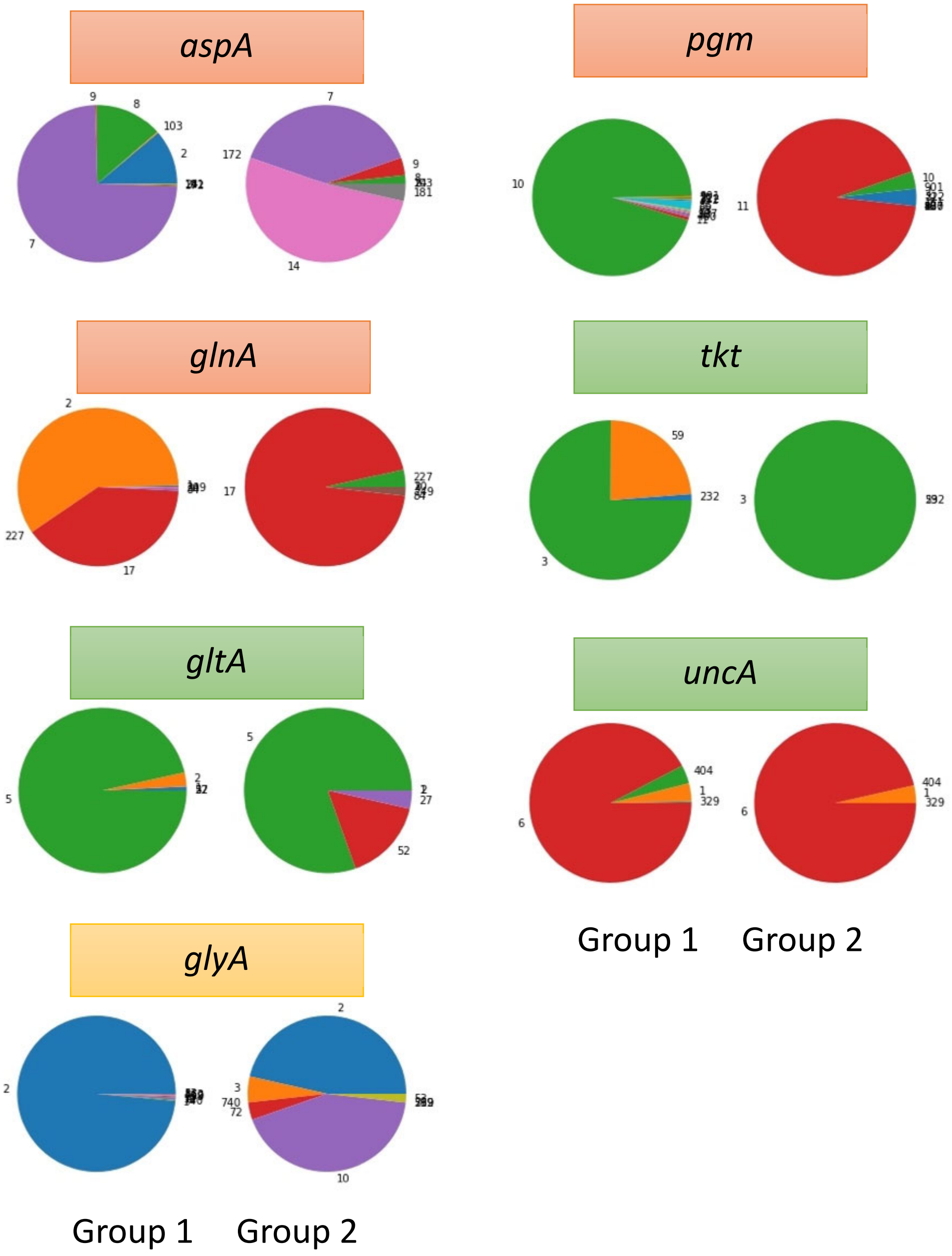
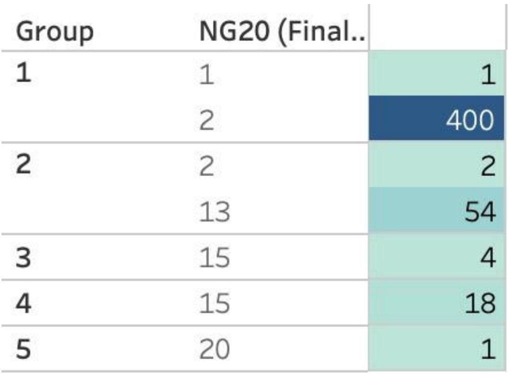


*Figure 19 CC353 Group 1 and Group 2's 7 MLST' loci allele composition comparison states how only three loci have the same majority allele, indicating that groups 1 and 2 should be separate. The red color shows the allele that are different in the majority. Green shows the loci that have the same allele majority. The yellow color (glyA) has the same allele 2, in both groups but did not meet the majority criteria.*

## 6.) Application of NeighbourGroups onto CC353



| Group | NG20 (Final.. | |
|---|---|---|
| 1 | 1 | 1 |
| | 2 | 400 |
| 2 | 2 | 2 |
| | 13 | 54 |
| 3 | 15 | 4 |
| 4 | 15 | 18 |
| 5 | 20 | 1 |

*Figure 20 NeighbourGroups applied onto clonal-complexes and their counts.*

Lastly, a previously established classification methodology known as "NeighbourGroups" was applied to our isolates (Veltcheva *et al.*, 2023)(Figure 20). A marginal number of isolates in Groups 1 and 2 are assigned to different NeighbourGroups (ranging from 1 to 2 isolates), and most groups are consistently associated with a specific NeighbourGroup assignment. Specifically, Group 1 is predominantly classified within NeighbourGroup 2, whereas Group 2 aligns mainly with NeighbourGroup 13, Groups 3 and 4 are primarily affiliated with NeighbourGroup 15, and Group 5 exhibits alignment with NeighbourGroup 20. This outcome supports the findings that CC353 can be effectively partitioned into four distinct groups by applying the NeighbourGroups classification method. Although we believed it would be five groups initially, through NeighbourGroups, they are separated into four groups. The last 1 isolate can be omitted until we can expand the sample sizes and see if there are other isolates in this group. Thus, the findings provide compelling evidence for reclassifying clonal-complex 353 into multiple specific groups.

## Discussion

In this investigation, we have undertaken a comprehensive analysis employing five distinct methodologies to re-evaluate the classification of *Campylobacter jejuni* clonal-complex 353 (CC353). It is worth noting that other clonal-complexes, such as 21, 45, and 206, have also been found in various regions of these phylogenetic trees. These complexes, often called generalists, have previously been subject to investigation regarding the potential presence of multiple clades within their classification(Sheppard et al., 2014; Woodcock et al., 2017). Given the sigmoidal pattern of fluoroquinolone resistance exhibited by CC353, the possibility arose that this complex included distinct genetic variants that have, erroneously, been grouped under a singular classification. One notable gap in the clonal-complex classification is the absence of inquiries into the persistence of the central genotypes initially established in the 2001(Dingle *et al*., 2002). Thus, this study re-evaluated the classification of clonal-complex 353 within the framework of the Multi-locus Sequence Typing (MLST) initially. However, due to the fact that it is not solvable within the MLST scheme, an alternative classification methods NeighbourGroups have been used.

### Beyond Clonal-complex classification

The utilisation of clonal-complexes as a classification system for *Campylobacter jejuni* (*C. jejuni*) is a widely accepted practice within the scientific community and has been extensively applied in numerous studies (Dearlove *et al*., 2016; Ramonaite *et al*., 2017; Yamada *et al*., 2019; Arning *et al*., 2021). Efforts to enhance the resolution of this classification scheme have led to notable advancements in understanding the temporal distribution of endemic patterns. Notably, the extension of the Multi-locus Sequence Typing (MLST) scheme to incorporate markers such as *gyrA* and *porA* has been attempted (Nennig *et al*., 2021). Additionally, identifying *flaA* short variable regions and *porA* gene variants as informative loci has proven valuable for distinguishing outbreaks (Dingle *et al*., 2008). It is noteworthy that the application of whole-genome MLST data for outbreak identification has been emphasised in several studies. Nevertheless, a continuing challenge is dealing with historical isolates lacking whole-genome sequencing data (Llarena, Taboada and Rossi, 2017; Epping, Antão and Semmler, 2021). Moreover, complementary research endeavours have further substantiated these findings by focusing on allelic variations within key loci such as *dnaE, ftsX*, and *pycB,* which are known to contribute to the ecological niches within host organisms (Epping *et al*., 2021).

Our study has provided evidence that core and accessory genomic components play important roles in determining host specificity. The interplay of evolutionary processes governs the presence or absence of genes and dictates their relative expression levels. Although our study primarily relies on genomic data, integrating phenotypic data, particularly of resistance levels could, offer a deeper insights into emerging trends.

### Clonal-complex 353 in a global context

Clonal-complex 353 (CC353) is a prominent and widely distributed variant with a global distribution. Notably, in Lithuania, studies have revealed the dominance of CC353 and CC21 among paediatric clinical variants, with environmental variants seldom encountered in clinical cases (Ramonaite *et al*., 2014). Furthermore, a separate Lithuanian study was consistent with these findings, underscoring the prevalence of ST5 within CC353 isolates, with 96.9% of these isolates displaying resistance to ciprofloxacin, indicating a substantial resistance burden (Aksomaitiene *et al*., 2019).

Expanding our geographical scope to Japan, while the sample size was only ten isolates, CC353 were identified. Within CC353, some divergence was observed in pulsed-field gel electrophoresis (PFGE) patterns, with ST3052 and ST400 emerging as notable subtypes. Interestingly, ST400 was the only shared ST between Japanese and UK studies (Yabe *et al*., 2010). Furthermore, a recent investigation employed Turkey's newly established CC581 classification, highlighting the utility of the updated 45 clonal-complexes. (Aydin *et al*., 2023). In this Turkish study, CC353 emerged as one of the most prevalent clonal-complexes, with strong associations observed between CC353 and clinical cases and other animal sources. CC581 was predominantly linked to chicken origins, demonstrating the rapid adoption of this novel variant classification in ongoing epidemiological studies (Aydin *et al*., 2023).

In China, findings indicated that CC353 and CC464 were the predominant variants in central China, while in the east of the country, CC21 accounted for a substantial 39.3% of isolates. The study further emphasised the strong correlation between fluoroquinolone resistance and CC353 in the Chinese context, frequently occurring in clinical cases in China. These observations align with the hypothesis that poultry is the primary transmission route to clinical infections (Wang *et al*., 2022). Likewise, Luxembourg's isolates corroborate the prevalence of CC21 as the most prominent variant, constituting

24% of isolates, while CC353 represented 5.8%. In Luxembourg, CC353 also emerged as a variant strongly associated with clinical cases, while CC61 was notably linked to bovine origins(Ragimbeau *et al*., 2008).

## Conclusion

In summary, our study reveals the challenges of reclassifying CC353. Altering the classification system or central genotype to ST5 does not provide a straightforward solution. Clonal-frame analysis shows that CC353's groups 1 and 2 are closely related at the core MLST level. However, examining only the 7 MLST housekeeping genes reveals divergence, identifying two distinct CC353 sub-clusters. Group 4, characterised by ST-581, stands out due to its phylogenetic and resistance differences, making a case for introducing the new central genotype, CC581. Lastly, NeighbourGroups refines our classification, consolidating the initial five groups into four distinct ones, a distinction we could not make just through clonal-complex classification.

## Supplementary

*Table 5 MLSTs from groups 1 and 2 in which the randomisation influences the clonal-complex assignment. Across the 480 CC353 isolates, only 28 have a potential effect due to this randomisation, leading to around a 5.8% error rate. nAssignments illustrates that it can be classified up to other groups.*

| MLST | Mode | Mode Frequency | Observed | Frequencies | Clonal_complex | nAssignments | Group | Count |
|---|---|---|---|---|---|---|---|---|
| **1036** | 574 | 0.341 | ('353', '574', '581') | (0.341, 0.34, 0.319) | 353 | 3 | 1 | 2 |
| **8334** | 460 | 0.361 | ('353', '460', '354') | (0.361, 0.33, 0.309) | 353 | 3 | 1 | 11 |
| **7865** | 353 | 0.5 | ('353', '607') | (0.5, 0.5) | 353 | 2 | 2 | 1 |
| **2281** | 353 | 0.5 | ('353', '607') | (0.5, 0.5) | 353 | 2 | 2 | 1 |
| **7198** | 464 | 0.504 | ('353', '464') | (0.504, 0.496) | 353 | 2 | 1 | 1 |
| **1075** | 460 | 0.505 | ('353', '460') | (0.505, 0.495) | 353 | 2 | 1 | 1 |
| **3696** | 353 | 0.506 | ('353', '52') | (0.506, 0.494) | 353 | 2 | 2 | 1 |
| **5247** | 574 | 0.516 | ('353', '574') | (0.516, 0.484) | 353 | 2 | 2 | 1 |
| **2036** | 574 | 0.516 | ('353', '574') | (0.516, 0.484) | 353 | 2 | 2 | 9 |

| | |
|---|---|
| **Total error** | 28.00 |
| **Total** | 480.00 |
| **Error rate** | 0.06 |
| **100% scale** | 5.83 |

Exploring CC353 and its reclassification, we see that in some groups 1 and 2, MLSTs could be classified into another clonal-complex (Table 5). However, overall, the total amount of isolates that could be reclassified are 28 out of 480, which means that only 5.8% of the isolates have potential changes based on randomisations in the CC353 isolates.

*Table 6 Error Rate based on 5 Groups indicating how group 2 has the highest error rate based on randomisation. Groups 3,4 and 5 do not have an issue even with randomisation.*

| **Group** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Misclassified** | 15 | 13 | 0 | 0 | 0 |
| **Correct** | 386 | 43 | 4 | 18 | 1 |
| **Total Count** | 401 | 56 | 4 | 18 | 1 |
| **Error rate** | 0.04 | 0.23 | 0 | 0 | 0 |
| **Percent** | 3.74 | 23.21 | 0 | 0 | 0 |

Groups 1 and 2 would slightly impact the randomisations (Table 6). On the other hand, groups 3,4 and 5, although lower in the count, will not have any issues. Group 2 has the highest reclassification, with 23% reclassified into another clonal-complex.

From this analysis, we can show that the ordering of the central genotypes impacts group 2's classification. Still, only about 20% and only about 5% of the MLST types could be reclassified into another clonal-complex, proofing that this method is insufficient to reclassify CC353 or to understand why there are five separate groups.

*Table 7 Choosing CC353, CC21 and CC61 and identifying their error rates based on randomisations.*

| **Mismatch** | **2** | **3** | **4** | **5** | **Total** | Isolates proportion from total | Missclasify 2 or more |
|---|---|---|---|---|---|---|---|
| **CC353 Only** | **80** | **23** | **3** | **0** | **573** | | |
| Percentage | 0.139616056 | 0.040139616 | 0.005235602 | 0 | 0.184991274 | 0.08 | 0.01 |
| 100% scale | 13.96160558 | 4.013961606 | 0.523560209 | 0 | 18.4991274 | 7.98 | 1.11 |
| | | | | | | | |
| **CC21 only** | **104** | **7** | **0** | **0** | **922** | | |
| Percentage | 0.112798265 | 0.007592191 | 0 | 0 | 0.120390456 | 0.13 | 0.01 |
| 100% scale | 11.27982646 | 0.759219089 | 0 | 0 | 12.03904555 | 12.84 | 1.45 |
| | | | | | | | |
| **CC61 only** | **15** | **4** | **1** | **1** | **207** | | |
| Percentage | 0.072463768 | 0.019323671 | 0.004830918 | 0.004830918 | 0.101449275 | 0.03 | 0.00 |
| 100% scale | 7.246376812 | 1.93236715 | 0.483091787 | 0.483091787 | 10.14492754 | 2.88 | 0.21 |

Some other clonal-complexes were also looked at to see whether a specific clonal-complex has a high impact based on this current system (Table 7). Most isolates are from CC21 (12.8%). However, only 1.45% of the clonal-complex classification will be changed if the order is dynamic. CC353 will be the only MLST currently available. CC61, one of the MLSTs (2951) with potential five reclassifications available, shows that only 0.28% of the complete clonal-complex assignment is due to this clonal-complex. Randomisation could improve the current system, but we can still trust the results.

## Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

### Chapter 5:

| | |
|---|---|
| Title of Paper | *Campylobacter jejuni* clonal-complex 353 re-evaluated through clonal-complex assignment methods: re-classification to multiple NeighbourGroups and proposal of CC581 as a new clonal-complex |
| Publication Status | ☐Published ☐ Accepted for Publication<br>☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Authors: Dessislava Veltcheva, Margaret Varga, Michael B, Bonsall, Martin C. J. Maiden<br>Paper Style: Microbial Genomics from Microbiology Society (Research Article) |

Student Confirmation

| | | | |
|---|---|---|---|
| Student Name: | Dessislava Veltcheva | | |
| Contribution to the Paper | e.g. performed analysis on all samples, interpreted data, wrote manuscript.<br>• Conceptualization of the research project<br>• Establishing Methodology<br>• Software and Script creation<br>• Formal analysis and Investigation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project management | | |
| Signature Dessy | | Date | 02.10.2023 |

Supervisor Confirmation

| | | |
|---|---|---|
| Supervisor name and title: Professor Michael Bonsall | | |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. | | |
| Signature | Date | 2nd October 2023 |

# Chapter 6: Trends in Fluoroquinolone and Tetracycline Resistance of *Campylobacter jejuni* using NeighbourGroups

# Trends in Fluoroquinolone and Tetracycline Resistance of *Campylobacter jejuni* using NeighbourGroups

Dessislava Veltcheva, Michael B, Bonsall, Martin C. J. Maiden

***Affiliation*** University of Oxford

***Corresponding author***: dessislava.veltcheva@spc.ox.ac.uk

***Keywords**: Campylobacter jejuni, multi-locus typing scheme,* clonal-complex, phylogenetic tree, variants, re-classification, microbiology, pathogen

***Repositories**:* The data is available on public project in the PubMLST: No.127 Campyjejuni_UK_1997_2022(n=14,481))
[https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=127&submit=1]

## Abstract

This study examined *Campylobacter jejuni* (*C.jejuni*) variants from diverse sources from the UK between 1997 to 2022. Fluoroquinolone and tetracycline resistance patterns were investigated using the NeighbourGroup classification system. Findings indicate significant antibiotic resistance for NeighbourGroup 1 and 2 (CC353, CC354, and CC464). On the other hand, NeighbourGroup 13 (CC443, CC574) exhibited high tetracycline resistance but not to the fluoroquinolone. Moreover, NeighbourGroup 15 (a sub-group of CC21) was resistant to fluoroquinolones but not tetracycline. On the contrary, NeighbourGroup 7

(CC45) was susceptible to both antibiotics, showing further decreasing resistance to both across the study. Overall, the NeighbourGroup classification revealed evolving antibiotic resistance patterns within *C. jejuni* populations, shedding light on resistance mechanisms and temporal dynamics.

## Data Summary

Data are open-accessible and can be found on the PubMLST database (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018). The data can be searched using No.127 Campyjejuni_UK_1997_2022(n=14,481) (Queried on the 5th of March 2023). [https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=127&submit=1]

## Gap statement/Aim

While NeighbourGroups have been established for *C. jejuni,* their application in revealing trends in antimicrobial resistance remains unexplored. This study investigated chromosomally mediated fluoroquinolone resistance and plasmid-mediated tetracycline resistance within NeighbourGroups. The objective is to identify potential variations in resistance mechanisms among different groups and ascertain any similarities through trend analysis within clonal-complexes.

## Impact Statement

This research demonstrates the pivotal role of NeighbourGroups in unveiling compelling trends in fluoroquinolone and tetracycline resistance dynamics within *C. jejuni*. Leveraging NeighbourGroups, which utilise cgMLST, shows intricate resistance patterns, which advanced our understanding compared to prior studies. Previous investigations identified six distinct fluoroquinolone resistance patterns within clonal-complexes; our study extends and deepens this understanding by encompassing tetracycline resistance, yielding invaluable insights into the evolving landscape of antimicrobial resistance.

## Introduction

*Campylobacter jejuni (C. jejuni)* is a primary causative agent of bacterial gastroenteritis, substantially burdening healthcare systems. In the United Kingdom it is estimated that approximately 500,000 infections annually, prompting over 80,000 general practitioner consultations yearly (Tam *et al*., 2012; Kaakoush *et al*., 2015a). This aerophilic microorganism thrives in the intestinal poultry, which is an ideal habitat for proliferation (Facciolà *et al*., 2017). Embedded within the intricate food chain, *Campylobacter* can persist through abattoir and eventually contaminate poultry products available in retail settings. Suboptimal cooking practices or cross-contamination during food handling can lead to its transmission to the human gastrointestinal tract, leading to foodborne illnesses (Clarke and Ajlouni, 2021). The symptoms can range from bloody diarrhoea and abdominal discomfort to vomiting and fever (Haldenby *et al*., 2020). Generally, these symptoms are self-limiting and antibiotic prescriptions are typically reserved for individuals who are immunocompromised, elderly, young, or those with underlying medical conditions (Kaakoush *et al*., 2015b).

*C. jejuni* has been recognised by the World Health Organization as one of the twelve priority microorganisms displaying a concerning rise in fluoroquinolone resistance, presenting a significant challenge to public health (Whelan *et al*., 2019). This resistance phenomenon is attributed to a specific point mutation localised within the Quinolone Resistance Determining Region (QRDR) of the *gyrA* gene, a regulator of DNA supercoiling (Han *et al*., 2012). The C257T mutation leads to an amino acid substitution at position T86I within the protein sequence, strongly correlating with heightened fluoroquinolone resistance (Elhadidy *et al.*, 2020; Haldenby *et al.*, 2020). Recent studies indicate that fluoroquinolone-resistant isolates of *C. jejuni* may possess an enhanced virulence potential (Abraham *et al.*, 2020; Inglis *et al*, 2021). Moreover, these resistant variants have demonstrated the ability to form viable biofilms under aerobic conditions, aiding their invasion of epithelial cells (Whelan *et al*., 2019). Understanding the genetic basis and potential implications of fluoroquinolone resistance in *C. jejuni* is crucial for devising targeted therapeutic interventions and formulating effective public health strategies to address the escalating trends in resistance observed on a global scale.

Divergent trends in antimicrobial resistance within variants of the same bacterial species have been well-documented in various studies (Prestinaci *et al*, 2015; Munk *et al*., 2022). Specific variants, such as CC464, exhibit predominant resistance to fluoroquinolones, while others, such as CC 61, predominantly display susceptibility (Haldenby *et al*., 2020; Veltcheva *et al*., 2022). Moreover, projects such as IID2, INTEGRATE, and ENIGMA have provided evidence of the increasing prevalence of specific point mutations within these variant populations in the clinical cases (Haldenby *et al*., 2020). *Campylobacter* demonstrates a high genotypic-to-phenotypic mapping correlation, indicating a solid concordance between corresponding antimicrobial resistance genes and phenotypic resistance (Ocejo *et al*., 2021). Consequently, utilising bioinformatics to predict antimicrobial resistance from Whole Genome Sequencing (WGS) is very important for *C. jejuni* and *C. coli* within public health surveillance system (Painset *et al*., 2020).

Establishing Multi-locus Sequence Typing (MLST) schemes has significantly advanced epidemiological investigations (Maiden et al., 1998). Leveraging the housekeeping genes of *C. jejuni* and the formation of clonal-complexes, substantial progress in source attribution, understanding virulence factors, and studying global epidemiology (Dingle *et al*., 2001). A previous investigation focused on clinical isolates from the United Kingdom between 1998 and 2018 for fluoroquinolone resistance, utilising clonal-complexes as the classification basis  (Veltcheva *et al*., 2022). This study identified six distinct patterns in fluoroquinolone resistance within the clonal-complex classification. The analytical scope of this current study is expanded by employing the NeighbourGroups classification to discern variances in resistance patterns within the same dataset (Veltcheva *et al*., 2023). Additionally, by investigating tetracycline resistance patterns, an effort is made to elucidate whether specific variants have potentially developed resistance to both antibiotics.

Understanding the dynamics and mechanisms of antimicrobial resistance in *C. jejuni* is crucial for managing infections and devising effective preventive strategies. This study exploring the trends and patterns of antimicrobial resistance in *C. jejuni*, employing the NeighbourGroup classification system to enhance our understanding of the evolving landscape of this significant public health concern.

## Materials and Methods

### Meta Data Acquisition

Data are publicly available from the PubMLST database (Jolley, Bray and Maiden, 2018). The project used in this study can be accessible here: https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=127&submit=1. Using whole-genome sequence data for each of the isolates chosen from the PubMLST database, the gyrase A (*gyrA*) (CAMP0950) locus was queried. The *gyrA* nucleotide sequences for all isolates were aligned to one another using MAFFT/7.305. The nucleotide alignment was translated to an amino acid alignment using Transeq. Due to the absence of phenotypic data for many of the samples, the presence or absence of the Thr86Ile amino acid substitution was used to infer resistance or susceptibility, respectively. Although TetO is not part of the cgMLST scheme in the PubMLST database you can also query for presence or absence of the *TetO* gene. Thus, the presence was queried and used for this analysis. For this study only *TetO* were queried but *TetM* could also be looked at in the future.

### Software usage

The project uses Python (Van Rossum, G. , Drake, 2009). The temporal dynamics were visualised using Tableau ('Tableau (version. 9.1)', 2016). Finally, tables were created using Microsoft Excel (version 16.63). Code for the methods described is available here: https://github.com/bgrdessislava/Chapter6.

**Linear regression and heatmaps:**

Logistic regression has been used to investigate the fluoroquinolones and tetracycline resistance, with 1 indicating resistance and 0 indicating susceptibility. The log ratio is calculated through a Python package. Heatmaps are used to standardise the significance of the resistance between different NeighbourGroups, and Z-scores were used to identify this level. "NeighbourGroups" is created through a generalisable machine learning approach, that reproducibly, robustly, and rapidly classifies multi-locus sequence types with defined precision using CatBoost algorithm. The method is described here: https://github.com/bgrdessislava/NeighbourGroups

## Results

In this study, we have 14,481 isolates from the UK spanning from 1997 to 2022, from different sources: clinical, poultry, cattle, and environmental sources and more. Here, we observed trends in fluoroquinolones and tetracyclines resistance through the NeighbourGroups. This study uses logistic regression, heatmaps and clustering to investigate the trends of fluoroquinolones and tetracycline resistance.

### 1.) Overall antimicrobial resistance dynamics:

The logistic regression analysis of fluoroquinolones and tetracycline resistance demonstrated an escalation in resistance levels of *C. jejuni* to both antibiotics (Figure 21). The odds of *C. jejuni* being resistant to fluoroquinolones increased at a log ratio of 0.063 annually. For tetracycline, the odds increased at a log ratio of 0.026 per year within the UK. Consequently, the resistance to fluoroquinolones is rising slightly more than that of tetracycline each year. In 1997, the resistance probability for fluoroquinolones was approximately 0.18, while tetracycline stood at around 0.22. This disparity suggests a higher resistance to tetracycline than fluoroquinolones in the samples obtained in 1997. However, by 2022, the resistance probability for fluoroquinolones had surpassed the 0.4 mark, whereas tetracycline was just below 0.4. This trend suggests an increasing resistance within *C. jejuni*, with fluoroquinolone resistance escalating at a comparatively higher rate.

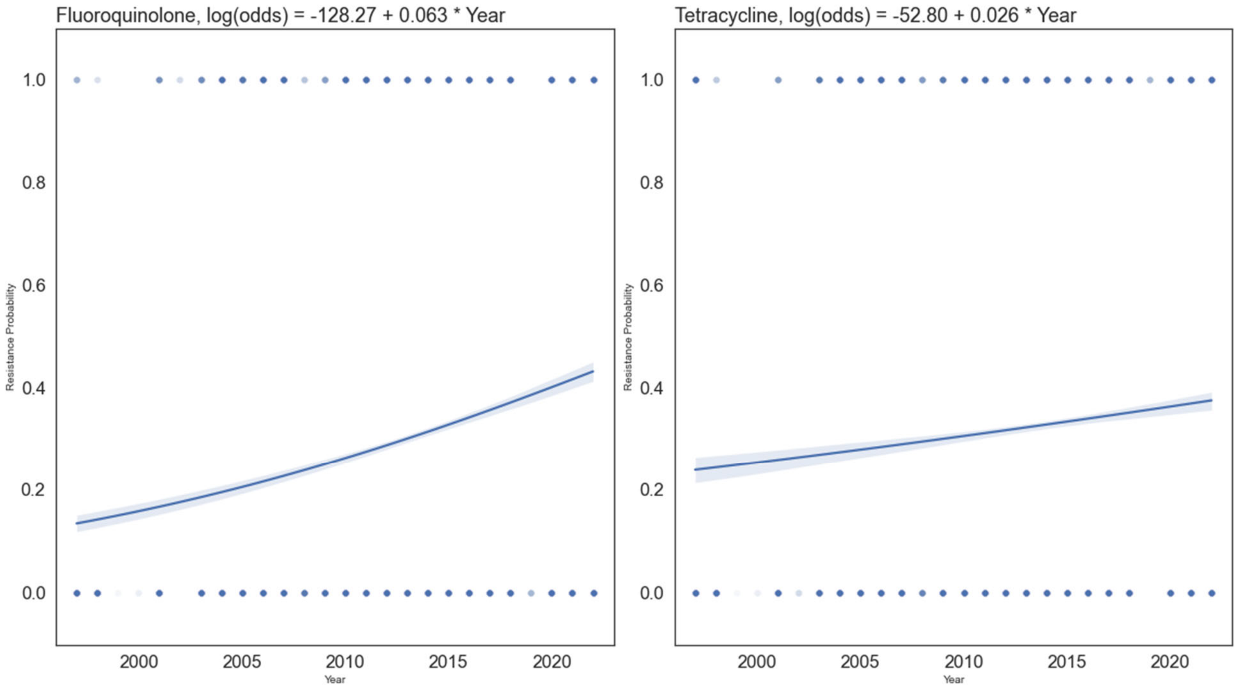


*Figure 21 Logistic regression of fluoroquinolone and tetracycline resistance across 1997 to 2022 in the UK in resistance probability. 1 Indicates resistance, and 0 indicates susceptible isolates. The confidence interval is shown as a shade around the logistic regression line. The shade of the dots illustrates the abundance of samples on the time point. Isolates that are resistant to both are not excluded from this figure.*

### 2.) NeighbourGroups pattern in antimicrobial resistance

Subsequently, an assessment was conducted to ascertain the antibiotic resistance profiles of NeighbourGroups—whether they demonstrate resistance to one antibiotic, both, or none (Figure 22). The analysis employed permutation testing utilising the Z score to ascertain the significance of observed differences. NeighbourGroup 1 and 2 exhibited a propensity to resist fluoroquinolones and tetracycline. NeighbourGroup 1 and 2 corresponded to clonal-complexes CC353, CC354, and CC464 (Supplementary: Figure 5). NeighbourGroup 7 displayed susceptibility to both antibiotics, primarily comprised of CC45 (Supplementary: Figure 5). NeighbourGroup 2 and 15 notably elevated resistance to fluoroquinolones but not to tetracycline, aligning with clonal-complexes CC353, CC464, and a sub-variant group of CC21 (Supplementary: Figure 5). Conversely, NeighbourGroup 9 (encompassing CC206 and CC48) and NeighbourGroup 13 (includes CC443, CC574, and others) demonstrated heightened tetracycline resistance but not fluoroquinolones resistance.

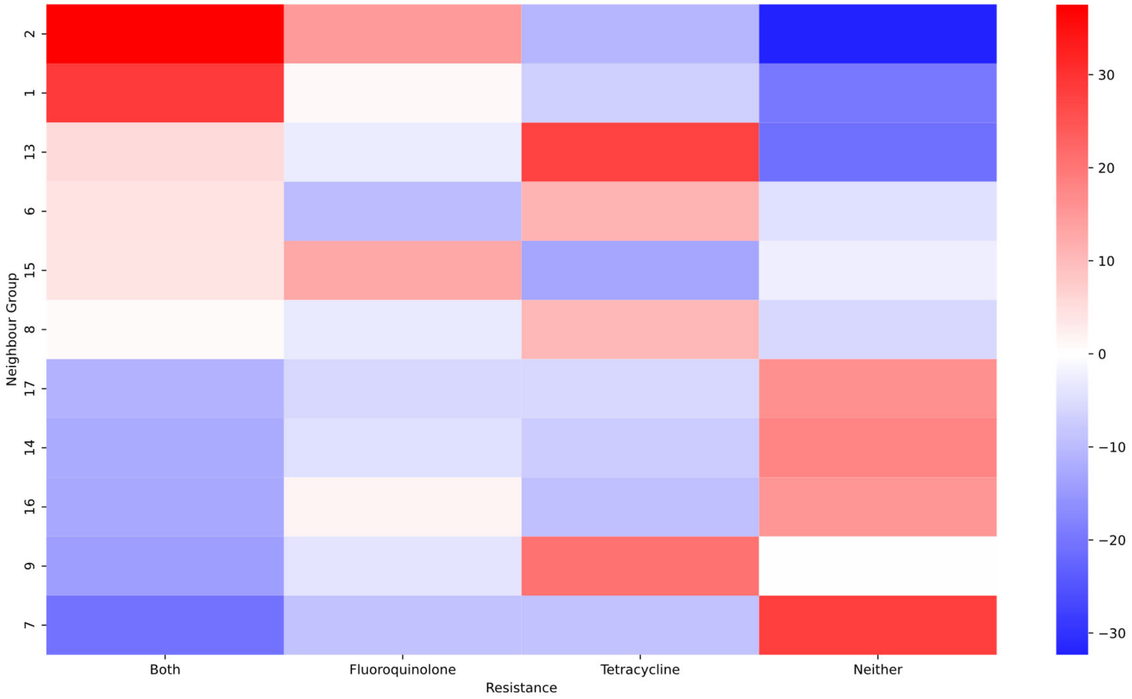
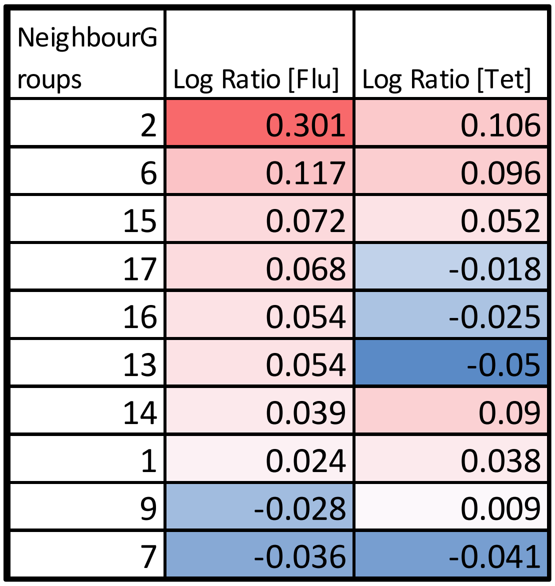


*Figure 22 Heatmap of fluoroquinolone only, tetracycline only, both resistance and neither demonstrated for NeighbourGroups. The scale is Z-score, with red indicating higher abundance and blue showing the lowest abundance3. The test uses a permutation test, showing the significance as a Z score.*

As previously discussed, the classification of CC21 has been split into NeighbourGroups 15 and 16 (Veltcheva *et al*., 2023). In this analysis, NeighbourGroup 16 exhibited susceptibility to both antibiotics, while NeighbourGroup 15 demonstrated resistance to fluoroquinolones, though not significantly for tetracycline (Figure 22). This finding supports the subdivision of CC21 into two distinct sub-groups at the variant level.

Next, the trends among NeighbourGroups over time, focusing on NeighbourGroups 1, 2, 6, 7, 9, 13, 14, 15, 16, and 17 rather than addressing every NeighbourGroup were completed. The summarised log ratio of these NeighbourGroups is presented (Figure 23). Four patterns were identified among NeighbourGroups: an increase in resistance to both antibiotics, an increase in fluoroquinolone resistance but a decrease in tetracycline resistance, a decrease in fluoroquinolone resistance but an increase in tetracycline resistance, and a decrease in resistance to both antibiotics. NeighbourGroup 2 displayed the highest increase rate in fluoroquinolone and tetracycline resistance; however, the rate of tetracycline resistance increase was comparatively less significant than that of other NeighbourGroups. On the contrary, NeighbourGroup 7 demonstrated a declining trend in resistance to both antibiotics. Additionally, although certain NeighbourGroups, such as NeighbourGroup 14, showed a rise in resistance, most isolates remained susceptible. Hence, further examination was conducted through logistic regression to understand these trends better.



| NeighbourGroups | Log Ratio [Flu] | Log Ratio [Tet] |
|---|---|---|
| 2 | 0.301 | 0.106 |
| 6 | 0.117 | 0.096 |
| 15 | 0.072 | 0.052 |
| 17 | 0.068 | -0.018 |
| 16 | 0.054 | -0.025 |
| 13 | 0.054 | -0.05 |
| 14 | 0.039 | 0.09 |
| 1 | 0.024 | 0.038 |
| 9 | -0.028 | 0.009 |
| 7 | -0.036 | -0.041 |

*Figure 23 NeighbourGroups Log Ratio for fluoroquinolones and tetracycline resistance*

### 3.) Neighbour Groups that increase resistance to both antibiotics (Focusing on NGroup – 1, 2, 6 and 15)

NeighbourGroup 1 exhibited an increase in resistance to both antibiotics (Figure 24). The odds of displaying resistance to fluoroquinolones increased at a log ratio of 0.024 annually. In contrast, the odds of being resistant to tetracycline increased at a log ratio of 0.038 each year within NeighbourGroup 1.

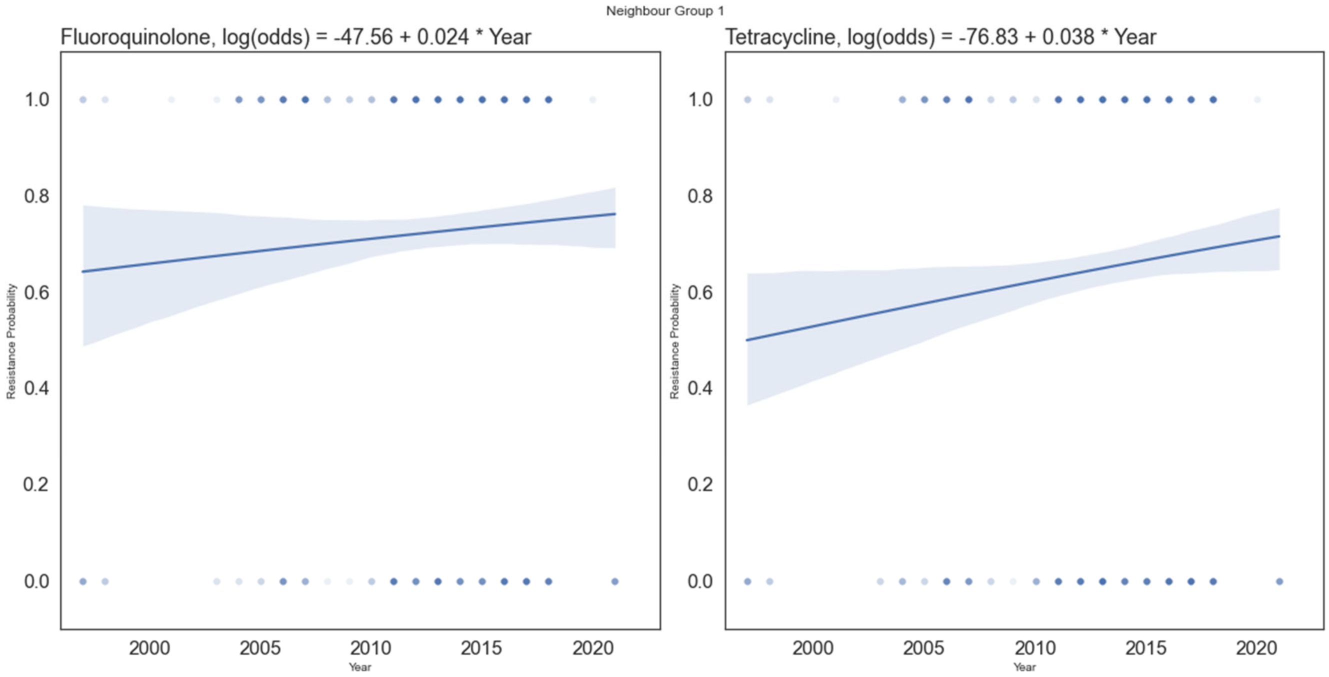


*Figure 24 NeighbourGroup 1 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Conversely, NeighbourGroup 2 displayed a sigmoidal increase in fluoroquinolone resistance and a substantial rise in tetracycline resistance (Figure 25). The odds of manifesting fluoroquinolone resistance escalated at a log ratio 0.301 per annum. In contrast, the odds of tetracycline resistance increased significantly at a log ratio of 0.106 annually for NeighbourGroup 2. A comparison between NeighbourGroup 1 and NeighbourGroup 2 reveals that in 1997, NeighbourGroup 2 demonstrated lower resistance to fluoroquinolones and tetracycline. In contrast, NeighbourGroup 1 already exhibited a higher resistance probability of 0.5 for both antibiotics.

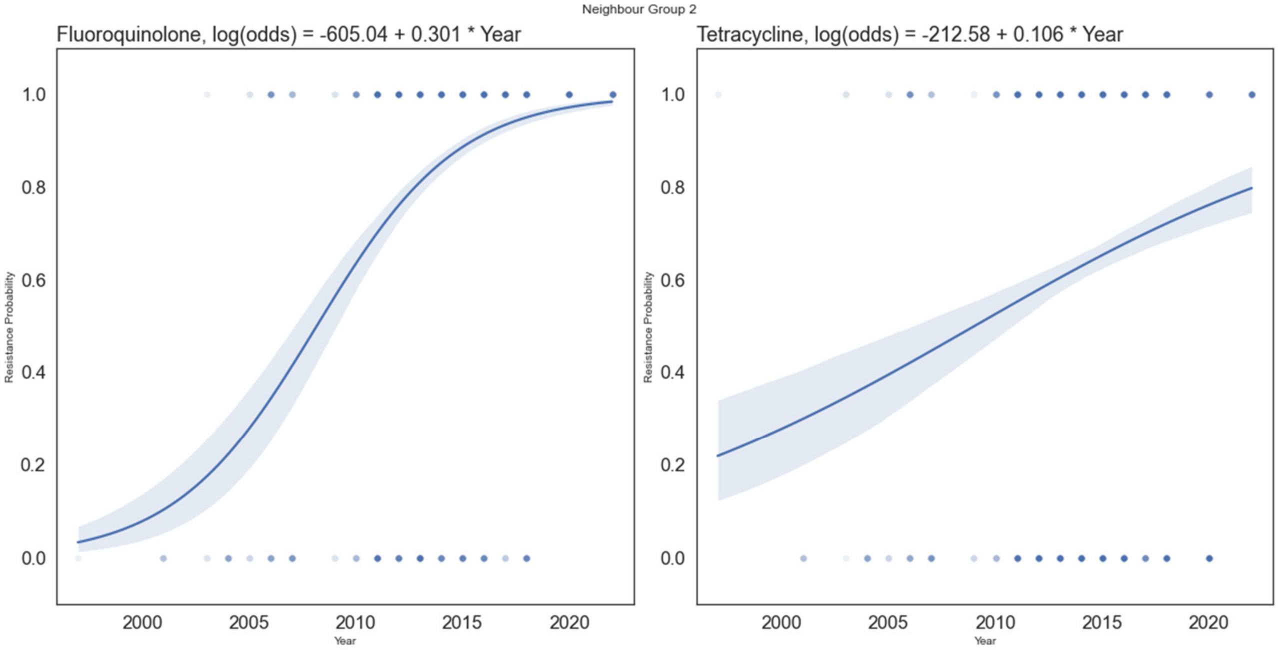


*Figure 25 NeighbourGroup 2 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Additionally, we analysed NeighbourGroup 6, which demonstrated increased resistance to both antibiotics. The odds of displaying resistance to fluoroquinolones increased at a log ratio of 0.117 annually, and for tetracycline, the odds of resistance increased at a log ratio of 0.096 each year (Figure 26). A comparative analysis with NeighbourGroups 1 and 2 revealed that NeighbourGroup 6 exhibited a similar log ratio to NeighbourGroup 1 and matched its 2022 resistance level. However, concerning changes in resistance, NeighbourGroup 6 displayed a pattern close to NeighbourGroup 2, maintaining a low resistance overall in 1997.

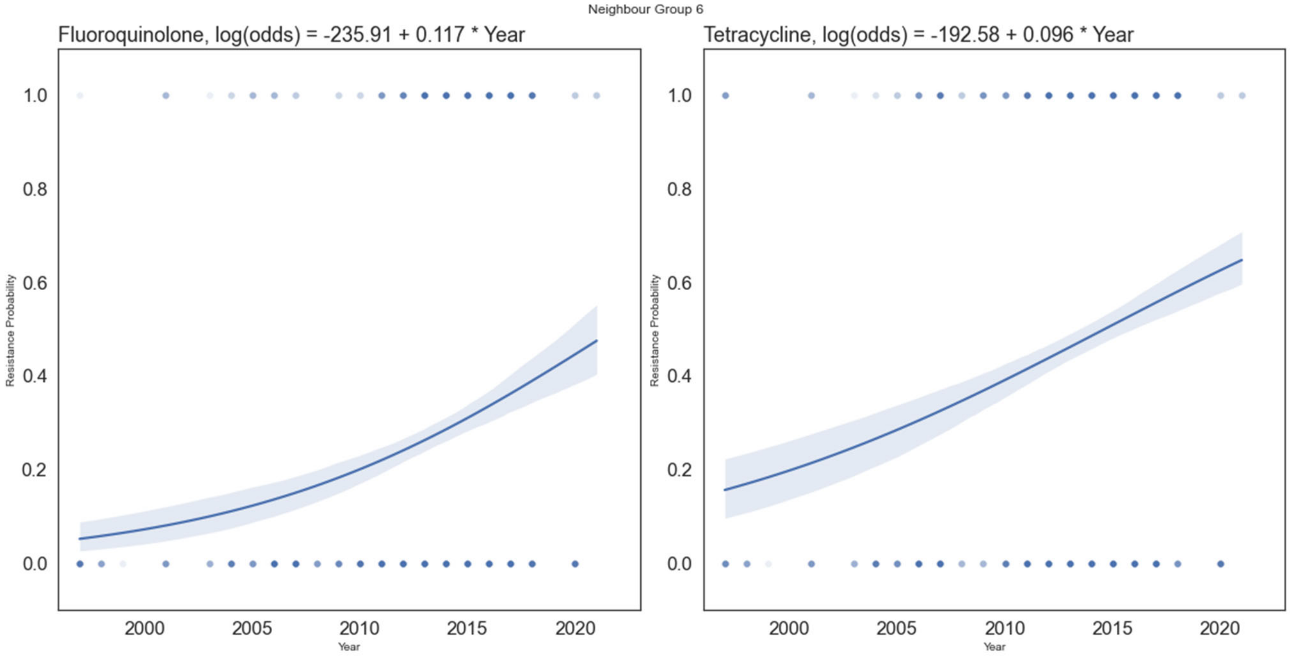


*Figure 26 NeighbourGroup6 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Another NeighbourGroup displaying a consistent increase in resistance to both antibiotics is NeighbourGroup 15. The odds of developing fluoroquinolone resistance increased at a log ratio of 0.072 annually. In contrast, for tetracycline, the odds of resistance increased at a log ratio of 0.052 each year within NeighbourGroup 15. Despite the log ratio being only slightly higher than that of NeighbourGroup 6, NeighbourGroup 15 exhibited an overall lower resistance probability for both antibiotics, hovering around 0.5. At the same time, NeighbourGroup 6 had a slightly higher resistance probability in 2022 (Figure 27 and Figure 26).

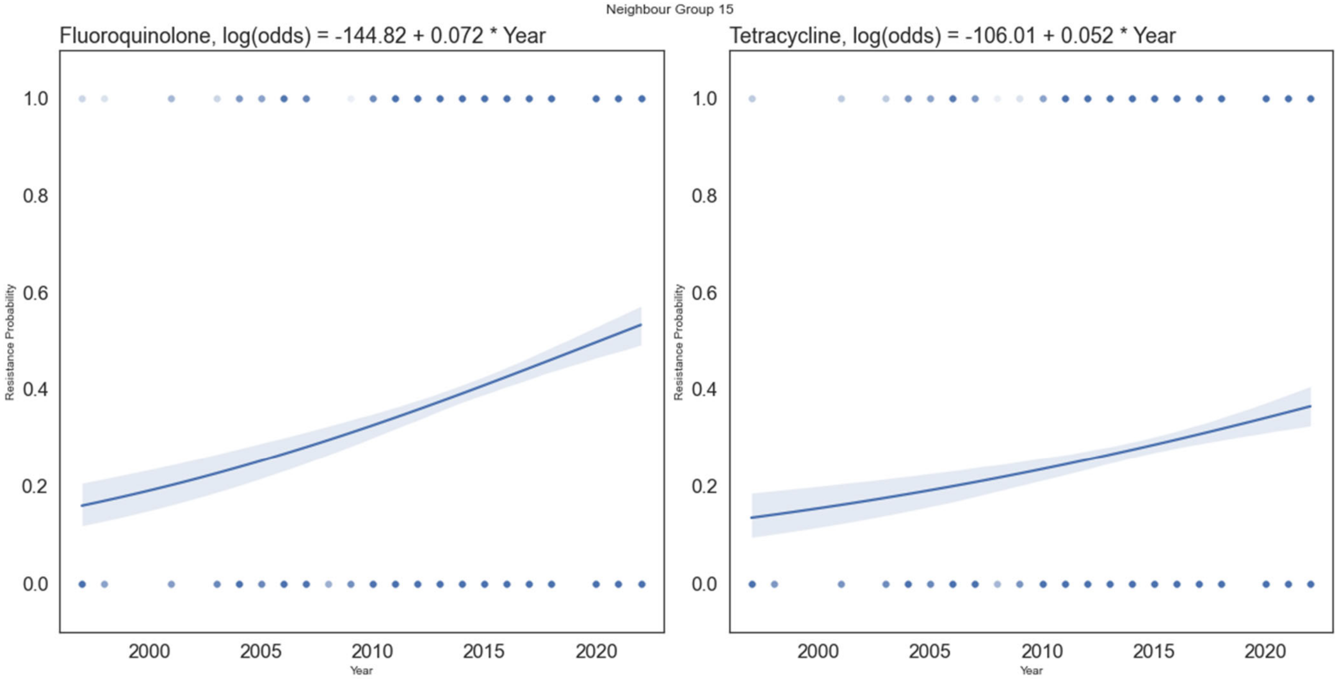


*Figure 27 NeighbourGroup 15 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Furthermore, an a bacterial populational dynamics were investigated by calculating each variants proportion of the whole year (Figure 28). Notably, NeighbourGroup 15 consistently increased the proportion of isolates across the years, signifying that in the last five years, NeighbourGroup 15 encompassed approximately 40% of all isolates. Conversely, some NeighbourGroups, such as NeighbourGroup 16 and NeighbourGroup 6, demonstrated a decline in their proportions over time. Additionally, eight NeighbourGroups consistently comprised less than 1% of isolates each year, highlighting an aspect of population dynamics for further exploration. This observation shows the importance of variant population dynamics when assessing the escalation of antimicrobial resistance.

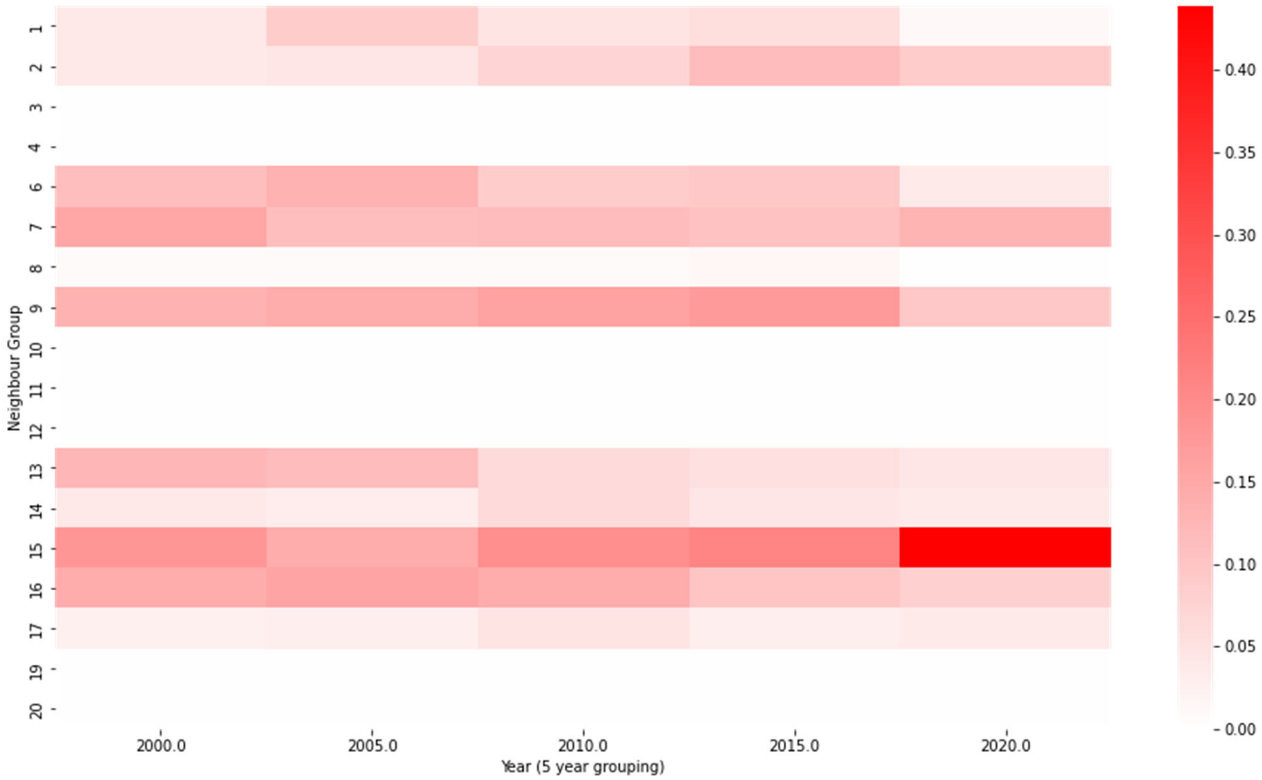


*Figure 28 Proportion of isolates within each NeighbourGroup at 5-year intervals. The scale is a percentage that equals 1 for each column. From 2017-2022 40% of isolates were from, NeighbourGroup 15*

## 4.) Neighbour Groups that are increasing in resistance but overall have a majority in susceptible isolates.

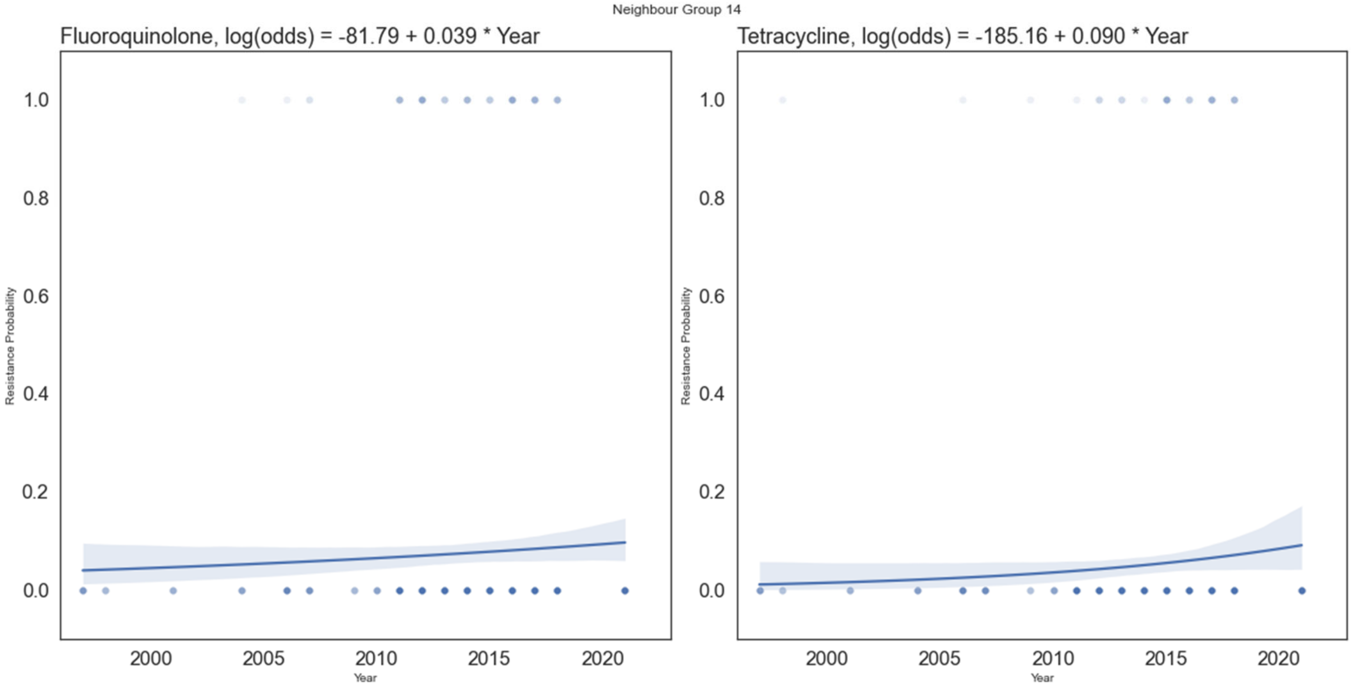


*Figure 29 NeighbourGroup 14 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Certain NeighbourGroups consistently demonstrated a low level of resistance, as evident in NeighbourGroup 14 and NeighbourGroup 17 (Figure 29 and Figure 30). For NeighbourGroup 14, the odds of developing resistance to fluoroquinolones increased at a log ratio of 0.039 per year, and for tetracycline, the odds of resistance increased at a log ratio of 0.090 annually. Despite a minor upward trend, the resistance probability did not exceed 0.2 by 2022, signifying this variant's overall high susceptibility profile.

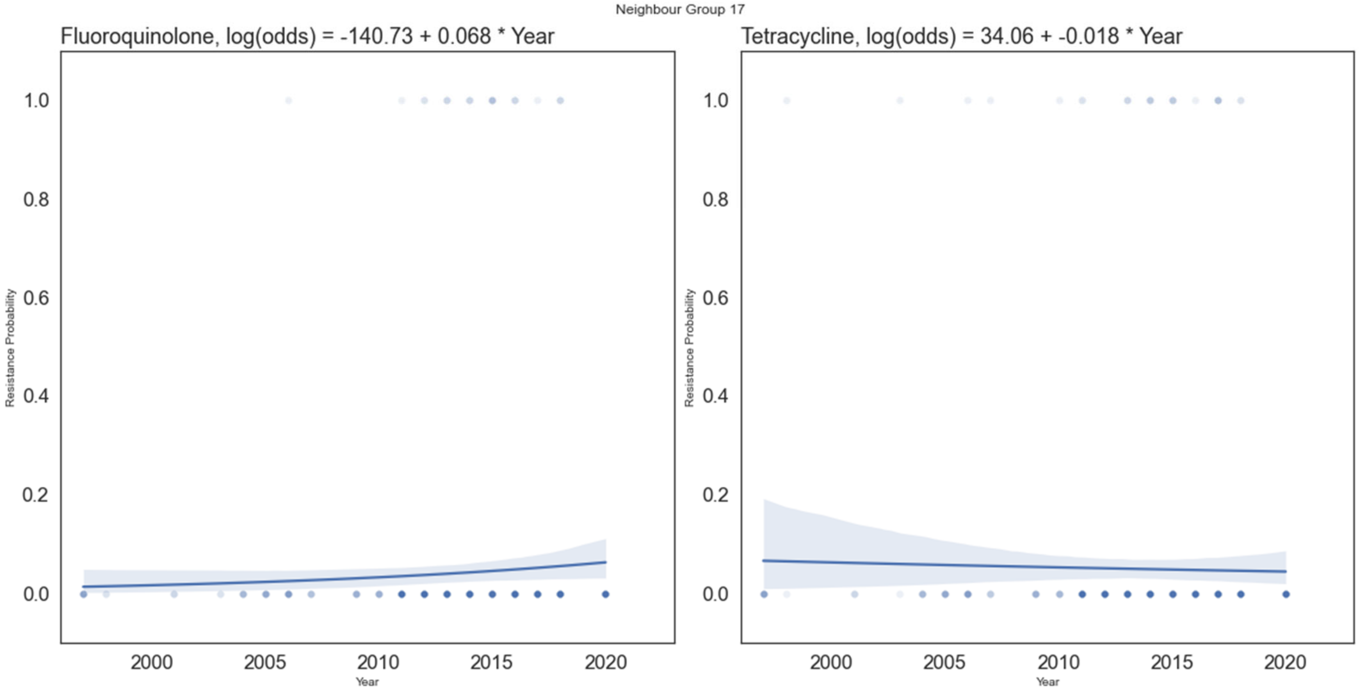


*Figure 30 NeighbourGroup 17 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

Furthermore, NeighbourGroup 17's odds of exhibiting resistance to fluoroquinolones increased at a log ratio of 0.068 annually. In contrast, for tetracycline, the odds of resistance displayed a minor decrease at a log ratio of -0.018 each year.

In contrast, NeighbourGroup 7 showed a slight decline in antibiotic resistance (Figure 31). The odds of being resistant to fluoroquinolones decreased at a log ratio of -0.036 annually. For tetracycline, the odds of resistance decreased at a log ratio of -0.041 each year within NeighbourGroup 7. Although the reduction was not substantial, this variant has a lower resistance level than other NeighbourGroups and follows a distinct trend.

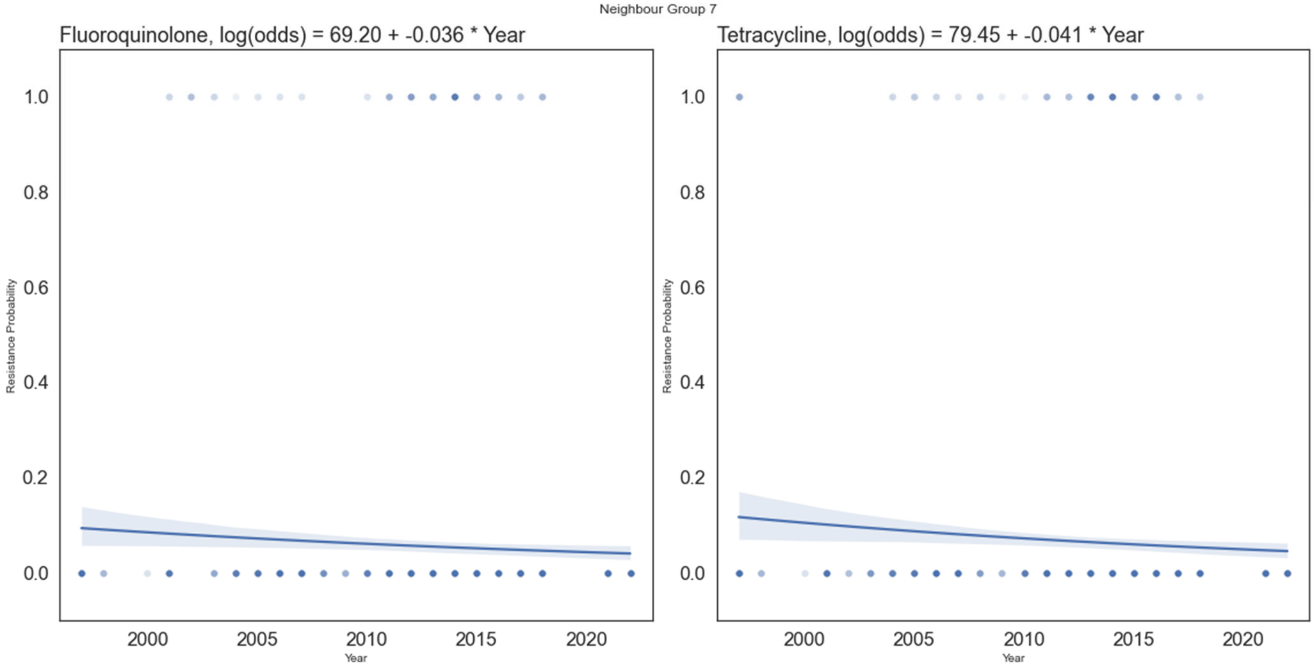


*Figure 31 NeighbourGroup7 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

### 5.) NeighbourGroups with different trends between antibiotics

Throughout the analysis, a consistent pattern emerged within certain NeighbourGroups, where trends in resistance either increased or decreased simultaneously, observed notably in NeighbourGroups 1, 2, 6, and 7. However, in the case of NeighbourGroup 9, a deviation from this pattern was apparent, with a slight decrease in fluoroquinolone resistance while maintaining a relatively stable resistance level for tetracycline, with a minor increase observed over time. The odds of being resistant to fluoroquinolones decreased at a log ratio of -0.028 each year, whereas tetracycline resistance increased at a log ratio of 0.009 annually (Figure 32).

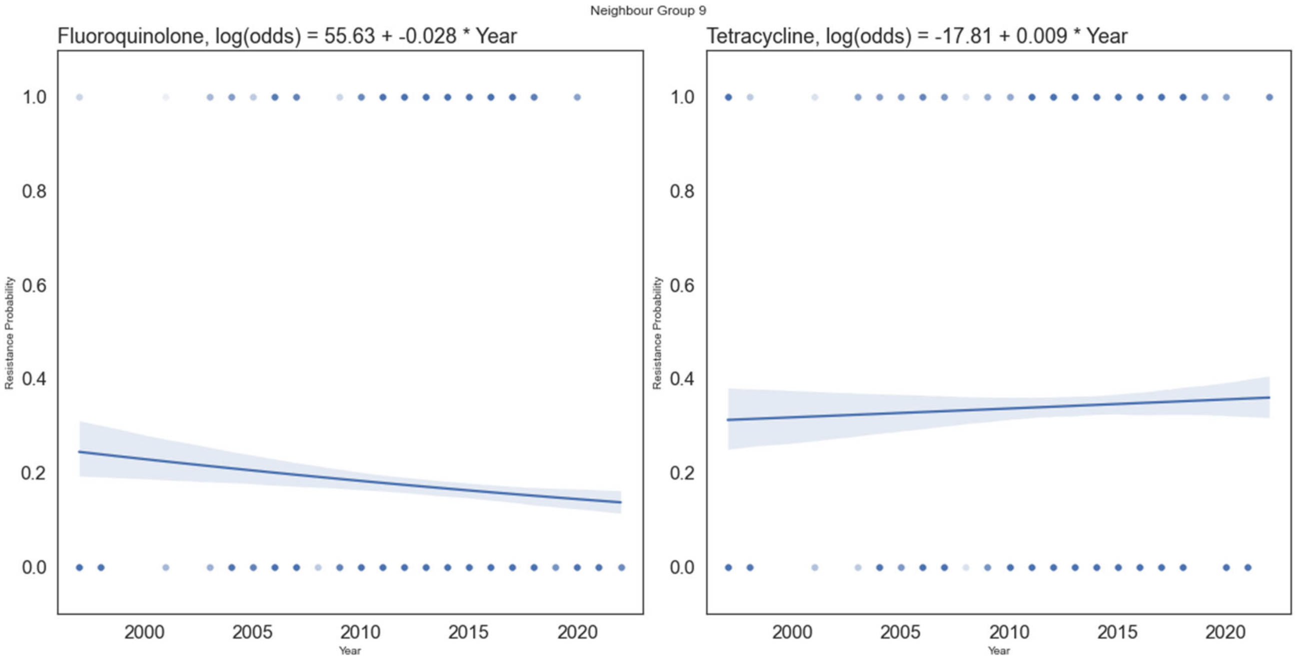


*Figure 32 NeighbourGroup 9 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

In contrast, NeighbourGroup 13 exhibited an opposing trend. The odds of displaying resistance to fluoroquinolones increased at a log ratio of 0.054 annually. Whereas for tetracycline, the odds of resistance decreased at a log ratio of -0.050 each year (Figure 33). Compared to NeighbourGroup 9, this demonstrates a contrary pattern, with fluoroquinolone resistance increasing while tetracycline resistance decreasing in NeighbourGroup 13. (Figure 33).

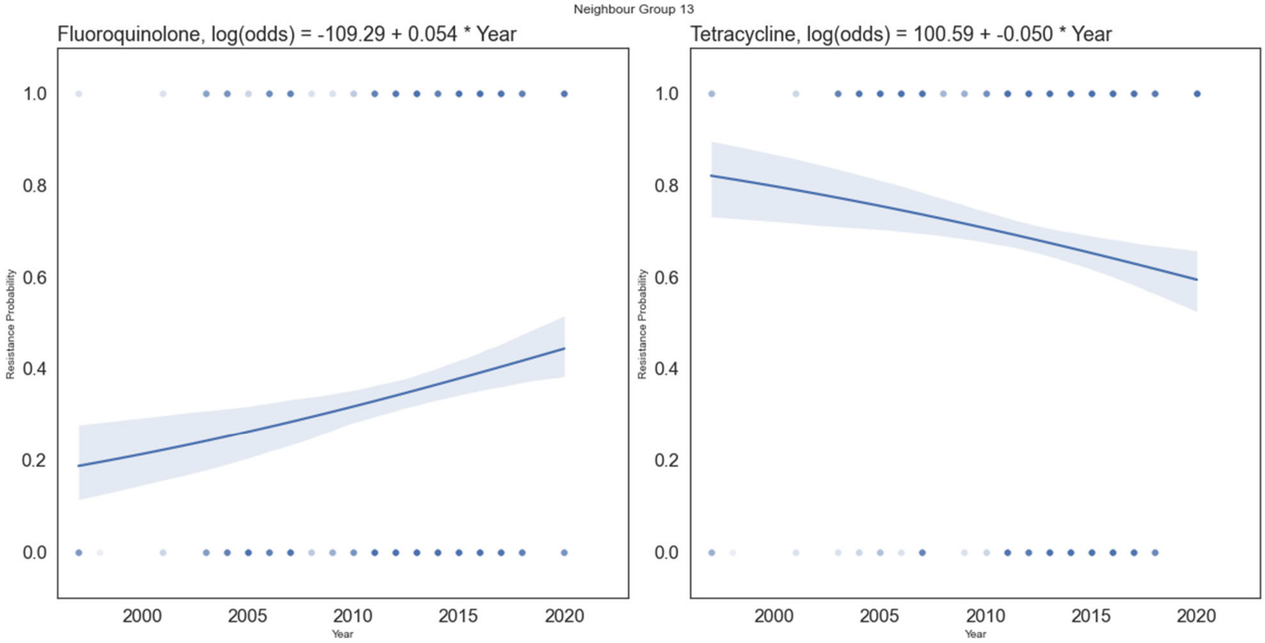


*Figure 33 NeighbourGroup 13 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

NeighbourGroup 16 displayed a subtle increase in resistance to fluoroquinolones and a minor decrease in resistance to tetracycline (Figure 34). The odds of developing fluoroquinolone resistance increased annually at a log ratio 0.054. In contrast, tetracycline's odds of resistance slightly decreased at a log ratio of -0.025 each year. Notably, although the yearly rate of change was comparable to that of NeighbourGroup 13, NeighbourGroup 16 maintained its overall resistance below 0.3. In comparison, its resistance level to tetracycline stayed high at approximately 0.6 by 2022 (Figure 33).

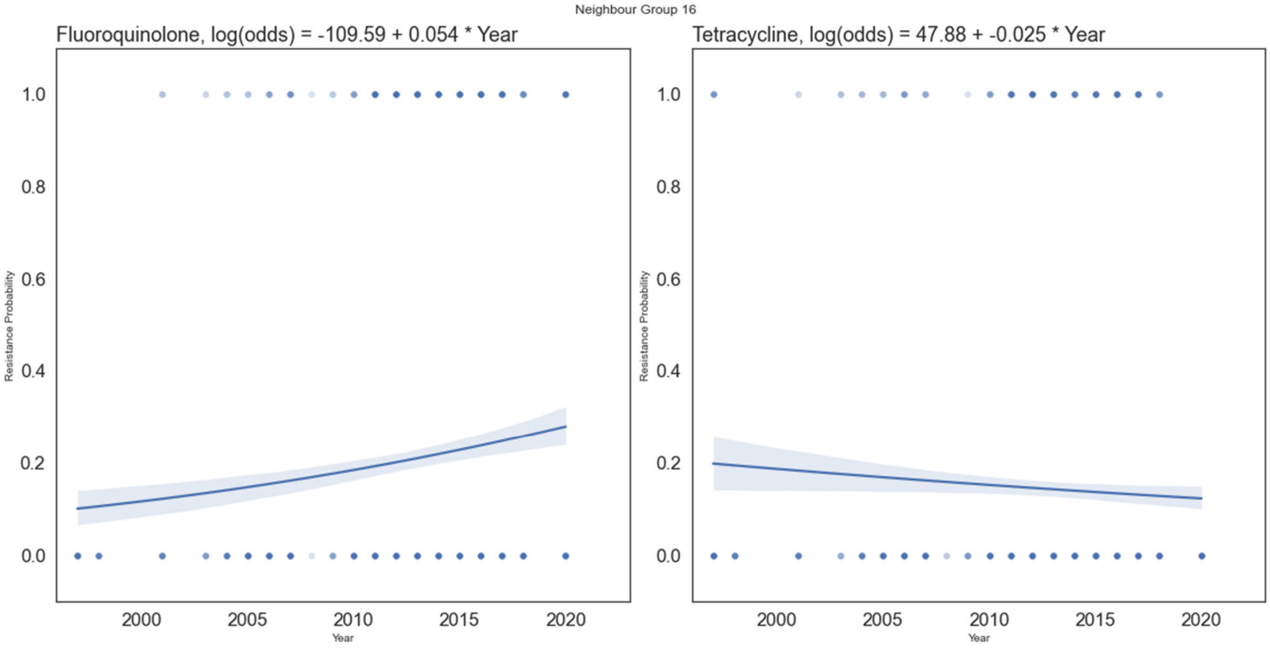


*Figure 34 NeighbourGroup 16 fluoroquinolone and tetracycline resistance logistic regression between 1997 to 2022. The light blue shade is the confidence interval.*

### 6.) Clustering antimicrobial resistant patterns in NeighbourGroups and Clonal-complex

The 16 clonal-complexes and 11 NeighbourGroups were analysed with a sufficient yearly sample size to depict resistance proportion over time (Figure 35). In the NeighbourGroups, two significant clusters were evident. The first cluster comprised NeighbourGroups 1 and 2, while the second cluster exhibited four sub-clusters. The first sub-cluster included NeighbourGroup 8, which appeared as an outlier due to limited isolate counts. The second and third sub-clusters represented mixed groups, lacking clear signals regarding fluoroquinolone resistance. The fourth sub-cluster formed a susceptible cluster encompassing NeighbourGroups 7, 14, and 17.

On the other hand, when considering clonal-complexes, two main clusters were observed. Although smaller branches within the clusters were apparent, the clonal-complex analysis lacked data coverage across earlier years. Reducing the number of classifications from 35 clonal-complexes to 20 NeighbourGroups facilitated a more granular view of data points in earlier years, spanning between 2003-2006.

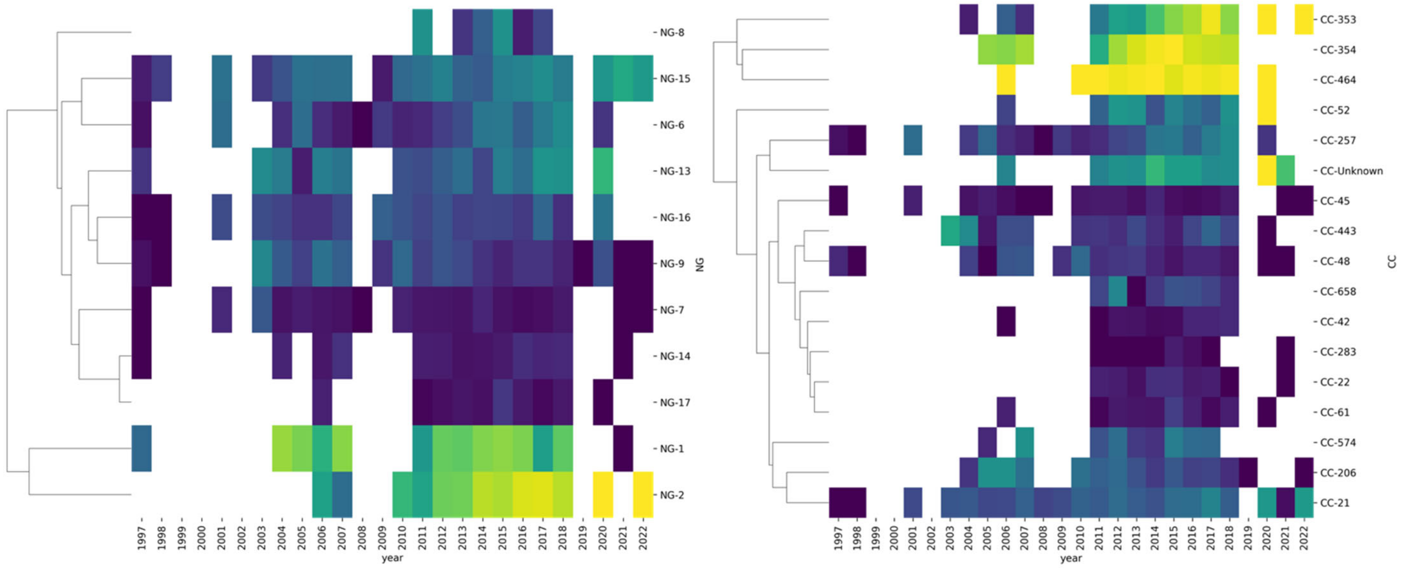


*Figure 35 Heat maps illustrating fluoroquinolone resistance proportion between 1997-2022 for clonal-complex and NeighbourGroups. Yellow indicates a high resistance proportion of 1, and purple indicates a low proportion of 0, showing low resistance. Isolates with more than 10 per clonal-complex per year have been included. White space is no samples, or it is <10, indicating that there are not enough samples for valid observation and have not been included in the visualisation.*

Tetracycline resistance presented a distinct pattern compared to fluoroquinolone resistance (Figure 36). A resistant cluster was observed in the clonal-complex analysis, prominently featuring CC464, CC354, and CC574. On the other hand, the NeighbourGroups analysis did not demonstrate a distinct, highly resistant cluster. Instead, a mixed range of resistance levels was evident, particularly in NeighbourGroups 1, 2, 8, and 13, with NeighbourGroups 2 and 13 displaying elevated resistance by 2022. Unlike clonal-complexes, the NeighbourGroups tended towards mixed or susceptible groupings, which revealed four distinct clusters: high resistance, moderate resistance, lower susceptibility, and susceptible cluster.

A significant advantage of utilising NeighbourGroups clustering was evident, particularly in revealing fluoroquinolone trends (Figure 35). The transition from clonal-complex level analysis to NeighbourGroup level allowed for retrieving data from the low isolate count between 1997-2007, providing a more precise pattern that was previously obscured.

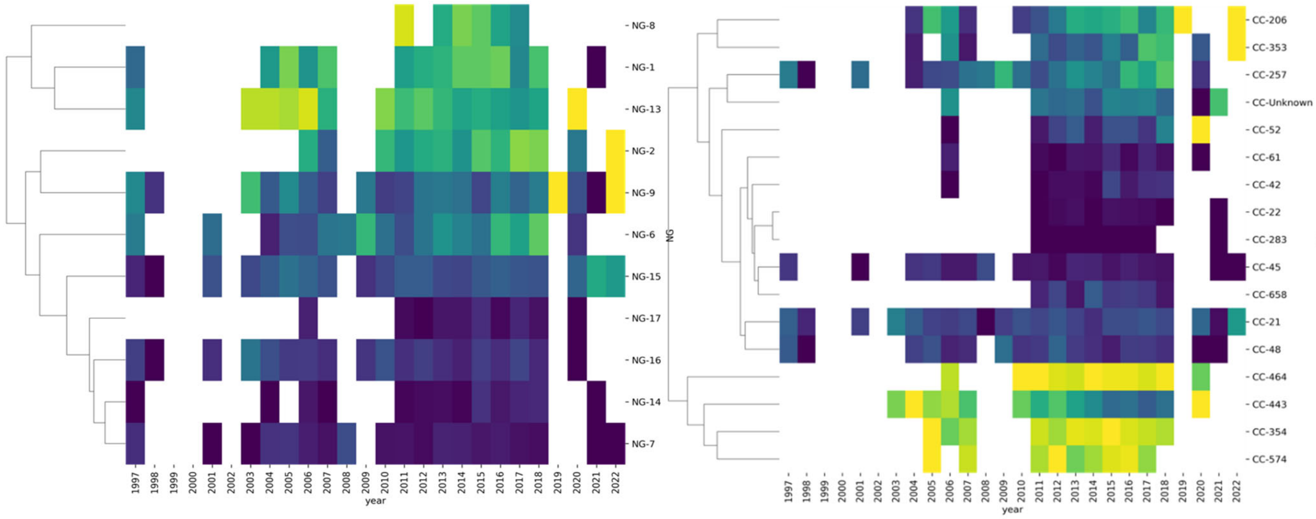


*Figure 36 Heat maps illustrating tetracycline resistance proportion between 1997-2022 for clonal-complex and NeighbourGroups. Yellow indicates a high resistance proportion of 1, and purple indicates a low proportion of 0, showing low resistance. Isolates with at least ten isolates for each clonal-complex have been included in the study. White space is no samples, or it is <10, indicating that there are not enough samples for valid observation and have not been included in the visualisation.*

In summary, our investigation of NeighbourGroups has revealed distinct patterns in resistance to fluoroquinolones and tetracycline. NeighbourGroups 1 and 2 exhibited significantly elevated resistance to both antibiotics, while NeighbourGroup 13 displayed substantial resistance to tetracycline but not fluoroquinolones. NeighbourGroup 15 demonstrated higher resistance to fluoroquinolones than tetracycline, constituting approximately 40% of NeighbourGroup proportions within the last five years. Hence, examining NeighbourGroups and their resistance profiles and population dynamics is crucial for future insights. We also observed that NeighbourGroup 7 did not have resistance to either antibiotic. From the perspective of logistic regression values, NeighbourGroup 2 exhibited the highest value for fluoroquinolones and tetracycline. Following closely, NeighbourGroup 6 displayed the second-highest rate. Notably, while NeighbourGroup 13 demonstrated the highest resistance among NeighbourGroups to tetracycline, it also showed the most significant reduction in tetracycline resistance. Hence, combining insights from logistic regression, resistance Z scores across time, and clustering is imperative for a comprehensive understanding.

## Discussion

Our study focused on *C. jejuni* resistance to fluoroquinolones and tetracycline. We found different trends within the same species. The odds of being resistant to fluoroquinolones increased by a log ratio of 0.063 per year, while resistance to tetracycline increased by 0.026 per year for *C. jejuni* in the UK. Initially, tetracycline resistance was higher than fluoroquinolone resistance in 1997, but by 2022, fluoroquinolone resistance had surpassed the overall resistance, reaching approximately 0.4 (Figure 21), in agreement with previous predictions (Veltcheva *et al*., 2022). This escalating resistance highlights the need for ongoing monitoring, strict antibiotic stewardship, and the development of effective strategies to manage *C. jejuni* infections.

### Why is NeighbourGroup 15 increasing in size? (CC21, CC658)

While NeighbourGroups have not been extensively discussed in existing literature, our analysis reveals a substantial increase in the size of NeighbourGroup 15 over time (Figure 28). This NeighbourGroup includes clonal-complex 21, recognised as a source generalist that can thrive in multiple hosts (Dearlove *et al*., 2016; Mouftah *et al*., 2022). In contrast, clonal-complex CC658, part of NeighbourGroup 15, has not been extensively studied, and patterns within this group remain unknown.

One plausible reason for the observed increase in NeighbourGroup 15 could be its generalist nature, allowing it to thrive in diverse hosts. NeighbourGroup 15 exhibits increased antibiotic resistance, suggesting its ability to outcompete susceptible isolates and acquire a fitness advantage to survive throughout the food chain. Previous research has shown that specific genes, such as cj1377c, are overrepresented in clonal-complex CC21 and CC45, aiding survival within poultry processing plants (Yahara *et al*., 2017). Local clonal spread of CC21, with a significant proportion displaying fluoroquinolone resistance, has been observed in multiple regions, indicating prevalent selection pressure (Kovač *et al*., 2014). In Japan, a study from 2000 to 2017 has shown that a specific sequence type within CC21, ST4526, used to not exist in the earlier years(2000-2008) but increased rapidly afterwards with high resistance to ciprofloxacin in 2017 (54.4%) (Yamada *et al*., 2019). Moreover, *in vivo* studies have demonstrated the competitive advantage of fluoroquinolone-resistant variants out competing susceptible isolates (Luo *et al*., 2005). These findings emphasise the complex interplay between resistance mechanisms, fitness advantages, and the adaptability of *C. jejuni* variants in diverse ecological niches.

### NeighbourGroup 7 – susceptible (CC45, CC283)

NeighbourGroup 7, conversely, clonal-complex 45 is a variant known to reside in cattle and some river waters (McCarthy *et al.*, 2012). A study has shown that CC45 and CC283 have seasonal summer peaks in the UK, New Zealand and Finland (Habib, Uyttendaele and De Zutter, 2010; Friedrich *et al.*, 2016). CC45 is known to be resilient to oxidation, making it survive better in freezing conditions (Hur *et al.*, 2022). Thus, studies have demonstrated that CC45 has a lower chance of being transmitted through food but through untreated water and indirect exposures from farm animals such as cattle or pets exposed to the variant. CC45 is known also to be part of the generalist variant that has been found in different sources (Thépault *et al.*, 2017).

Although UK isolates from CC45, do not mainly come from poultry origin, South Korean and Irish studies have shown that their highest variant is CC45 (Wei, Kang and Jang, 2019; Truccollo *et al.*, 2021). Although for both studies, the sample size was small (Ireland – 30 isolates, South Korea – 60 isolates), in South Korea, the CC45 isolates came mainly from Mallard and Mandarin ducks and in Ireland it was found from both clinical and poultry isolates (Wei, Kang and Jang, 2019; Truccollo *et al.*, 2021).

### NeighbourGroups 1 has the highest resistance increase, correlated to CC354 and CC52.

Until now, many studies grouped CC353, CC354, and CC464 into similar patterns in source attribution. However, in our research, CC354 and CC52 were grouped into the same group instead (NeighbourGroup 1). NeighbourGroup 1 has stronger resistance to fluoroquinolones and tetracycline resistance. Another study conducted in Sri Lanka has shown that CC52 was resistant to ampicillin, ciprofloxacin, tetracycline and nalidixic acid, leading to being one of the multidrug-resistant variants, causing significant concerns (Tegner *et al.*, 2019). At the same time, CC464 and CC353 were grouped into NeighbourGroup 2. This group is associated with being more resistant to fluoroquinolones and not tetracycline. NeighbourGroup 2’s CC353 and CC464 are the two clonal-complexes heavily related to being a poultry specialist. A source attribution study in Switzerland also showed that all isolates from CC464 were resistant to fluoroquinolones, agreeing with our results (Kittl *et al.*, 2013). Through our earlier study (chapter 5:CC353 re-evaluated), we have shown that CC353 is composed of multiple sub-groups. (Figure 6). Thus, CC353 is a variant that should not be used as one clonal-complex.

## Conclusion

This study comprehensively investigated the resistance patterns of *C.jejuni* to fluoroquinolones and tetracycline from 1997 to 2022. Our analysis revealed an overall increasing trend in fluoroquinolone and tetracycline resistance within the *C. jejuni* species. NeighbourGroups 1 and 2 (CC52, CC353, CC354, CC464, and others) exhibited significant antibiotic resistance. On the other hand, NeighbourGroup 13 (predominantly CC443 and CC574) displayed the highest tetracycline resistance, with a rapid decrease in resistance levels. The NeighbourGroup 15 (a sub-group of CC21) resistance to fluoroquinolones but not to tetracycline and its proportion within the population surged, constituting nearly 40% of the overall population during 2017-2022. NeighbourGroup 7 (CC45) also displayed a marked susceptibility to both antibiotics and consistently decreased its resistance. The NeighbourGroups classification proved to clarify diverse antimicrobial resistance patterns within *C. jejuni.* Further investigation into the differences underlying resistance patterns among variants is required for a deeper understanding of antimicrobial resistance dynamics in *C. jejuni*.

# Supplementary

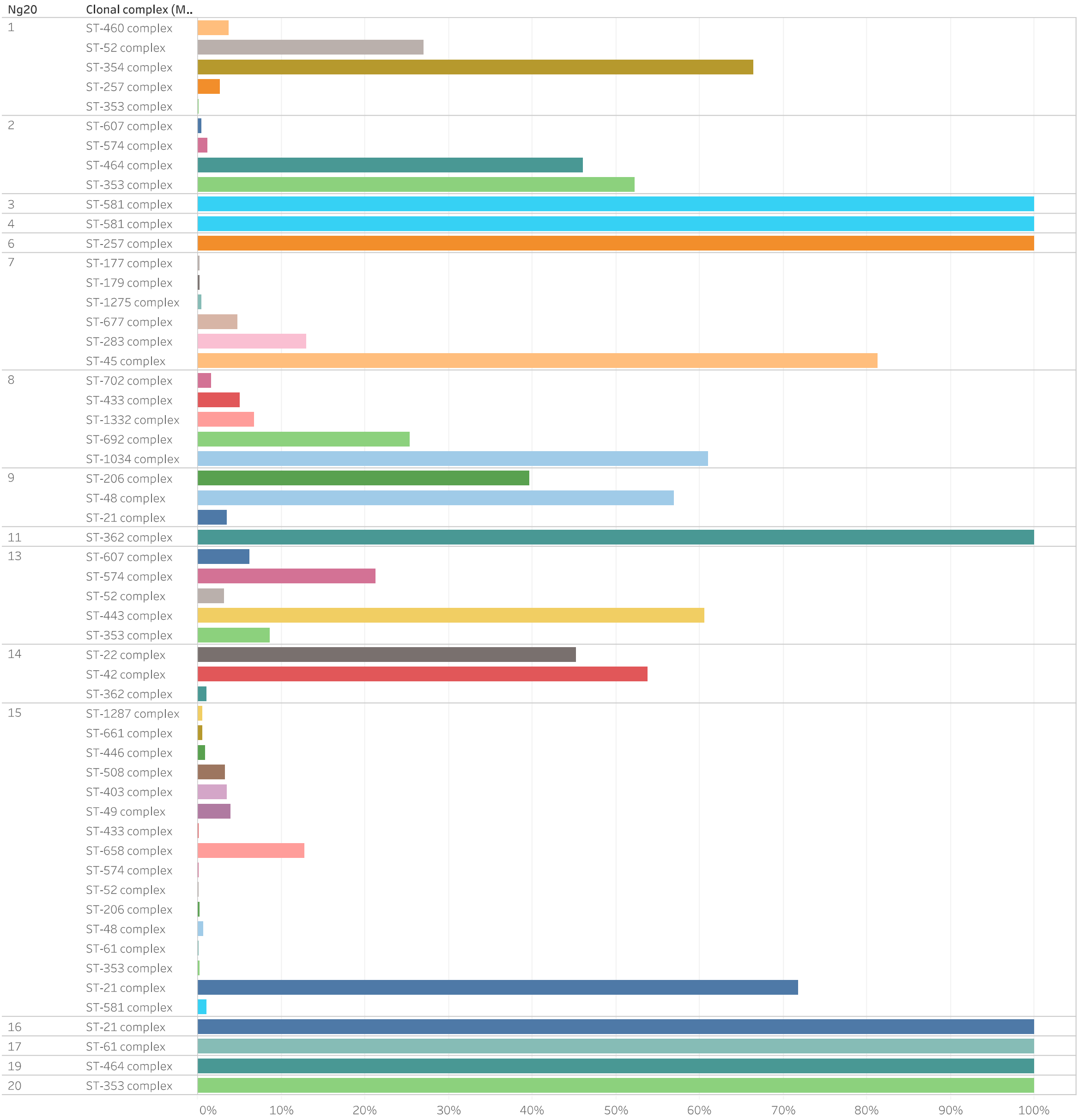


*Figure 37 NeighbourGroups and their clonal-complex isolates composition. The colour indicates the clonal-complex. Each of the NeighbourGroups adds up to 100%.*

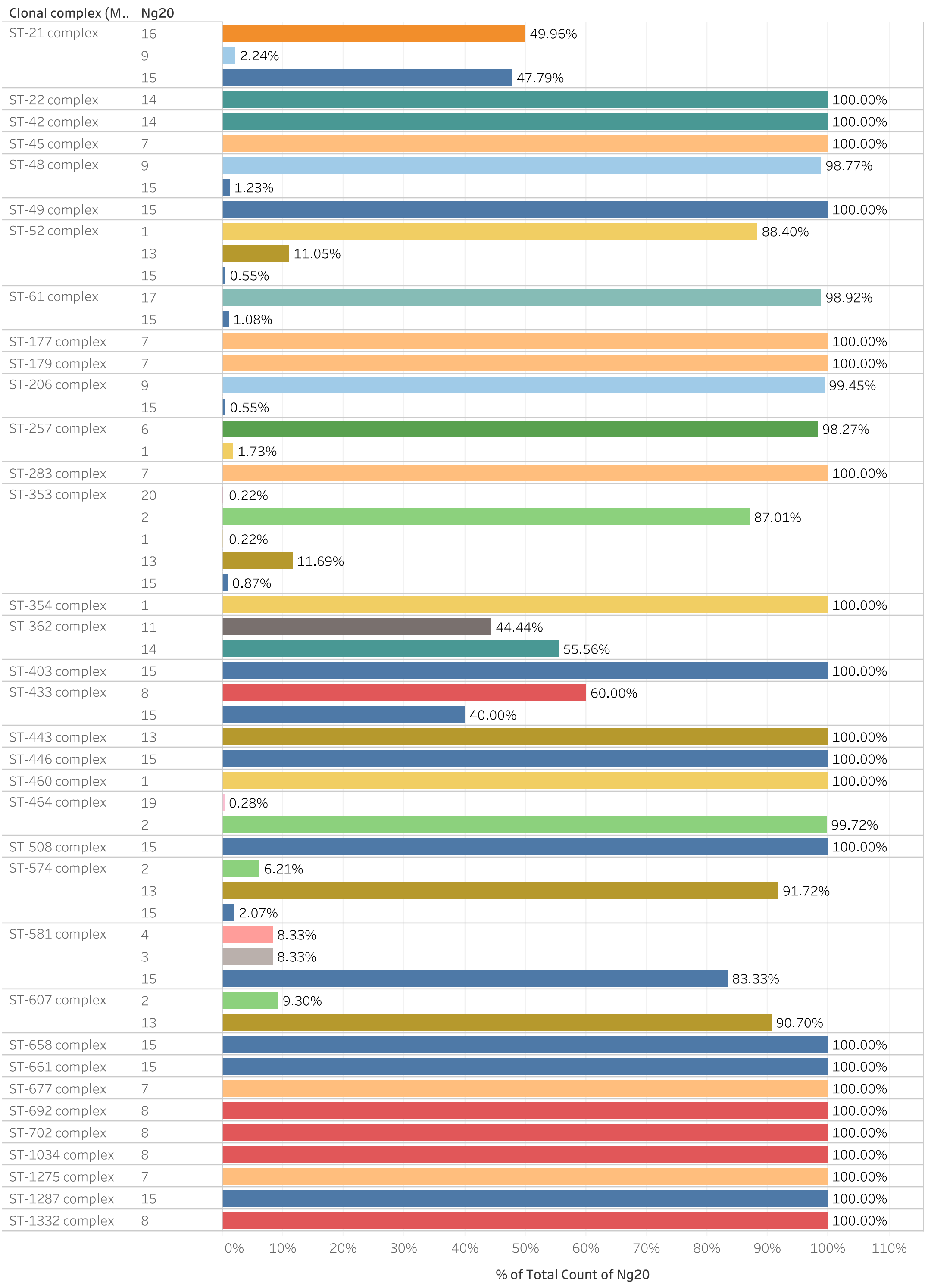


*Figure 38 Bar chart showing clonal-complex and their NeighbourGroups assignments. The colour indicates the NeighbourGroups. Each of the clonal-complex adds up to 100%.*

**Statement of Authorship for joint/multi-authored papers for PGR thesis**
To appear at the end of each thesis chapter submitted as an article/paper

# Chapter 6:

| | |
|---|---|
| Title of Paper | Trends in Fluoroquinolone and Tetracycline Resistance of *Campylobacter jejuni* using NeighbourGroups |
| Publication Status | ☐Published ☐ Accepted for Publication<br>☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Authors: Dessislava Veltcheva, Michael B, Bonsall, Martin C. J. Maiden<br>Paper Style: Microbial Genomics from Microbiology Society (Research Article) |

Student Confirmation

| | | | |
|---|---|---|---|
| Student Name: | Dessislava Veltcheva | | |
| Contribution to the Paper | e.g. performed analysis on all samples, interpreted data, wrote manuscript.<br>• Conceptualization of the research project<br>• Establishing Methodology<br>• Software and Script creation<br>• Validating results<br>• Formal analysis and Investigation<br>• Data curation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project management | | |
| Signature | | Date | 02.10.2023 |

Supervisor Confirmation

| | | | |
|---|---|---|---|
| Supervisor name and title: Professor Michael Bonsall | | | |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. | | | |
| Signature | MBBonsall | Date | 2nd October 2023 |

# Chapter 7: Clonal-complex revisited: Which clonal-complex will stay the same, will change, and which will retire

Clonal-complex revisited: Which clonal-complex will stay the same, will change, and which will retire.

Research Paper Title: Clonal-complex re-assignment for *Campylobacter jejuni* : The Good, The Split and The Retirement

Dessislava Veltcheva, Margaret Varga, Michael B, Bonsall, Martin C. J. Maiden

***Affiliation*** University of Oxford

***Corresponding author***: bgrdessislava@gmail.com

***Keywords***: *Campylobacter jejuni, multi-locus typing scheme,* clonal-complex, phylogenetic tree, variants, re-classification, microbiology, pathogen

***Repositories***: *Project ID:121* global_campyjejuni_1997=(n=17,338), which can be found in PubMLST public projects: https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=121&submit=1

## Abstract

*Campylobacter jejuni* (*C. jejuni*) research frequently utilises clonal-complexes as an informal way to define central genotypes and categorise isolates. *C. jejuni* contains 34 clonal-complexes, and these classifications have been used widely since their establishment in 2001. In this study, we re-evaluated these clonal-complexes by

exploring whether their central genotypes match the majority alleles for each clonal-complex. We used 16,273 isolates across various countries and multiple sources with an assigned clonal-complex. This study identified that 16 clonal-complexes have their most prevalent alleles matching their central genotypes. On the other hand, 12 clonal-complexes no longer align their original central genotypes. However, three clonal-complexes (CC21, CC353, and CC581) are in the caution category, where significant concerns have been raised about keeping them as clonal-complexes, even though their overall alleles match. Especially for CC21, two sub-variants have been discovered, and CC353 has been further identified to be made of multiple sub-variants in another study (Chapter 5). On the other hand, clonal-complexes with low overall account (8 clonal-complexes) have been discussed to be retired as a classification due to their lack of quantity. We further discuss the groupings using NeighbourGroups.

## Data Summary

Data are open and accessible and are found through PubMLST (https://pubmlst.org/organisms/campylobacter-jejunicoli/) (Jolley, Bray and Maiden, 2018). Specific isolate data used in this study is available through the PubMLST database with Project ID:121 global_campyjejuni_1997=(n=17,338): https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=121&submit=1

## Gap statement/Aim

No study has focused on re-evaluating the clonal-complex's robustness. This study revisits clonal-complex classification for *C. jejuni* by identifying whether the combinations of alleles, once known to be the central genotype for each clonal-complex, are still the majority for each group. The aim is to determine which clonal-complexes should remain the same, which need to be changed and which should retire due to low quantity.

## Impact Statement

This study analyzed the current clonal-complexes and their validity have not been ever done. The result indicated that 47.3% of the isolates (16 out of the 34 clonal-complexes) have central genotypes aligned to the original clonal-complex alleles. However, for the other 18, some have changed their central genotype allele, and some clonal-complexes are problematic for classifying isolates into one group. Furthermore, some have so few numbers of isolates that it would benefit to be kept as sequence type and not as clonal-complex.

## Introduction

Clonal-complexes and Multi-locus sequence typing (MLST) schemes have been used robustly in defining bacterial sub-variants (Maiden, 2006). Since bacteria constantly evolve, understanding bacterial sub-variants and identifying variant-specific characteristics is vital. One bacterium, *C. jejuni,* has extensively used the MLST scheme and clonal-complexes since its application to *C. jejuni* in 2002 (Dingle *et al.*, 2002). Initially, to establish central genotypes, the study used 194 isolates and combined subject matter expert views, with assistance from the BURST algorithm, UPGMA clusters and split decompositions implemented through SPLITSTREE (Dingle *et al.*, 2002; Feil *et al.*, 2004).

*C. jejuni* is a gram-negative bacterium that can cause gastroenteritis if ingested. Although many cases are sporadic, it can also cause outbreaks if the same source leads to food poisoning; for these cases, we can see that identical variants were the cause of the problem (Kang, Bang and Cho, 2019; Lai *et al.*, 2022). Poultry is the number one cause of Campylobacteriosis, but *C.jejuni* also inhabits different animal hosts from cows, pigs, geese, unpasteurized milk and is present in sewage water as well (Sheppard *et al.*, 2009; Cody *et al.*, 2015). Using clonal-complex, we have understood source attribution and global epidemiology (Maiden, 2006). Clonal-complex 21 is heavily circulated globally and inhabit different hosts commonly known as host generalist (Sheppard *et al.*, 2014; Dearlove *et al.*, 2016; Woodcock *et al.*, 2017). Clonal-complex 353, 354 and 464 are known to be poultry specialists(Alison J Cody *et al.*, 2012; Kittl *et al.*, 2013a; Ramonaite *et al.*, 2017a). Clonal-complex 42 and 61 have been known to be of cattle origin (Epping *et al.*, 2021). Further studies have tried distinguishing these clonal-complexes using accessory genomes and softcore genes (Thépault *et al.*, 2017a; Mourkas *et al.*, 2019). Recent papers focus on combining clonal-complexes with metadata to apply theoretical and machine-learning models to identify source attribution (Rawson, Dawkins and Bonsall, 2019; Arning *et al.*, 2021). However, there have been no studies that re-evaluate clonal-complex classification itself and its robustness. Thus, in this paper, we re-examine whether we have the correct central genotypes. For clonal-complexes that do not have the correct central genotypes, we have suggested which ones should change, which should be kept the same and which should retire due to their lack of isolate counts.

## Materials and Methods

### Meta Data Acquisition

Data is publicly available on the PubMLST database (Jolley, Bray and Maiden, 2018). The study uses *Campylobacter jejuni* isolates, which have been carefully chosen for having high-quality cgMLST data through the PubMLST database at Project ID:121 global_campyjejuni_1997=(n=17,338): https://pubmlst.org/bigsdb?db=pubmlst_campylobacter_isolates&page=query&project_list=121&submit=1. Isolates that only contained clonal-complex information (16,273) were used for the study (Accessed on 7th April 2023).

### Software usage

The coding language used for this study is Python (Van Rossum, G. , Drake, 2009). Code for the methods described is available here: https://github.com/bgrdessislava/Campyjejuni_gyrA. The minimum spanning tree was constructed using Grapetree software, included in the PubMLST tools to run the isolates (Zhou *et al.*, 2018).

### Category Assignment:

Clonal-complexes were classified into four categories: aligned, not-aligned, caution and retirement. Alignment were clonal-complex groups where the highest allele per 7 loci matched the original central genotypes allele (illustrated as asterisk sign *). Therefore, some of the clonal-complexes which do not have a majority (over 50%) from the central genotypes are also classified into this category. Not aligned are clonal-complexes where the highest proportion of alleles per loci are different from the central genotype alleles.

Thus, even if one of the seven loci is different, this was classified into a not-aligned category. Caution categories are clonal-complexes with the highest proportion aligning with the central genotype. The caution category is specifically critical since they are isolates found in high quantity, but through our study on NeighbourGroups, more insights indicate that there are multiple variants within. Retirement categories are isolates found in very low amounts where it is difficult to say if the central genotypes are still prevalent confidently. Furthermore, keeping it as a strain type could be better than grouping them into clonal-complexes.

## Results

The PubMLST query identified 17,338 *C. jejuni* isolates from any source between 1997 and 2022. Isolates without clonal-complex assignments were removed, leaving 16,273 isolates for analysis. First, we looked at each clonal-complex and observed their highest proportional alleles. Next, we evaluated all the ST types assigned to this clonal-complex (ex., for clonal-complex 21, we have ST21, ST50, ST53, ST19 and other STs) and constructed a pie chart per 7 MLST loci and their alleles. The relations of whether the clonal-complex is kept with the identical alleles were investigated (Figure 39). We can see that some clonal-complexes are congruent to their central genotypes (e.g., CC22, CC48 and CC257).

On the other hand, we have other clonal-complexes, such as CC206, CC464 and CC1034, that have a different majority of alleles to the central genotype alleles. Through the composition of the alleles, clonal-complexes have been divided into four categories: Not aligned, caution, aligned, and retirement (Table 8). If most MLST alleles match central genotype (CG) alleles, we categorize them into “aligned”. If the 7 MLST differs from CG, this is allocated to be in the “Not Aligned” category. We have also created a “retirement” category for clonal-complexes with less than ten isolates. Finally, we also have a “caution” category, a CC with unresolved issues, further discussed later.

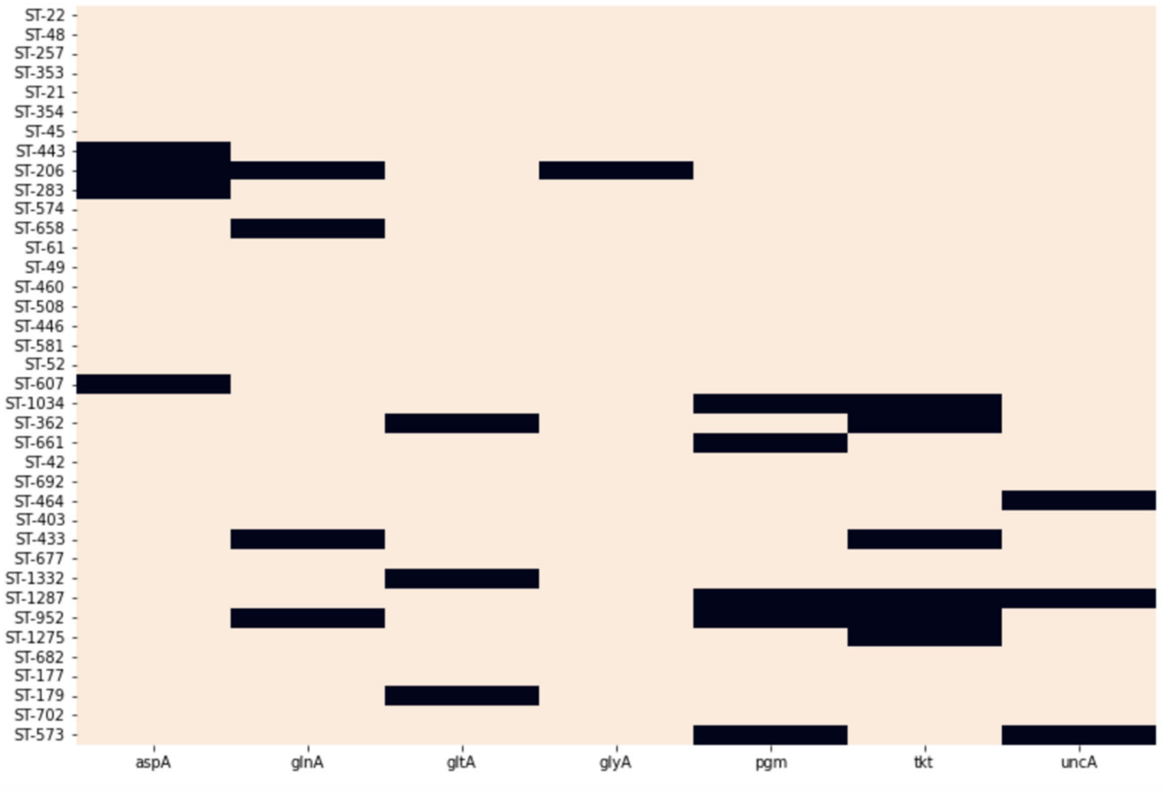


*Figure 39 Clonal-complexes with their 7 MLST loci and whether the majority central genotypes not aligned are shown in black. For 16 clonal-complexes out of 38 (42.1%), at least one of the seven housekeeping genes allele majority has changed. This indicates that these clonal-complexes need to be examined in more detail. However, for the other 19 clonal-complexes, the majority of each of the clonal-complexes central genotypes alleles has kept itself.*

We have 3089 isolates not aligned to the central genotypes, leading to 19% of the total isolates not having the central genotypes as their core group. On the other hand, we have 7700 isolates, which are 47.3% of the isolates with their central genotypes aligning to 16 clonal-complexes. We have a big group under caution, including CC21, CC353 and CC581. Although these clonal-complexes have aligned central genotypes, there is a big divide within their alleles. This group needs further investigation, especially since the caution category is 33.4% of all the isolates. The combined total of Not aligned and caution groups reached 52.4% of the total isolates, an astonishing number that cannot be ignored. The retirement category is 0.3% of the isolates, leading to a suggestion to keep them as ST types rather than keeping them as clonal-complex. Of course, this could be due to the sampling bias of the data available in PubMLST and needs to be further investigated with other databases in the future.

*Table 8 Clonal-complex to Central Genotypes assignment categories*

| MLST matching the central genotype alignment | Status | Clonal-complex | Isolate Count [% of Sampled Isolates] |
|---|---|---|---|
| Not Aligned to Central Genotypes | Not Aligned | 206, 283, 362, 443, 464, 607, 658, 661, 1034, 1287, 1332 | 3089 [19 % - 11 CC] |
| Aligned but needs further investigation | Caution | 21, 353, 581 | 5442 [33.4% -3 CCs] |
| Aligned | Aligned | 22, 42, 45, 48,49, 52, 61, 257, 354, 403, 446,460,508,574,677, 692 | 7700 [47.3% - 16 CCs] |
| Mix but Count <10 | Retirement | 177, 179, 433, 573, 682, 702, 952, 1275 | 43 [0.3% - 8 CCs] |
| | | **Total** | **16273 – 38 CCs** |

**Central Genotypes Investigated**

Here, we have demonstrated the allele composition of each clonal-complex by adding an asterisk (*) to their original allele used for central genotypes. Some clonal-complexes do not have most alleles aligned to the original clonal-complexes (CC206 have 3 out of 7 MLST loci differing from the original alleles).

### 1.) Clonal-complex that needs re-adjustment [Count >20] 206, 283, 353, 362, 443, 464, 607, 658, 661, 1034, 1287, 1332

In CC206, we can see that only 13.6% have the original ST206 alleles. For some clonal-complexes, the four and 5s dominate more than 6 or 7 matches (CC206, CC362, CC607,CC1034,CC1287 and CC1332). Furthermore, some clonal-complexes do not have any isolates with the core alleles (CC1287), where the original alleles are not found for two loci (pgm and *tkt*). This illustrates that although an ST type was initially picked as the central genotype, the variant is no longer seen.

For CC206's loci, *gltA*, *pgm*, *tkt,* and *uncA* have high levels of their alleles aligned to the central genotypes (Figure 40). However, for the other three central genotypes alleles, their composition is not the highest: *aspA* allele-*2*(36.9%), *glnA* allele-*21*(27.7%), and *glyA* allele-*37*(20.1%). 64.1% of the total CC206 isolates have the same four identical alleles to MLST alleles. On the other hand, only 13.6% are from ST206 isolates. On the other hand, for CC464, 6 out of 7 loci have matched the central genotypes (Figure 40). However, for uncA, only 22.5% of the isolates have the central allele of 1, and the rest seem to have allele 3. Although CC464 have more than six out of 7, the same as the central genotype, having only 17% of its isolates from the central genotype is relatively

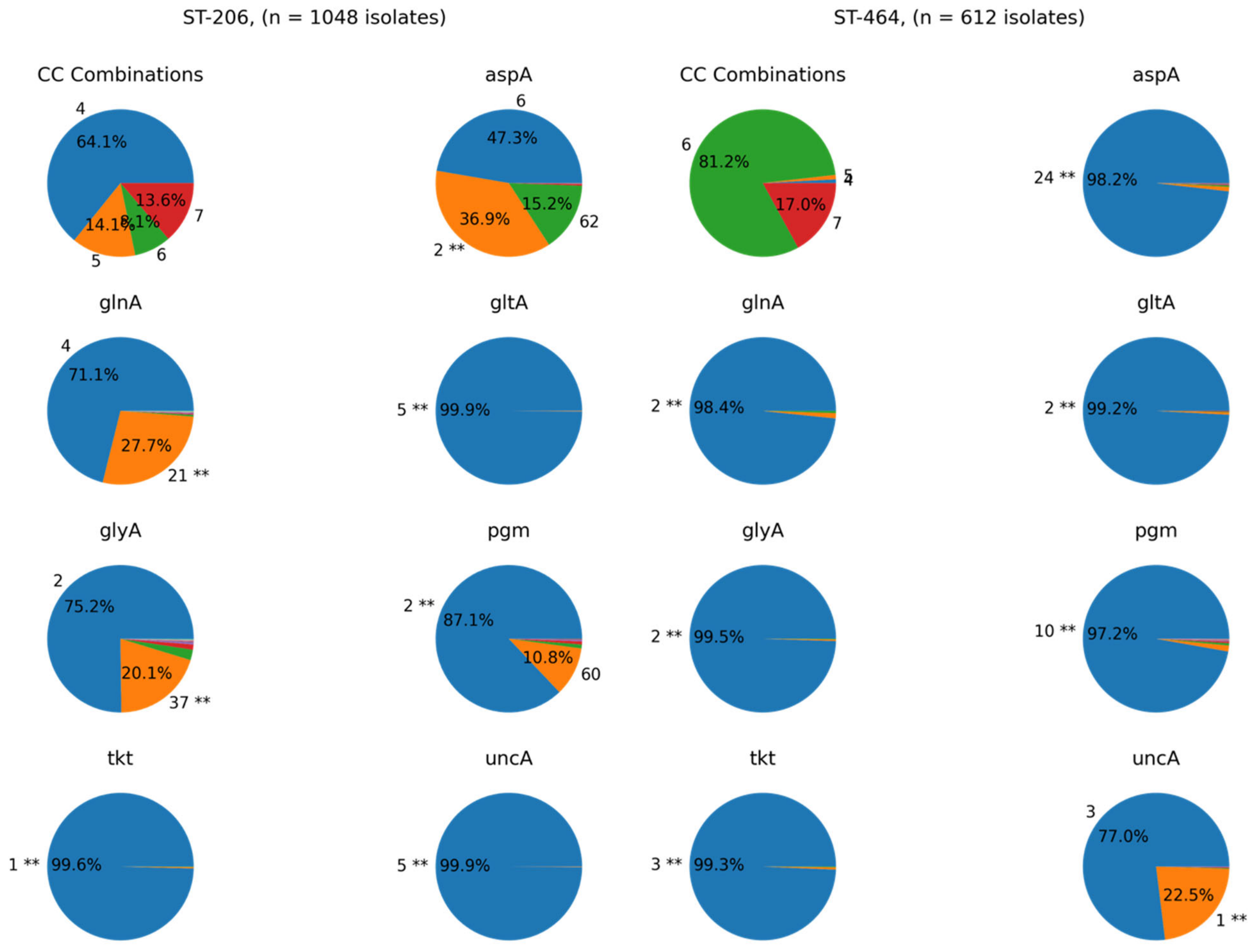


*Figure 40 CC206 and CC464 pie charts showing 7MLST allele compositions. The * indicates the central genotype's alleles. Top left pie chart indicates the overall compositions of the isolates (i.e 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third and red is fourth.)*

low.

Next, we will investigate clonal-complexes found with less quantity but still show signs that adjustments are needed (Figure 41). CC283 is a peculiar case where most alleles match with more than 80%. However, the *aspA* allele is not found even once in our samples. Moreover, for CC283, we do not have any isolates that are from ST283, and we only have 6 or 5 of the same in the central genotypes. CC1034, on the other hand, has a couple of different alleles per 7 MLSTs. We have two loci, *pgm* and *tkt,* that are not aligned with the central genotypes and diverse alleles for both loci. Within *pgm,* there are 12 different alleles grouped into the same clonal-complexes. For *gltA* and *uncA,* the allele matches 100% and 97% each to the central genotypes.

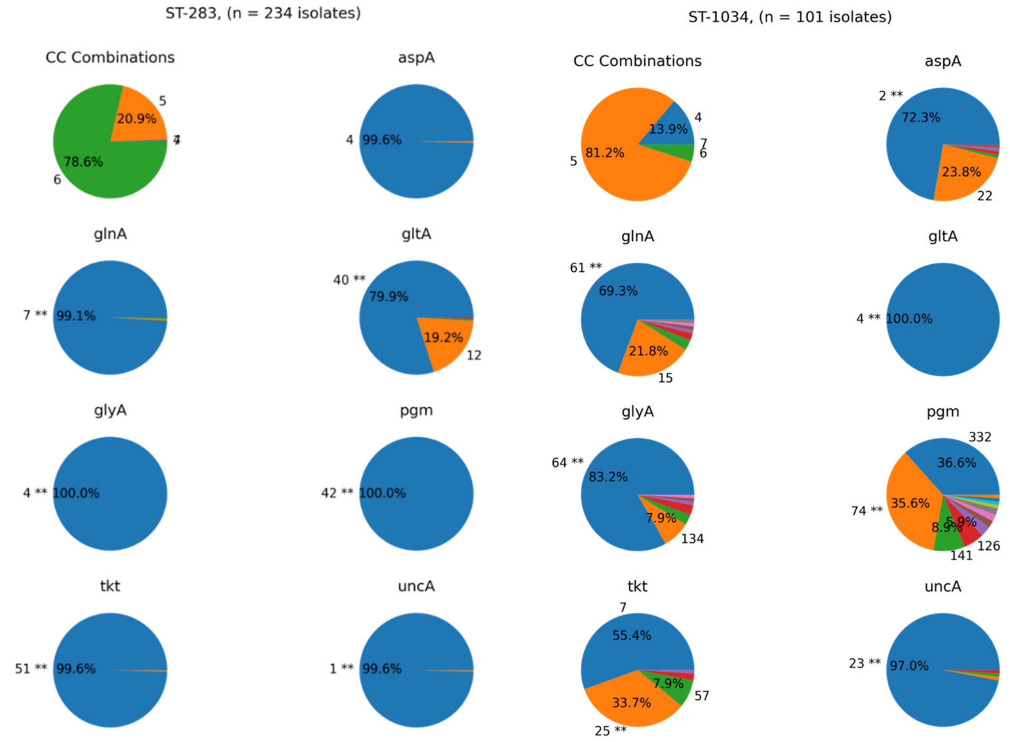


*Figure 41 CC283 and CC1034's 7MLST allele compositions and its MLST allele combinations. . The * indicates the central genotype's alleles. Top left pie chart indicates the overall compositions of the isolates (i.e 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third and red is fourth.)*

In previous work, we established a new classification, "NeighbourGroups", which uses the gradient boosting tree model "Catboost" that has trained to identify cgMLST(1343 loci) level classification from 7 MLST. For *C. jejuni,* we have used 10,359 isolates from the UK from 1998 to 2018 to train this classification. (Veltcheva *et al.*, 2023). The alignment between clonal-complex and NeighbourGroups is shown below (Figure 42). Some clonal-complexes such as CC206, CC464, and CC607 have some small amounts divided into different NeighbourGroups. However, most clonal-complex matches one NeighbourGroup (CC206 to NGroup 9). This illustrates that although the central genotype is not aligned, we see that, as a group, the central genotypes were, from the beginning, wrongly assigned, or it shows that the central genotype as a whole group has changed over time. We cannot speculate from here which is the answer, but we can say that, as a group, they do have cohesion.

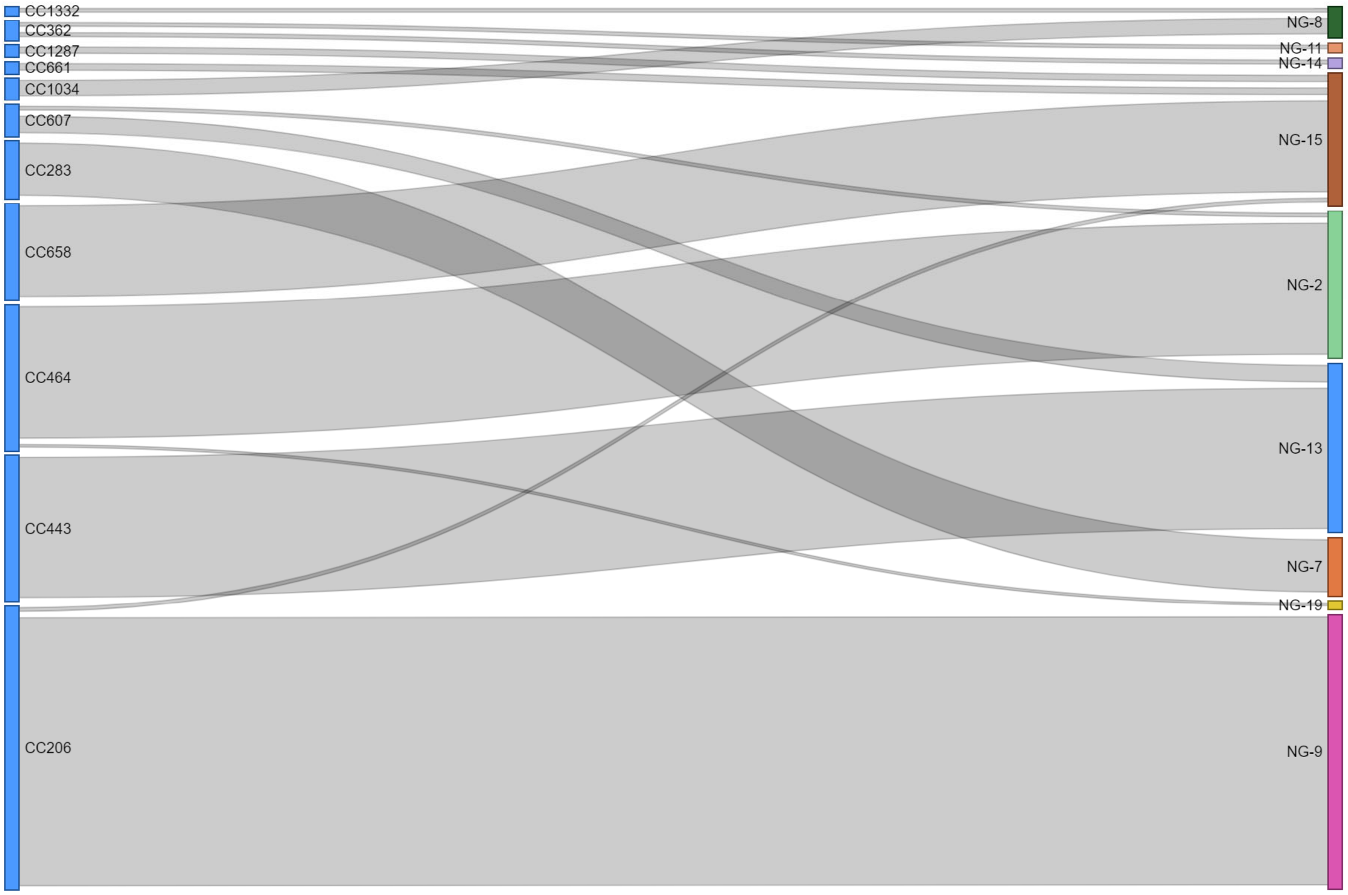


*Figure 42 Clonal-complexes to NeighbourGroups connections for clonal-complexes that are in the "Not Aligned Category", meaning they do not have the majority of the central genotypes aligned and have other alleles being dominant in at least one of seven MLST loci*

### 2.) The clonal-complex stays the same but needs close attention.

Looking at the central genotype's majority, these three clonal-complexes, CC21, CC353 and CC581, do not show signs of potential conflict. However, through the proximity of the top allele for *gltA* for CC21 and previous work on CC353, which has also established CC581, we know these two clonal-complexes also need further attention.

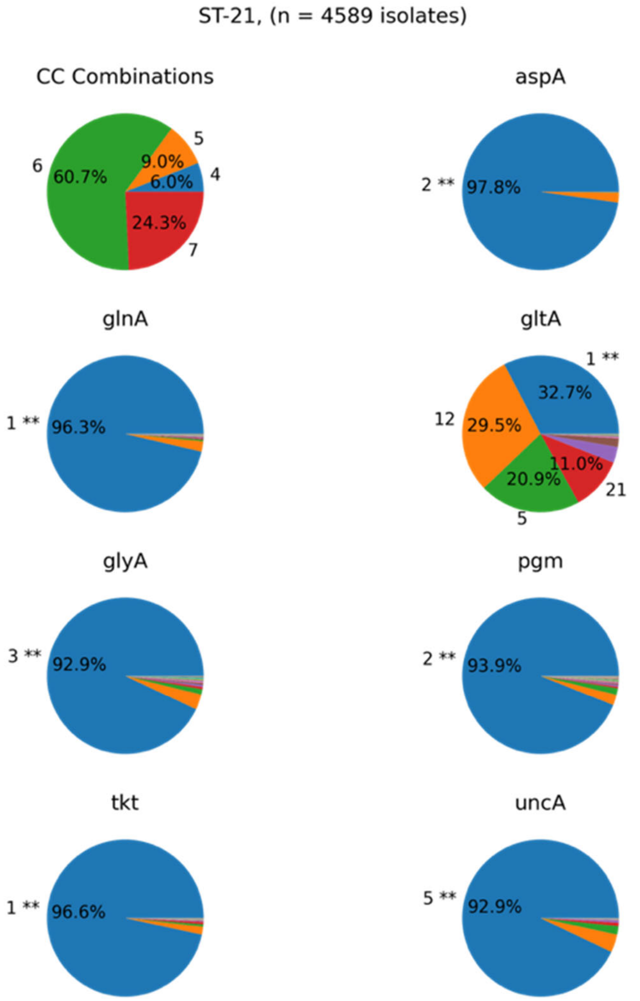


*Figure 43 Clonal-complex 21 Allele compositions strongly associate with central genotypes except for gltA, which has a couple of high-level alleles. . The * indicates the central genotype's alleles. The top left pie chart indicates the overall compositions of the isolates (i.e. 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third, and red is fourth.)*

Regarding CC21, there are six identical alleles (60.7%), and ST21 isolates are 24.3% (Figure 43). In terms of each of the alleles, almost all of them match the central genotype alleles except for *gltA.* We can see that there are multiple alleles for this locus: 11 (32.7%),12(29.5%), 5(20.9%) and 21(11%). To investigate this further, we have looked at what ST types are produced based on these changes, and the STs changes correspondingly lead to ST50, ST19 and ST5. There are two major clusters with ST-50 and ST-19 and ST-21 and ST53.

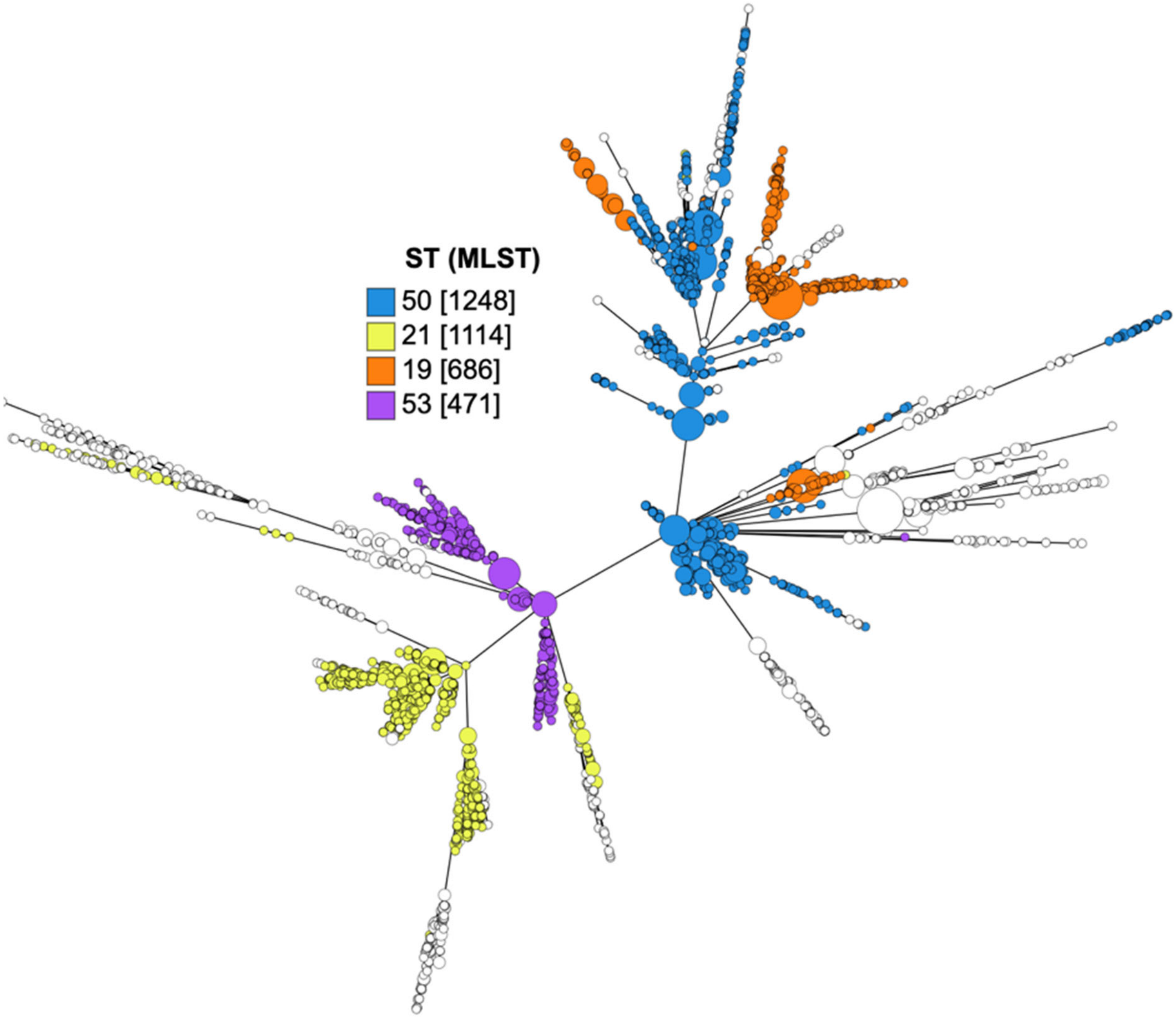


*Figure 44 Minimum spamming tree built from cgMLST showcasing CC21 and its Top 4 MLST types showcasing a big two divide with the top two, ST50 and ST19, in the similar cluster and top 3 and 4, ST53 and ST21, in the second cluster. Other isolates have not been annotated, and it is shown in white.*

Next, CC353 has been further investigated for re-classification in previous work (CC353 paper citation needed). This work shows that CC353 comprises five sub-groups, one displaying another behaviour, the ST581 (Chapter 5). Through this work, ST581 was added to form a new clonal-complex. However, none of the historic isolates have been re-run to see if they will be classified onto CC581. For CC353, the *glnA,* the second highest allele, is 2. If we change the combination of an allele for *glnA* to 2, we get MLST 5, the highest MLST type within CC353. However, if we change the central genotype to MLST 5, the problem of CC353 found across multiple branches continued.

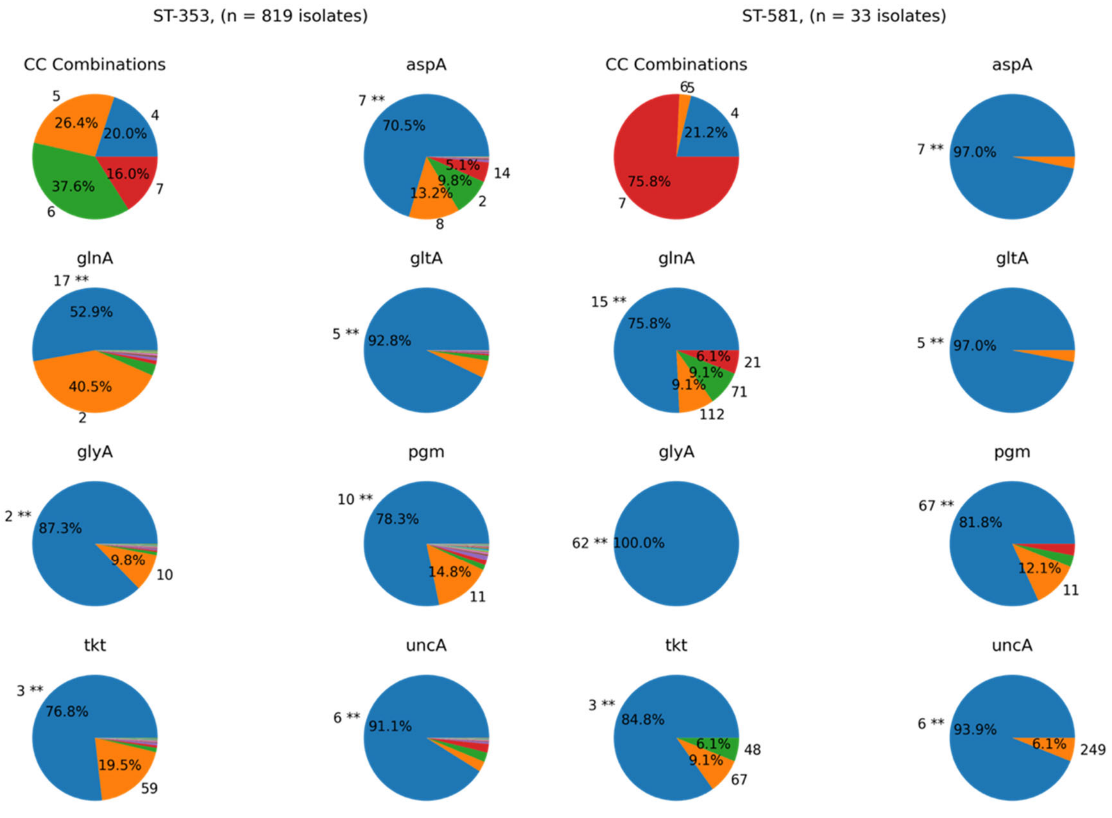


*Figure 45 CC353 and CC581 are the two clonal-complexes in the caution group. Although these two groups look to keep their majority, they have some issues beyond the 7 MLST loci. The * indicates the central genotype's alleles. The top left pie chart indicates the overall compositions of the isolates (i.e. 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third, and red is fourth.)*

The alignment of the Clonal-complex to NeighbourGroups associations for the caution category is illustrated below (Figure 46). The result indicates that we have multiple NeighbourGroups associated with these three groups. There is also a clear signal that CC21 comprises at least two main clusters, which we observed when we changed the ST types (Figure 44). Thus, as a clonal-complex, these three should not be classified into one group, and we need to be cautious about any findings we associate with them.

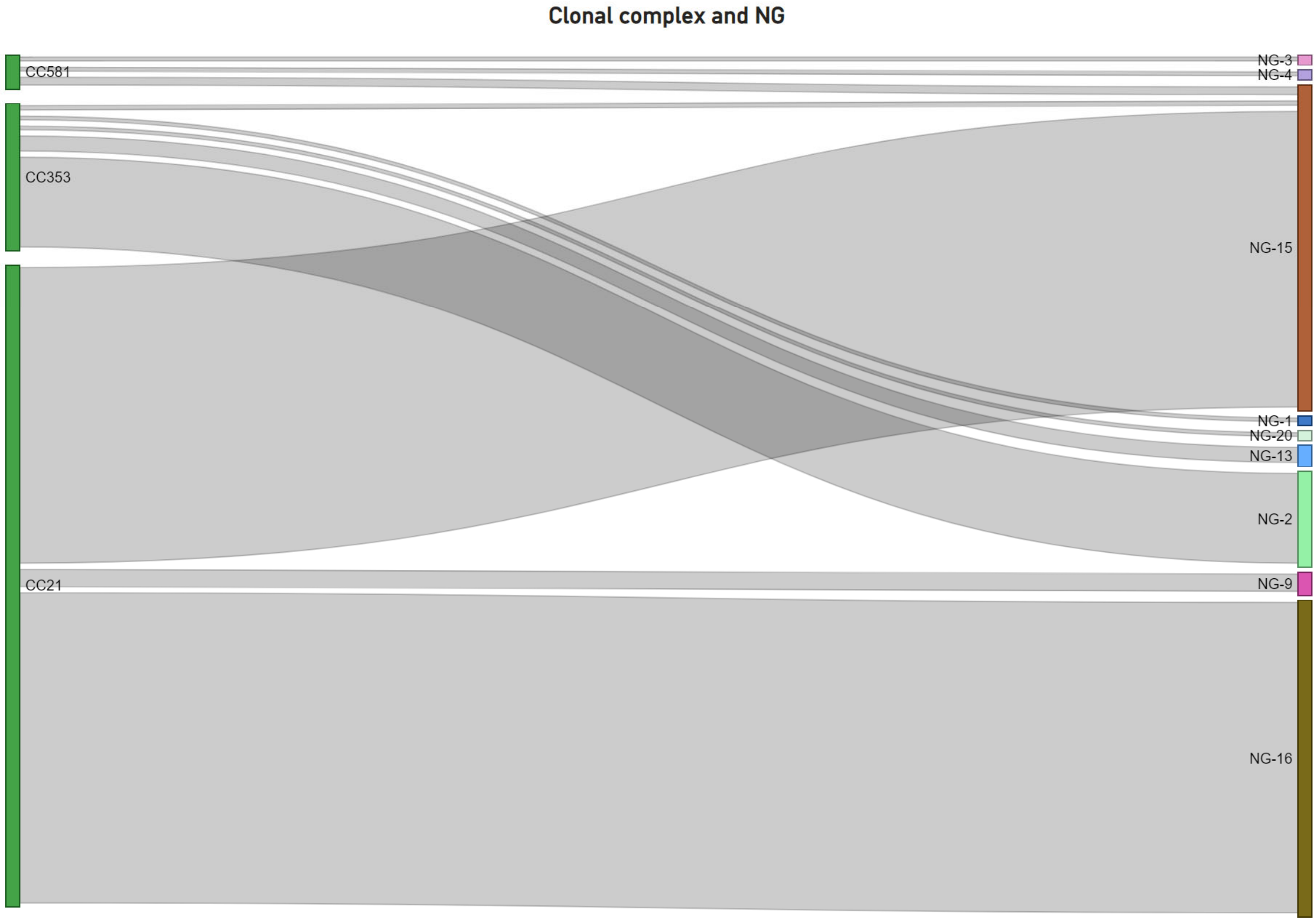


*Figure 46 Clonal-complexes to NeighbourGroups connections for clonal-complexes in the "caution" category are aligned to central genotypes but have some issues beyond the 7 MLSTs.*

### 3.) The clonal-complex is still going strong, and no re-adjustment is needed (=> 20 isolates per clonal-complex)

Here we list clonal-complexes [22, 42, 45, 48,49, 52, 61, 257, 354, 403, 446,460,508,574,677, 692] that have their 7 MLST alleles aligned with their central genotypes. Many clonal-complexes dominate all seven alleles, aligning with the dominant central genotypes. As a group, these clonal-complexes still have their central genotypes captured 21 years ago. Some of the clonal-complexes and the rest can be found in the supplementary figures (Figure 47). For CC42, all 7 MLSTs are aligned to the central genotype ST42 alleles, with all having above 86% of the majority. In CC45, although *pgm* has another allele, 42, assigned to 32.2% of the isolates within this group, allele 1 for *pgm* is the majority with 65.4%, and all their alleles are aligned.

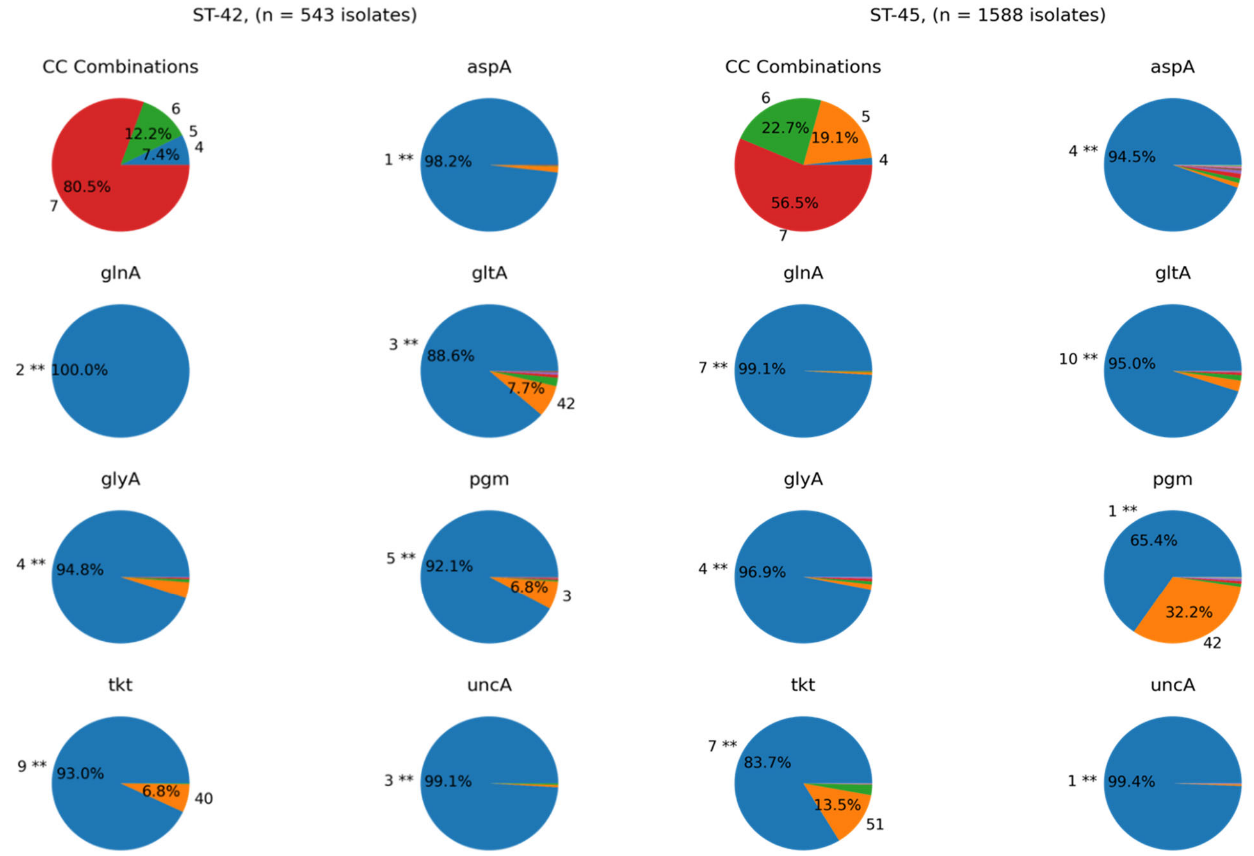


*Figure 47 CC42 and CC45 are clonal-complexes with aligned central genotypes for all 7 MLST alleles. The * indicates the central genotype's alleles. The top left pie chart indicates the overall compositions of the isolates (i.e. 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third, and red is fourth.)*

Regarding the clonal-complex to NeighbourGroups alignment, other than a small number of isolates in CC48, CC61 and CC257, the other CCs are all aligned to one NeighbourGroups. By shrinking the size of the groups from 35 clonal-complex to 20 NeighbourGroups, we can see a couple of clonal-complexes being assigned to the same NeighbourGroups such as CC22 and CC42 becoming NeighbourGroup 14 and CC52 and CC354 and a little bit out of some other clonal-complex into one NeighbourGroup.

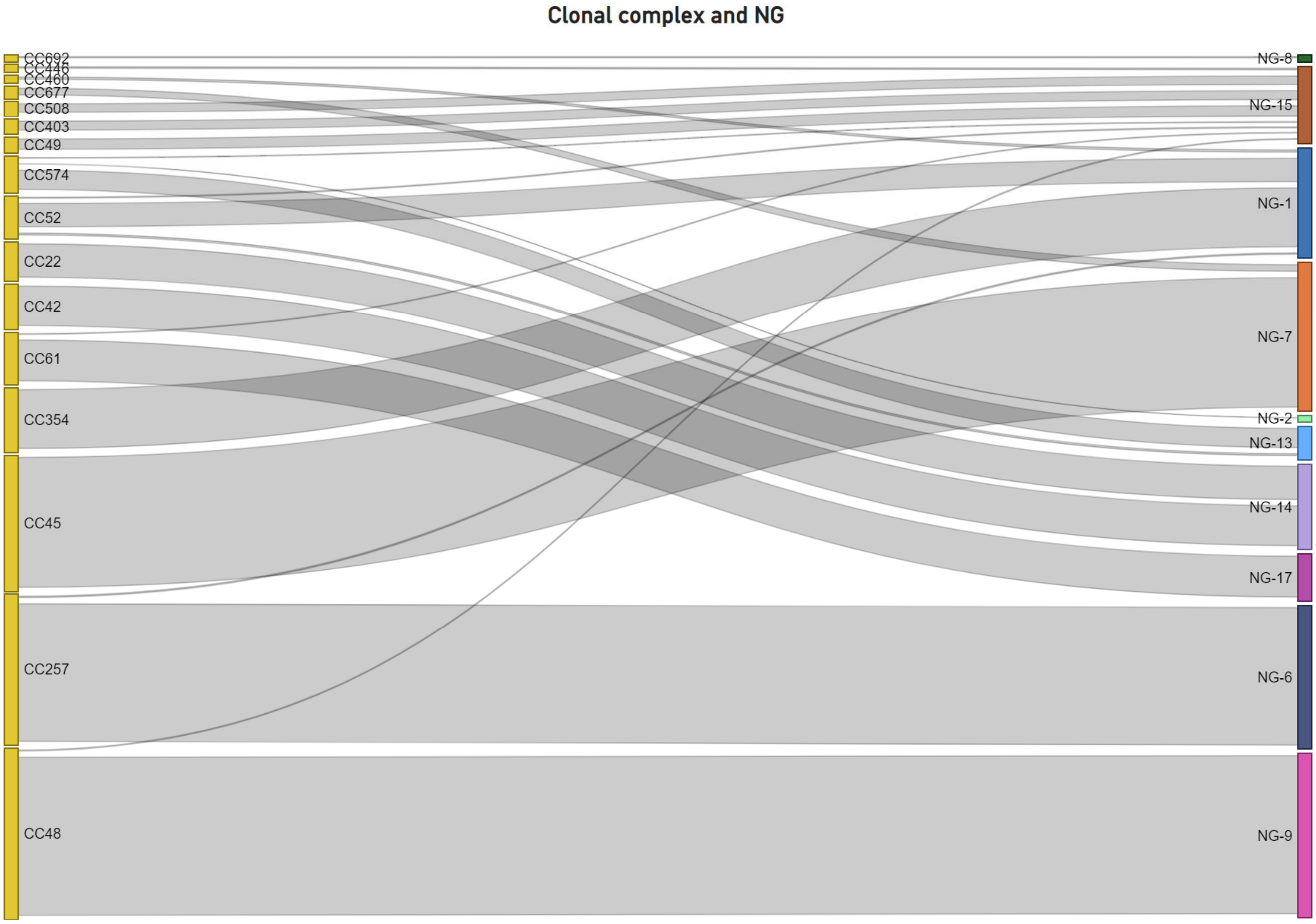


*Figure 48 Clonal-complexes to NeighbourGroups connections for clonal-complexes that are in "Aligned", meaning that the central genotypes of the 7 MLSTs are still aligned to the current groups.*

### 4.) Clonal-complex for retirement [very few numbers (<10)]

Clonal-complexes [177, 179, 433, 573, 682, 702, 952, 1275, 1287] are in low count (less than ten isolates per clonal-complexes), raises the issue of whether they should stay as ST-types and retire as clonal-complexes (Figure 49). For CC433, we only have seven isolates *glnA* is composed of multiple alleles, and the central genotype allele is not found. For CC179, a similar issue is found in *gltA,* where we can see that none of the alleles from central genotypes are found in our sample. For locus *glyA*, the top two alleles are the same, and we cannot see any ST179 isolates in our sample. These two clonal-complexes demonstrate that staying as strain types rather than clonal-complexes would be more beneficial when the isolate counts are very low. Thus, we suggest that these nine clonal-complexes, only 43 isolates from the total samples (0.3% of the total isolates), could be retired as clonal-complexes.

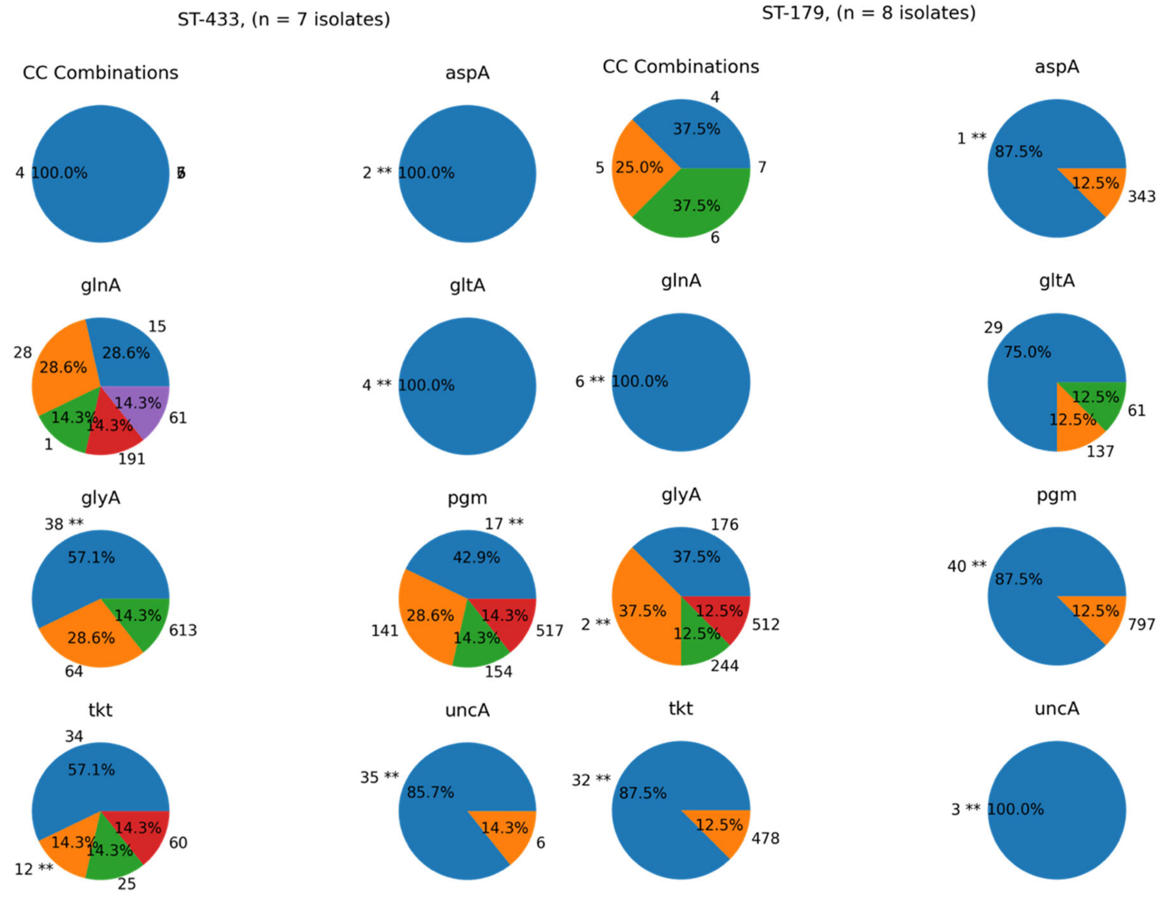


*Figure 49 CC433 and CC179 are two examples of clonal-complexes in the "retirement" group. These clonal-complexes are less than ten in isolates, suggesting that it would benefit MLST types more than having a clonal-complex. The * indicates the central genotype's alleles. The top left pie chart indicates the overall compositions of the isolates. i.e. 4 of the same alleles to central genotype up to 7 of the same indicating the central genotype). Blue indicates the highest proportion of the alleles, orange is second, green is third, and red is fourth.*

# Discussion

## Clonal-complex usages:

This paper uses clonal-complex and robustness to identify whether central genotypes still hold the majority per group. Results demonstrated that 16 clonal-complexes still have their central genotypes, the majority for all 7 MLST loci. On the other hand, 11 clonal-complexes of the central genotypes were not aligned. However, in almost all of them, it looks as when shown in NeighbourGroups, they keep their core group, indicating that it still has similar characteristics as a group. Nevertheless, their central genotype alleles have either changed or were not the correct central genotype initially. Clonal-complexes CC21, CC353, and CC581 are problematic since they separate into multiple groups when assigned to NeighbourGroups. CC21 and CC353 are among the most found clonal-complexes worldwide (Kovač *et al.*, 2014; Ramonaite *et al.*, 2017b). Through this study and NeighbourGroups, we have shown that re-adjustments are needed for CC21 and CC353.

## CC21 and CC353 literature

Our work supports previous research that indicates that CC21 has a high level of sequence type diversity. (Grove-White *et al.*, 2011). Our results show that CC21 has a high level of deviation from the central genotype in *gltA*, and these findings may explain why this CG has been previously classified as a 'generalist' in previous work (Gripp *et al.*, 2011; Sheppard *et al.*, 2014; Dearlove *et al.*, 2016). Furthermore, another study has shown that the most sequence-type diversity was found in CC21, CC45 and CC206.

Through our analysis, CC21 did show high diversity in *gltA*, but CC45 did not have high diversity in 7 MLSTs from our study. Although CC206 is in the not aligned category, we also see a not-too-diverse nature where it mainly gets classified into NeighbourGroup 9 and some in NeighbourGroup 15. Another study found that CC21 comprises two groups of ST-21 and ST-50, which we have also seen in our research (Skarp *et al.*, 2017). The study verified that the difference lies in the capsule loci and that their accessory genetic content and plasmids are the reasons behind these differences (Skarp *et al.*, 2017). It has also been stated that ST-50 has further shown microevolution into two distinct clusters, which we have not found in our study, but we can see that ST-19 is a close group to ST-50 (Skarp *et al.*, 2017).

**Specific characteristics for clonal-complexes:**

Specific clonal-complexes can lead to particular phenotypes, and CC22 is a significant group leading to Guillain–Barré syndrome (GBS) (Nielsen *et al.*, 2010). However, another paper states that CC403 were the group associated with GBS in Bangladesh (Islam *et al.*, 2009). Our research demonstrated that CC22 and CC42 merge into the same NeighbourGroups, so identifying their similarities could be key.

In terms of antimicrobial resistance, a previous study has shown five different patterns in fluoroquinolone resistance (Veltcheva *et al.*, 2022). This work has verified that some clonal-complexes are firmly resistant across time, such as CC464, but some, such as CC42, CC45 and CC61, are predominantly susceptible. Other studies have also agreed with this finding, such as a paper that one of the CC464, an ST5136, was multi-drug resistant (Lopes *et al.*, 2019). Another study agreed that CC45 was predominantly susceptible and CC464 was resistant to fluoroquinolones (Kittl *et al.*, 2013a).

We can also start to see some source attribution to specific clonal-complexes, such as CC61 and CC42, closely related to cattle and sheep (Colles *et al.*, 2003; Kwan *et al.*, 2008). Regarding poultry origin, we can see clonal-complex 45, 353, 354 and 464 have been thought to be associated with poultry (Alison J. Cody *et al.*, 2012). Multiple methods have used source attribution studies, such as STRUCTURE software (Kittl *et al.*, 2013b). This program tries identifying potential host source populations using allele information at 15 host-segregating loci (Thépault *et al.*, 2017b). Another method that has attempted to use a machine learning approach to source attribution is comparing different modes, proofing that the random forest model can improve the source attribution using MLST loci (Arning *et al.*, 2021).

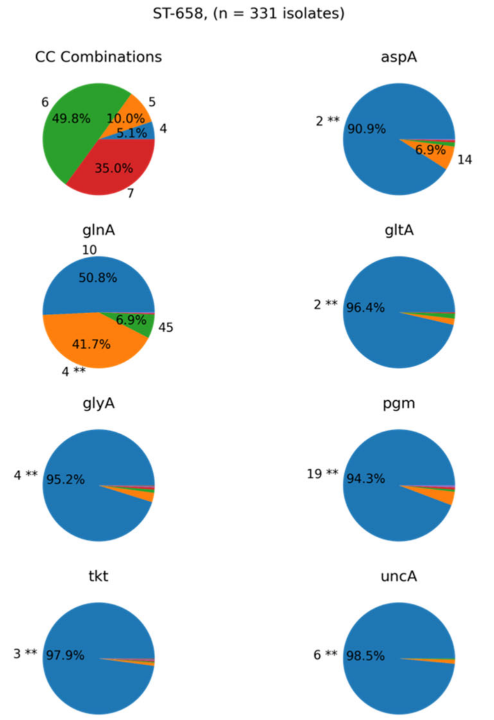
Furthermore, if we use cgMLST with XGBoost, the power of source attribution is even better (Arning *et al.*, 2021). However, if we are not currently using the most robust clonal-complex schemes by revisiting their central genotypes, any analysis used for further research will not be compatible. Thus, through our revisit to clonal-complexes, we have shed light on clonal-complex 21 and 353 that need to be re-classified into different groups. This could be the end of stating that CC21 is a generalist or that it is very complex.

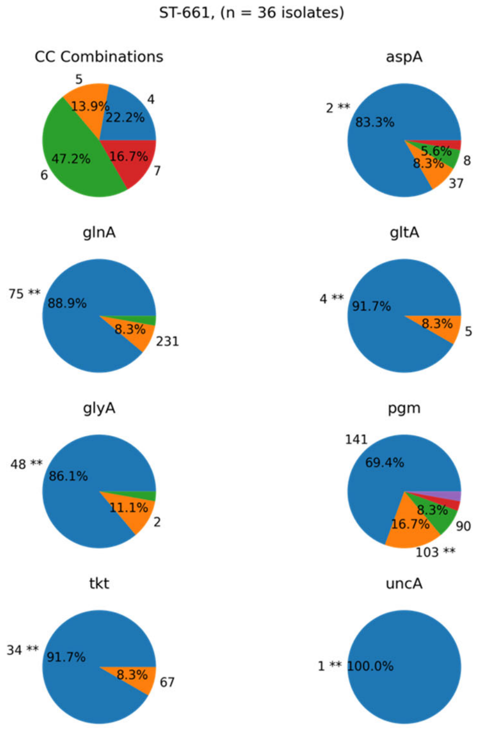
## Conclusion

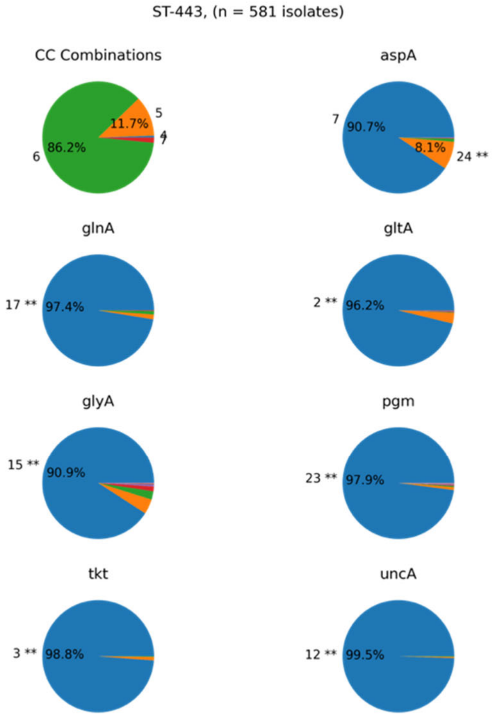
This study demonstrated four types of groups of clonal-complexes by re-evaluating the clonal-complex classification and its central genotypes. The first groups align perfectly with the central genotype and do not need to change. The second group still aligns with the central genotypes (CC21, CC353 and CC581) but raises concern for other reasons, and we strongly advise these three clonal-complexes not to be used in the future. The third group do not align with central genotypes, but the majority of them have still evolved to be relatively cohesive with some mismatched alleles. The fourth group are clonal-complexes with minimal data, which leads to whether these clonal-complexes should be retired and left as sequence types. Until now, no study has revisited the clonal-complex, and this work has reviewed the scheme and exhibited which groups are working well and which need to be changed through the application of NeighbourGroups.

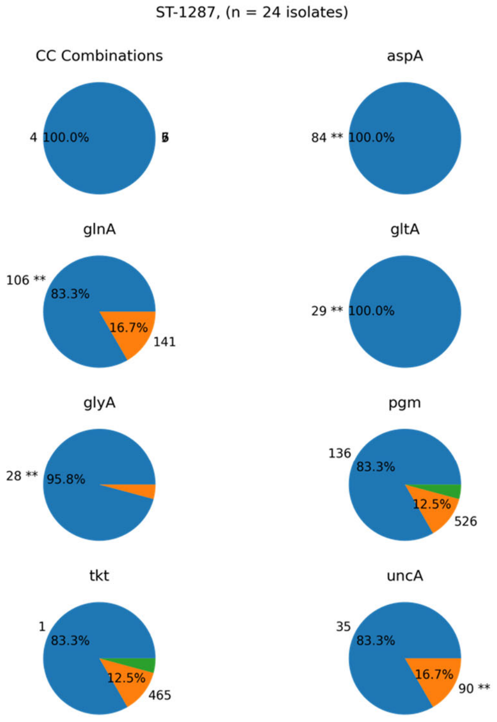
# Supplementary

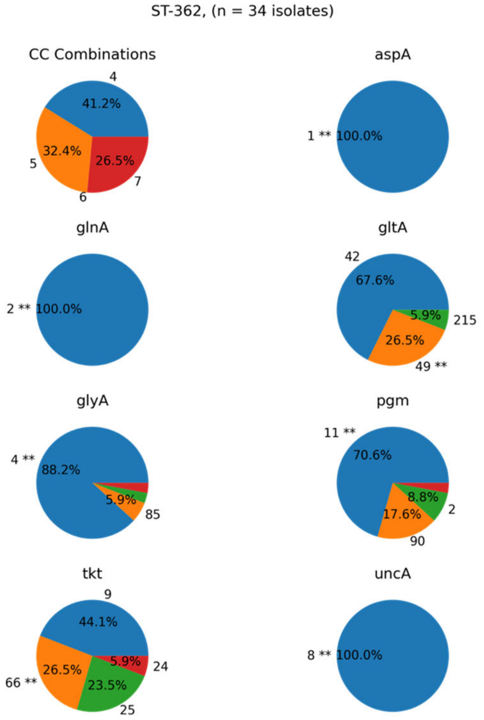
## Not Aligned

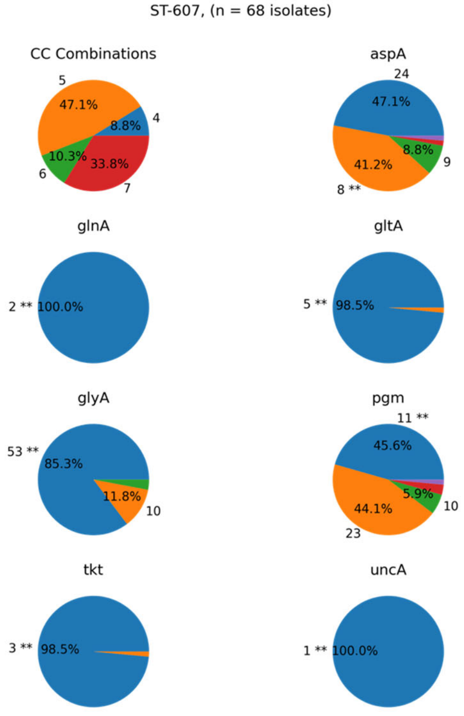

ST-658, (n = 331 isolates)

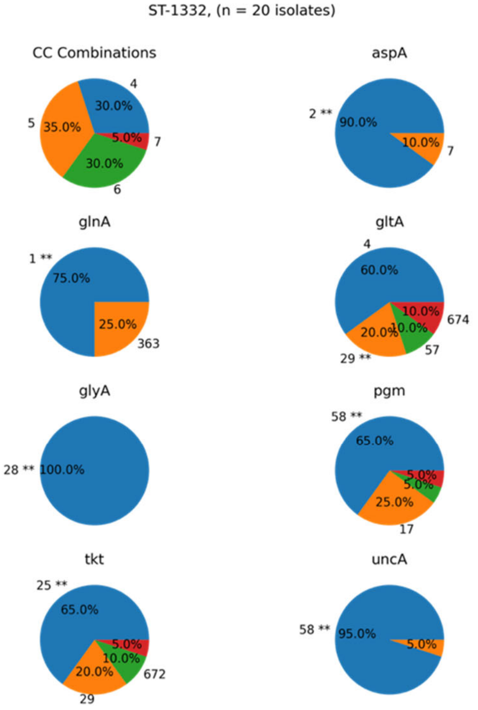

ST-661, (n = 36 isolates)

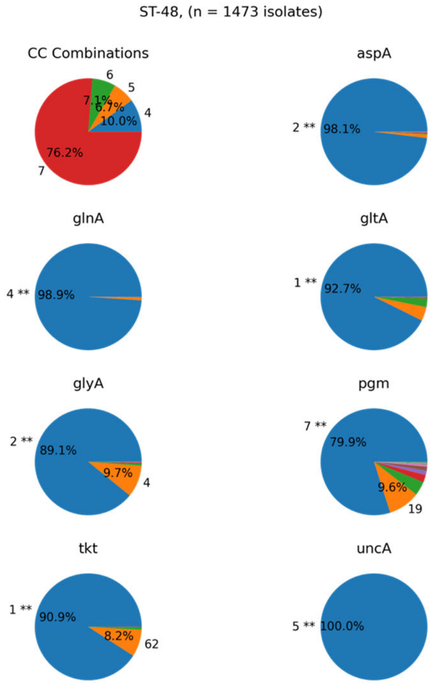

ST-443, (n = 581 isolates)

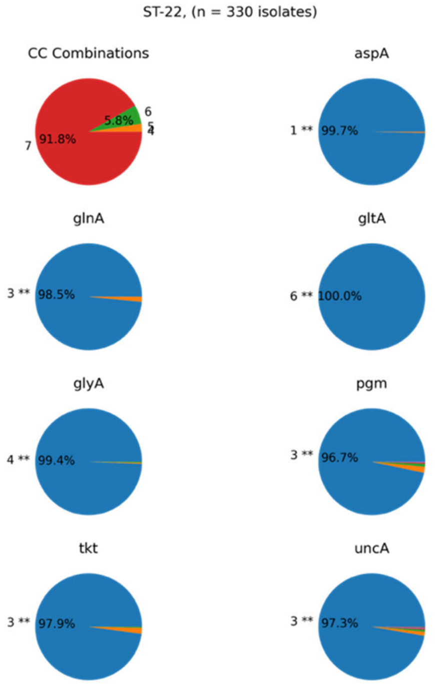

ST-1287, (n = 24 isolates)

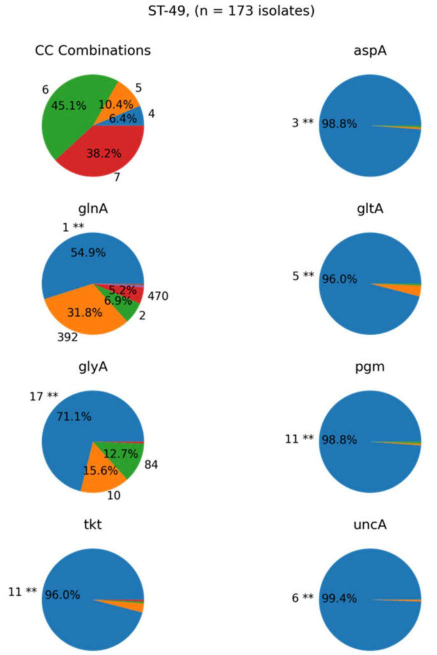

ST-362, (n = 34 isolates)

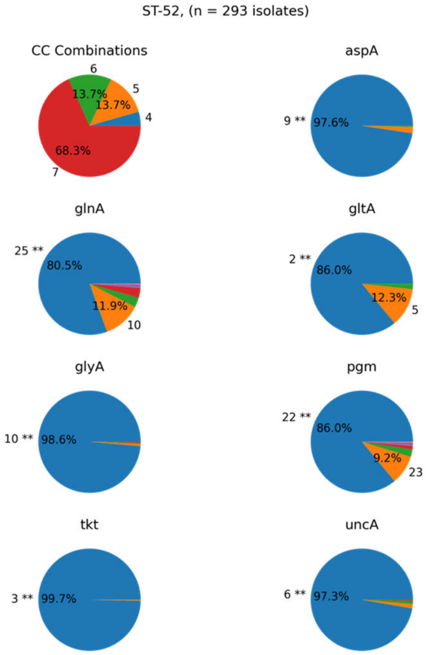

ST-607, (n = 68 isolates)


ST-1332, (n = 20 isolates)

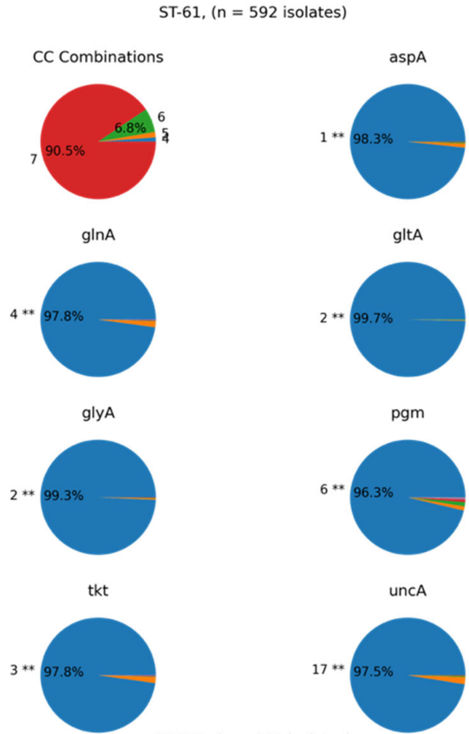

ST-48, (n = 1473 isolates)
CC Combinations
6
5
4
7.1%
6.7%
10.0%
76.2%
7
aspA
2 **
98.1%
glnA
4 **
98.9%
gltA
1 **
92.7%
glyA
2 **
89.1%
9.7%
4
pgm
7 **
79.9%
9.6%
19
tkt
1 **
90.9%
8.2%
62
uncA
5 **
100.0%


ST-22, (n = 330 isolates)
CC Combinations
6
5
4
5.8%
91.8%
7
aspA
1 **
99.7%
glnA
3 **
98.5%
gltA
6 **
100.0%
glyA
4 **
99.4%
pgm
3 **
96.7%
tkt
3 **
97.9%
uncA
3 **
97.3%

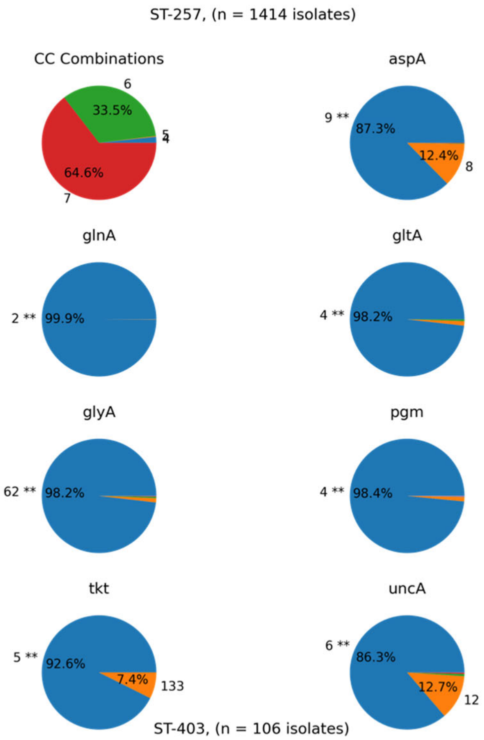

ST-49, (n = 173 isolates)
CC Combinations
6
5
4
45.1%
10.4%
6.4%
38.2%
7
aspA
3 **
98.8%
glnA
1 **
54.9%
5.2%
6.9%
31.8%
470
2
392
gltA
5 **
96.0%
glyA
17 **
71.1%
12.7%
15.6%
84
10
pgm
11 **
98.8%
tkt
11 **
96.0%
uncA
6 **
99.4%


ST-52, (n = 293 isolates)
CC Combinations
6
5
4
13.7%
13.7%
68.3%
7
aspA
9 **
97.6%
glnA
25 **
80.5%
11.9%
10
gltA
2 **
86.0%
12.3%
5
glyA
10 **
98.6%
pgm
22 **
86.0%
9.2%
23
tkt
3 **
99.7%
uncA
6 **
97.3%

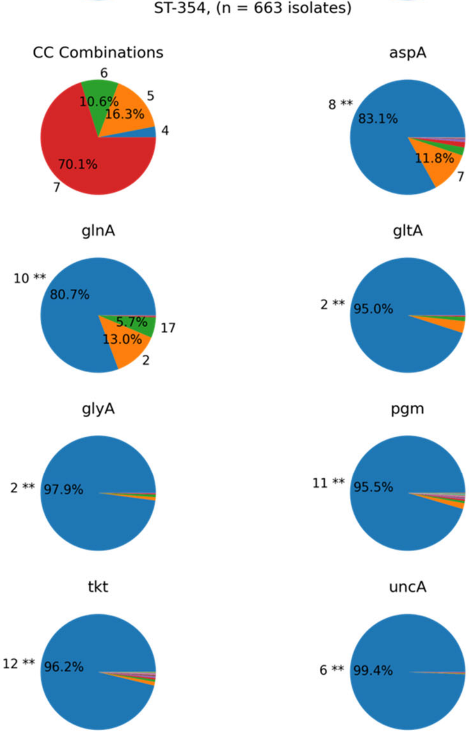
## CC Aligned



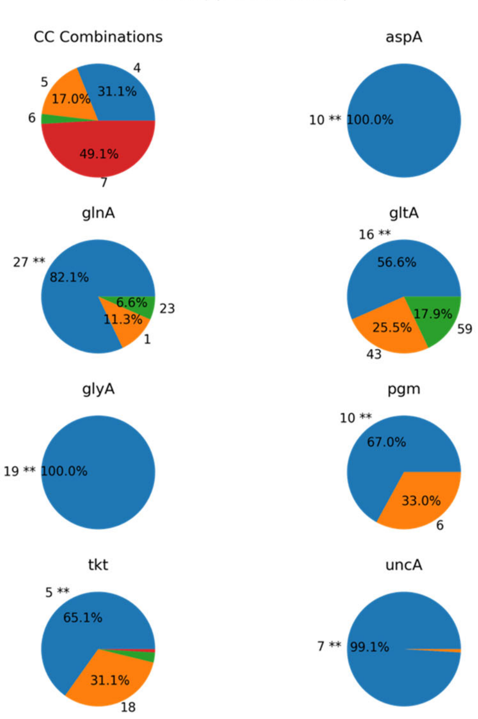
ST-61, (n = 592 isolates)



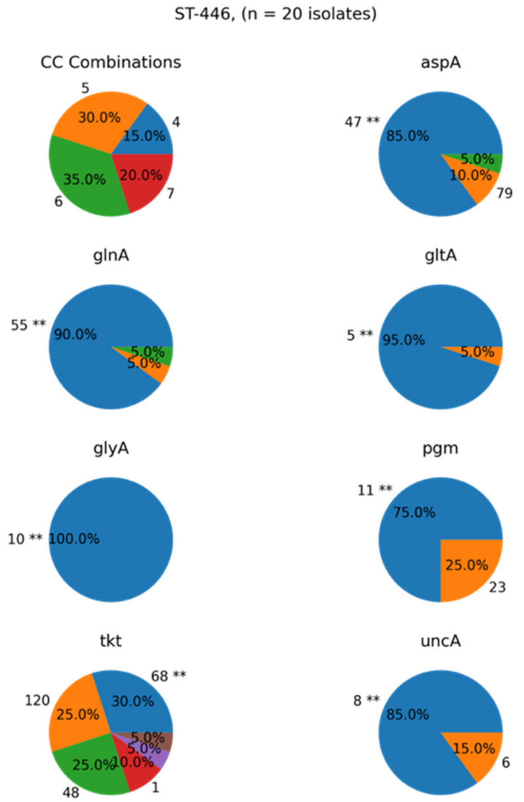
ST-257, (n = 1414 isolates)

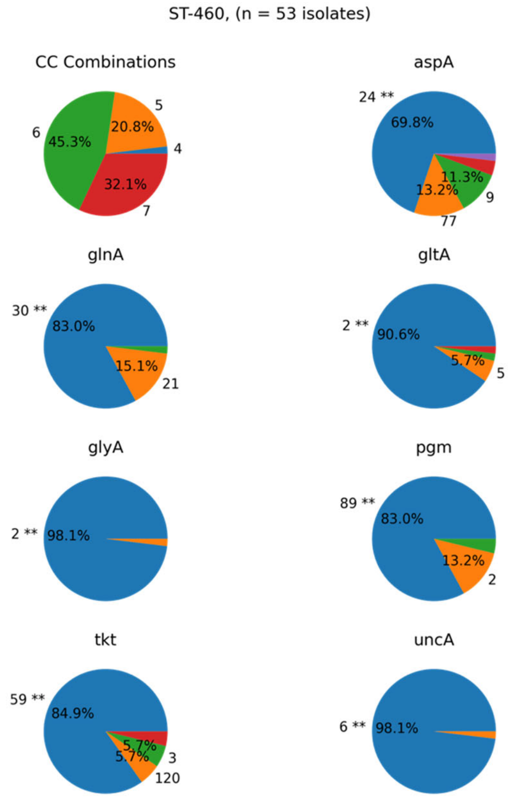


ST-354, (n = 663 isolates)

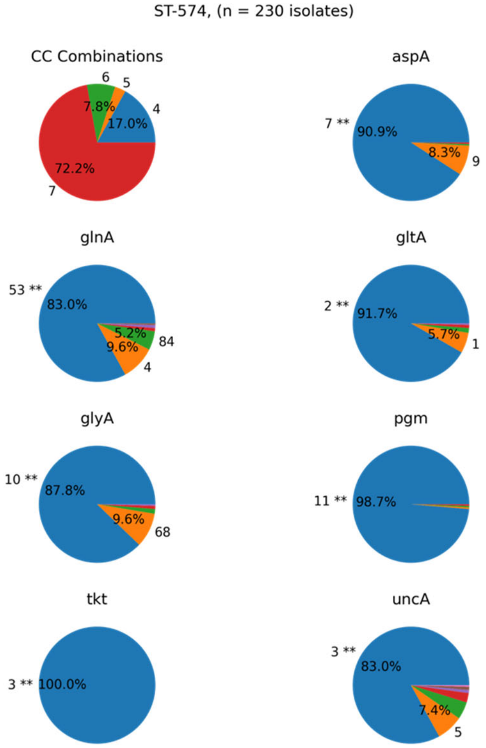


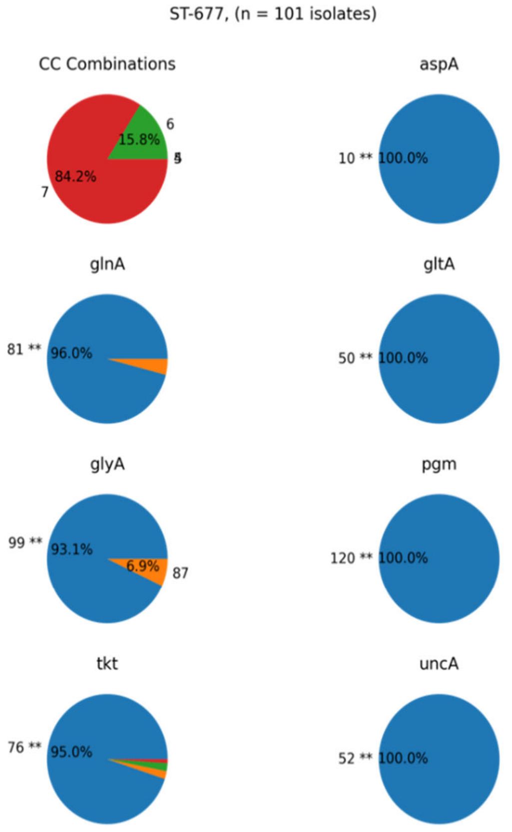
ST-403, (n = 106 isolates)

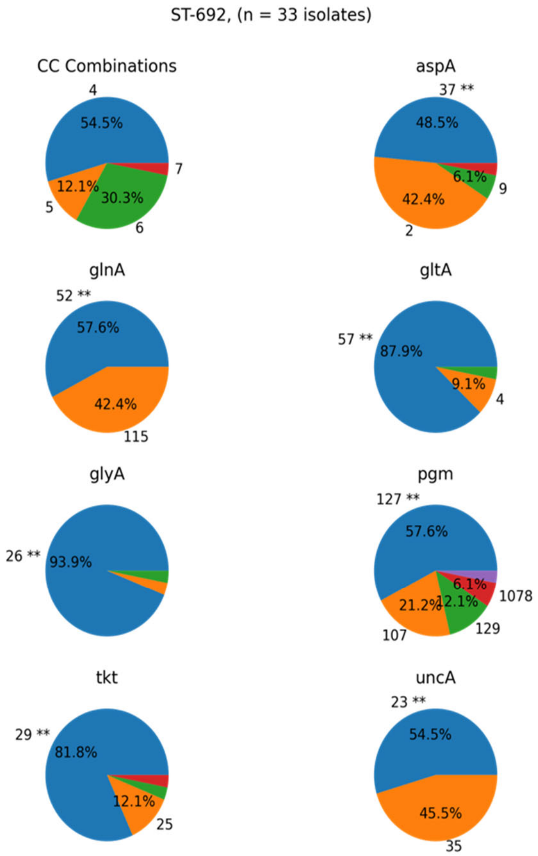

ST-446, (n = 20 isolates)
CC Combinations
5
30.0%
4
15.0%
35.0%
20.0%
6
7
aspA
47 **
85.0%
5.0%
10.0%
79
glnA
55 **
90.0%
5.0%
5.0%
gltA
5 **
95.0%
5.0%
glyA
10 **
100.0%
pgm
11 **
75.0%
25.0%
23
tkt
68 **
120
30.0%
25.0%
5.0%
5.0%
25.0%
10.0%
48
1
uncA
8 **
85.0%
15.0%
6

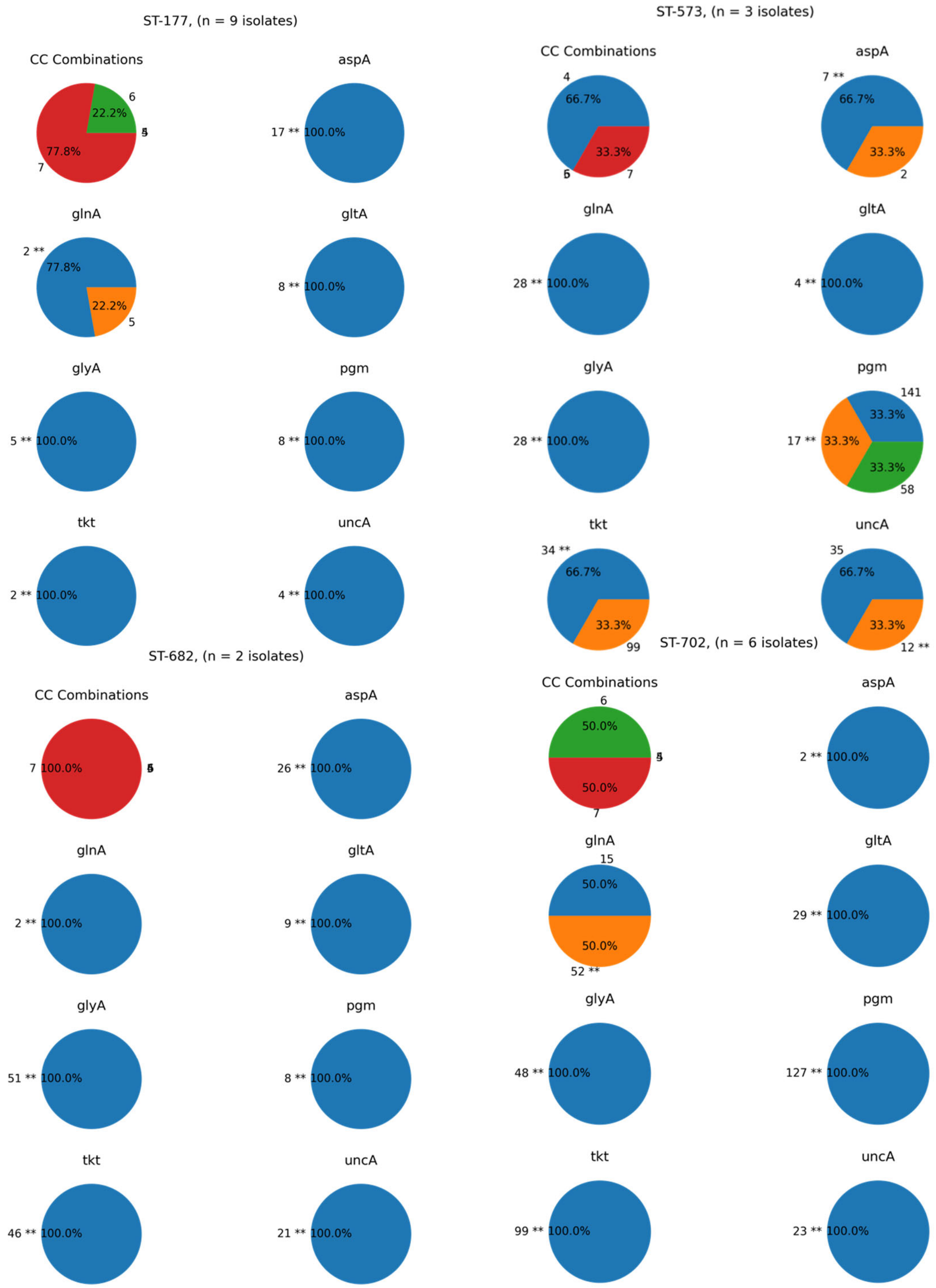

ST-460, (n = 53 isolates)
CC Combinations
5
6
20.8%
45.3%
4
32.1%
7
aspA
24 **
69.8%
11.3%
13.2%
9
77
glnA
30 **
83.0%
15.1%
21
gltA
2 **
90.6%
5.7%
5
glyA
2 **
98.1%
pgm
89 **
83.0%
13.2%
2
tkt
59 **
84.9%
5.7%
5.7%
3
120
uncA
6 **
98.1%

ST-508, (n = 88 isolates)
CC Combinations
6
34.1%
5
4
64.8%
7
aspA
1 **
100.0%
glnA
6 **
100.0%
gltA
60 **
65.9%
33.0%
22
glyA
24 **
97.7%
pgm
12 **
100.0%
tkt
28 **
100.0%
uncA
1 **
100.0%

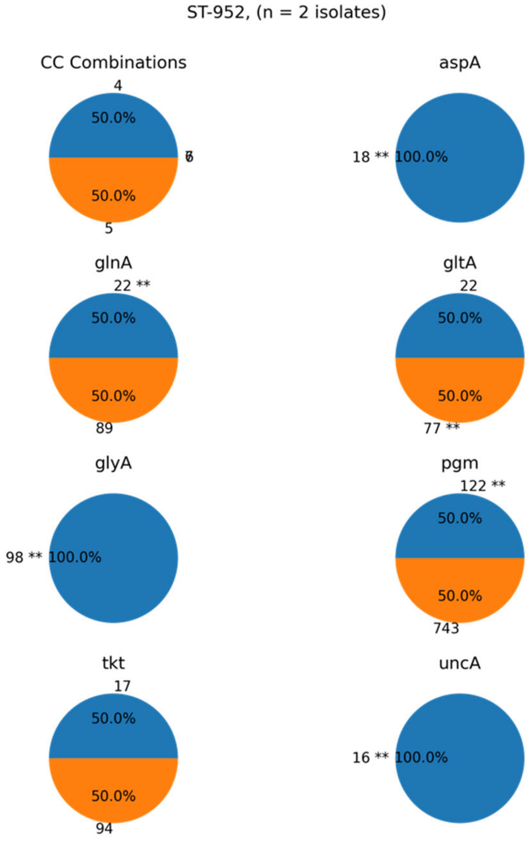

ST-574, (n = 230 isolates)
CC Combinations
6
5
7.8%
4
17.0%
72.2%
7
aspA
7 **
90.9%
8.3%
9
glnA
53 **
83.0%
5.2%
84
9.6%
4
gltA
2 **
91.7%
5.7%
1
glyA
10 **
87.8%
9.6%
68
pgm
11 **
98.7%
tkt
3 **
100.0%
uncA
3 **
83.0%
7.4%
5

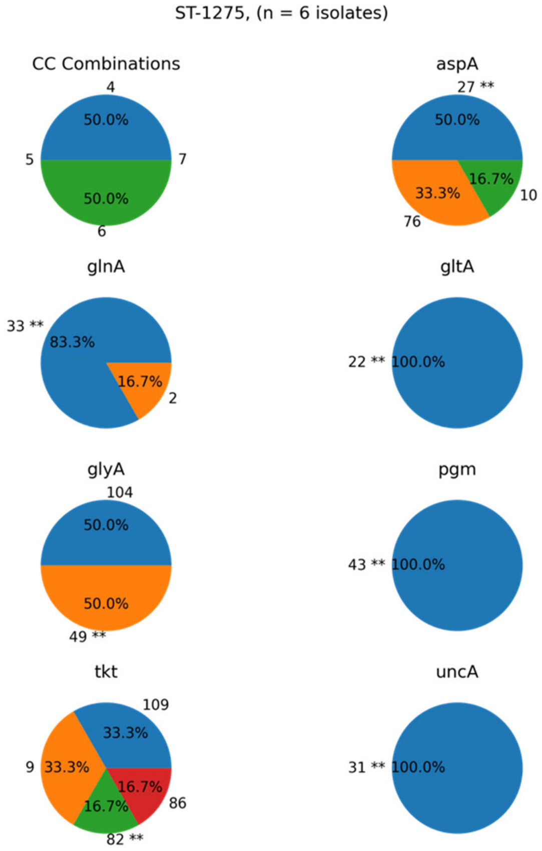

ST-677, (n = 101 isolates)
CC Combinations
6
15.8%
5
4
84.2%
7
aspA
10 **
100.0%
glnA
81 **
96.0%
gltA
50 **
100.0%
glyA
99 **
93.1%
6.9%
87
pgm
120 **
100.0%
tkt
76 **
95.0%
uncA
52 **
100.0%


ST-692, (n = 33 isolates)
CC Combinations
4
54.5%
7
12.1%
30.3%
5
6
aspA
37 **
48.5%
6.1%
9
42.4%
2
glnA
52 **
57.6%
42.4%
115
gltA
57 **
87.9%
9.1%
4
glyA
26 **
93.9%
pgm
127 **
57.6%
6.1%
1078
21.2%
12.1%
107
129
tkt
29 **
81.8%
12.1%
25
uncA
23 **
54.5%
45.5%
35



Isolates potentially for retirement:


ST-952, (n = 2 isolates)
CC Combinations
4
50.0%
7
6
50.0%
5
aspA
18 **
100.0%
glnA
22 **
50.0%
50.0%
89
gltA
22
50.0%
50.0%
77 **
glyA
98 **
100.0%
pgm
122 **
50.0%
50.0%
743
tkt
17
50.0%
50.0%
94
uncA
16 **
100.0%


ST-1275, (n = 6 isolates)
CC Combinations
4
50.0%
5
7
50.0%
6
aspA
27 **
50.0%
16.7%
10
33.3%
76
glnA
33 **
83.3%
16.7%
2
gltA
22 **
100.0%
glyA
104
50.0%
50.0%
49 **
pgm
43 **
100.0%
tkt
109
33.3%
9
33.3%
16.7%
86
16.7%
82 **
uncA
31 **
100.0%

**Statement of Authorship for joint/multi-authored papers for PGR thesis**
To appear at the end of each thesis chapter submitted as an article/paper

# Chapter 7:

| | |
|---|---|
| Title of Paper | Clonal-complex re-visited: Which clonal-complex will stay the same, will change, and which will retire for *Campylobacter* Alternatively: Clonal-complex re-assignment for *Campylobacter jejuni*: The Good, The Split and The Retirement. |
| Publication Status | ☐Published ☐ Accepted for Publication<br>☐Submitted for Publication ☒Unpublished and unsubmitted work written in a manuscript style |
| Publication Details | Authors: Dessislava Veltcheva, Margaret Varga, Michael B, Bonsall, Martin C. J. Maiden<br>Paper Style: Microbial Genomics from Microbiology Society (Research Article) |

Student Confirmation

| | | | |
|---|---|---|---|
| Student Name: | Dessislava Veltcheva | | |
| Contribution to the Paper | • Conceptualization of the research project<br>• Establishing Methodology<br>• Software and script creation<br>• Formal analysis and Investigation<br>• Writing – original draft preparation<br>• Writing – review and editing<br>• Data Visualization<br>• Project management | | |
| Signature Dessy | | Date | 02.10.2023 |

Supervisor Confirmation

| | | | |
|---|---|---|---|
| Supervisor name and title: Professor Michael Bonsall | | | |
| Supervisor comments<br>I can confirm that the candidate has ownership of the work and and substantial contributions to the research presented here. | | | |
| Signature | | Date | 2nd October 2023 |

# Chapter 8: Discussion and Conclusion

# Discussion

This discussion is split into three sections. The first section summarises the take-home message from each chapter and how they are connected to other parts of the thesis. The second part focuses on how this thesis and its findings impact the broader literature, leading to future progressions and suggestions on what can be done next in this field.

## Chapter 2

Chapter 2 [Emerging patterns of fluoroquinolone resistance in *C. jejuni* in the UK [1998–2018]] discusses *C.jejuni* and its fluoroquinolone resistance using logistic regression, clustering, and predictive analysis. *C. jejuni* is increasingly resistant to fluoroquinolone antibiotics, but very few studies have investigated variant-specific patterns of resistance across time. This study used 10,359 UK clinical isolates sampled between 1998 to 2018. The result showed six distinct patterns of fluoroquinolone sensitivity/resistance in *C. jejuni* across time. Trends show that some are fully resistant, some shifted from susceptible to resistant following a sigmoidal shape, and some remained susceptible over time.

This chapter demonstrated that investigating fluoroquinolone resistance on a species level hides much information. For example, the susceptible clusters (CC22, CC283, CC42, CC45 and CC61) do not match the overall species trend, indicating that some clinical cases could potentially be treated with fluoroquinolones. Furthermore, although there is a slight decrease, CC206 and CC48 are reducing resistance to fluoroquinolones, a trend not seen in other studies. On the other hand, the poultry specialists, which are CC353, CC354 and CC464, have had quite a different trend from each other. CC464 is predominantly resistant and has been only seen in high quantity since 2006. CC354, on the other hand, has been increasing in resistance, but even in 1998, it had high resistance. CC353 has a sigmoidal increase in resistance, further discussed in Chapter 5. Furthermore, the prediction of the future level of resistance has shown that not all clonal-complex could be predicted on its future trends. Although Group 2(CC353, CC354), Group 4(CC21, CC257, CC574, CC658) and Group 6(CC48, CC206) have tight confidence intervals, showing high confidence in their future resistance level.

## Chapter 3

In Chapter 3 [NeighbourGroups: a machine learning classification tool that assigns microbial Multi-locus genotypes to clusters], we have defined the methodology of the new classification “NeighbourGroups”. NeighbourGroups is a generalisable machine learning approach that reproducibly, robustly, and rapidly classifies Multi-locus sequence types with defined precision. This classification emerged after an issue in re-classifying clonal-complex 353 (Chapter 5). We have addressed this problem by leveraging the availability of large numbers of whole genome sequences available and implementing machine learning techniques.

We have used a minimum spanning tree to establish the NeighbourGroups for *C. jejuni*, but this could be created from other tree algorithms, such as neighbour joining and maximum likelihood. Then, a supervised machine learning algorithm called CatBoost was used to optimally predict the isolates to the group from fewer loci, in our case, using the

MLST loci. To choose the optimal number of groupings, the full tree (all isolates) and training tree (80% for training and 20% for testing) were used to identify whether the predictions were correct. This classification can be applied to any other bacterial species. It is also beneficial when whole genome sequence data are unavailable. For historic isolates where only their MLST loci are sequenced, we can predict their NeighbourGroups if the model of that bacterial species has been developed.

This chapter filled the gap in epidemiological classification that uses the cgMLST but is trained on fewer loci (MLST). By establishing the *C. jejuni* groups, we have assigned *C. jejuni* isolates to their corresponding NeighbourGroups, leading to our further analysis in chapters 4,5,6 and 7. This method can be applied to any other species, and we hope to establish NeighbourGroups classifications for other species soon.

### Chapter 4

In chapter 4 [CatBoost classifier "NeighbourGroups" classification applied to *Campylobacter jejuni*], we used "NeighbourGroups" onto *C. jejuni*. We found that 20 NeighbourGroups were optimal to describe the *C. jejuni* diversity. In this chapter, we also compare the NeighbourGroups and Clonal-complex. Here, we have shown that some clonal-complexes match with one NeighbourGroup, but some are split between multiple NeighbourGroups, indicating that we have various variants within one clonal-complex. For example, CC21 comprises two NeighbourGroups, 15 and 16.

Furthermore, by changing the number of groupings from 35 clonal-complexes of *C. jejuni* to 20 NeighbourGroups, we can see that some clonal-complexes merge into one NeighbourGroups, such as NeighbourGroup 14 containing CC22, CC42, and some CC362. The same type of merge can be found in poultry specialists CC353 and CC464, predominantly classified to NeighbourGroup 2. However, the other poultry specialist, CC354, is assigned to NeighbourGroup 1. Another aspect we have discovered in this chapter is the combination of 7 loci MLST and its adjusted Rand score. Although individually, they are insufficient to identify the NeighbourGroups correctly with high accuracy (ranging between 0.356 to 0.482), the combination of 7 loci leads to the adjusted Rand score of 0.895. However, we also found a single locus, hypothetical protein Cj1009c (CAMP0932), with an adjusted Rand score of 0.91.

The application of NeighbourGroups onto *C. jejuni* and through the comparisons with clonal-complex, we discovered more of its population dynamics and complexity. This chapter demonstrates the importance of the NeighbourGroups approach to bacterial epidemiology. By applying NeighbourGroups to *C. jejuni,* we have seen how this approach could be implemented in other bacterial species.

### Chapter 5

In Chapter 5 [*Campylobacter jejuni* clonal-complex 353 re-evaluated through clonal-complex assignment methods: re-classification to multiple NeighbourGroups and proposal of CC581 as a new clonal-complex], we investigate one of the problematic variants, CC353 and identify whether we can re-classify this variant using clonal-complex classification. In conclusion, we have found that CC353 comprises five groups: two primary and two small sub-groups, with one isolate in another group. A new clonal-complex, CC581, has been added to the original classification, making it 35 clonal-complexes for *C. jejuni*.

This chapter has not only shown that CC353 is polyphyletic, but it has also described the method upon which current clonal-complex classification is built. Although the way clonal-complex is formed is mentioned in the earlier papers (Maiden, 2006b), no papers have deconstructed the mechanisms of clonal-complex assignments from new STs in the way Chapter 5 has achieved. It has shown that some clonal-complexes have identical alleles across multiple clonal-complexes, which leads to some clonal-complexes being more closely related than others, such as CC52 and CC574 having three of their MLST alleles identical. It has demonstrated the difficulties of re-classifying one of 35 clonal-complexes since changing the central genotype from ST353 to ST5 will change the isolates in question and other isolate classifications. By applying NeighbourGroups, it has been shown that CC353 can be divided into 5 NeighbourGroups (NGroup 1,2,13,15 and 20). This chapter and its inability to re-classify within the clonal-complex was the reason NeighbourGroups was established (Chapter 3).

### Chapter 6

In Chapter 6: [Trends in Fluoroquinolone and Tetracycline Resistance of *Campylobacter jejuni* using NeighbourGroups], we have combined all the previous chapters and their findings. This study extensively examined *C. jejuni* variants from diverse sources from the

UK between 1997 and 2022, leading to the most extensive sample sizes in this thesis of 17,338 isolates. This chapter investigates fluoroquinolones and tetracycline resistance using the NeighbourGroup classification system. Findings indicate significant antimicrobial resistance for NeighbourGroup 1 and 2 (CC353, CC354, and CC464). NeighbourGroup 13 (CC443, CC574) exhibited high tetracycline resistance but not to the fluoroquinolone. NeighbourGroup 15 (a sub-group of CC21) was resistant to fluoroquinolones but not tetracycline. NeighbourGroup 7 (CC45) was susceptible to both antibiotics, showing further decreasing resistance to both.

This study combined Chapter 2's logistic regression and Chapter 3's NeighbourGroups method using Chapter 4's application to *C. jejuni*. By expanding it onto tetracycline, we have identified further patterns in *C. jejuni* and revealed evolving antimicrobial resistance patterns within *C. jejuni* populations. In this chapter, we shed light on the proportion of isolates in the year. It has shown that NeighbourGroups 15 is increasing in numbers, dominating the *Campylobacter* population from 2017 to 2022 by occupying 40% of overall isolates in our samples. It shows the complexity of *C. jejuni* and its antimicrobial resistance patterns.

### Chapter 7

In Chapter 7: [Clonal-complex re-visited: Which clonal-complex will stay the same, will change, and which will retire for *Campylobacter* ], we re-visit the clonal-complex classification. This chapter used the NeighbourGroups assignment from Chapter 4 but specifically focused on re-evaluating the clonal-complex as a classification. *C. jejuni* contains 35 clonal-complexes, and these classifications have been used widely since their establishment in 2001.

In this study, we re-evaluated these clonal-complexes by exploring whether their central genotypes match the majority alleles for each clonal-complex. We used 16,273 isolates across various countries and multiple sources with an assigned clonal-complex. This study identified that 16 clonal-complexes have their most prevalent alleles matching their central genotypes. However, 12 clonal-complexes no longer align with their original central genotypes. Also, three clonal-complexes (CC21, CC353, and CC581) are in the caution category, where significant concerns have been raised about keeping them as clonal-complexes, even though their overall alleles match. Especially for CC353, these have been further investigated in Chapter 5. On the other hand, clonal-complexes with low overall account (8 clonal-complexes) have been discussed to be retired as a classification due to their lack of quantity.

Furthermore, 11 clonal-complexes have their highest percentage of not matching the original central genotypes, which raises questions on why this has occurred. It could be that the wrong central genotypes were chosen for their groups in the initial assignment. On the other hand, the clonal-complex is assigned to one major NeighbourGroup (CC206 to NeighbourGroup 9 and CC464 to NeighbourGroup 2). The result indicated that these clonal-complexes are still behaving as a group, but their central genotypes could have evolved to another, or the initial central genotypes needed re-adjustments. Although CC581 was formed in chapter 5, in chapter 7, we have shown that it is one of the caution category variants that, although a small number, will be separated into multiple NeighbourGroups. Chapter 7 has solidified the problem of CC21, CC353 and CC581. Although 16 clonal-complexes are aligned to central genotypes, 11 are not, indicating whether we should still use clonal-complexes as a classification.

## Impact on broader literature:

### Antimicrobial Resistance

The increase in fluoroquinolones resistance of *C.jejuni* has been known in the broader literature across time (Wimalarathna *et al.*, 2013; Zawack *et al.*, 2016; Sproston, Wimalarathna and Sheppard, 2018; Park *et al.*, 2022). If antibiotics are prescribed for patients, macrolides are the choice of antibiotics. Through Chapter 2, we have shown that when discussing fluoroquinolone resistance, we should investigate *C. jejuni* on a variant level (clonal-complex) rather than identify trends as one species. Furthermore, chapter 6 revealed the further implication of how some variants, such as NeighbourGroup 1 and 2, will be resistant to both fluoroquinolones and tetracycline. On the other hand, NeighbourGroup 7 is susceptible and is decreasing in antimicrobial resistance. Therefore, this thesis has filled in the gap of variant-level resistant patterns that have not been done in the past.

### Bacterial classification and machine learning

This study has re-evaluated the fundamental concept of the bacterial classification method through chapters 5 and 7. Much literature focuses on distinguishing clonal-complexes using accessory genomes pan-genomes and applying complicated machine-learning algorithms to attribute sources (Mourkas, Bayliss, *et al.*, 2019; Arning *et al.*, 2021c). However, in this thesis, we have returned to the basics to deconstruct how MLST classification was built in the first instance (Chapter 5). Early findings revealed a sigmoidal increase in resistance in CC353, and the isolates were found across different branches in a phylogenetic tree (Chapter 2 and Chapter 5). Although all efforts were made to re-classify CC353, this was impossible within the clonal-complex scope. Thus, a creative solution was invented using cgMLST and MLST loci. Through the innovation of NeighbourGroups, we have noticed the significance of what this will entail for the whole *Campylobacter* community and other researchers working on different bacterial species that could benefit from this method. The available method to be built upon any phylogenetic trees also allows a large-scale project to train a model that could be used in the different bacterial species.

The generalisable and accessible tool for variant-level classification using a machine learning model has not been seen in *C. jejuni*. In the establishment phase of MLST schemes in 1998, housekeeping genes were selected for different bacteria (Kate E. Dingle *et al.*, 2002; Maiden, 2006b). Using these housekeeping genes and the abundant cgMLST from PubMLST, we can further apply the NeighbourGroups method to other organisms in the same method as in Chapter 4. As a side project, I established NeighbourGroups for *Acinetobacter baumannii*, *Neisseria gonorrhoeae* and *Haemophilus influenzae*. Through collaboration with other colleagues on different bacterial species, we hope to be able to implement NeighbourGroups in the PubMLST database.

## Future progressions

### NeighbourGroups: Sub-species and other organisms

Currently, NeighbourGroups have been applied to *C. jejuni*. Although we can assign NeighbourGroups directly to any *C. jejuni* isolates, this is not yet applicable to other species such as *C.coli*, *Campylobacter lari* and other less-known subspecies. Especially since 45 clonal-complexes contain 35 *C. jejuni* and 10 *C. coli* classifications. It is also known that there are three sub-clades in *C. coli*, where clade 1 is more clinical, and clades 2 and 3 are from environmental sources (Sheppard *et al.*, 2011). Furthermore, clade one is known to have introgression with *C. jejuni*, which could have a profound result in training a model that uses both *C. jejuni* and *C. coli* (Sheppard *et al.*, 2013). Thus, by adding *C. coli*, we could be able to predict both the sub-species and even different clades as well.

NeighbourGroups can be further applied to other bacterial organisms given the availability of cgMLST and known MLST or housekeeping genes. We discuss establishing NeighbourGroups with other species ranging from *Acinetobacter baumannii*, *Neisseria gonorrhoea* and *Haemophilus influenzae*. The application of NeighbourGroups onto these species will not only help establish the NeighbourGroups classification, but it will also help to create a backwards-compatible classification. Thus, we could also classify historical isolates that do not have cgMLST data using NeighbourGroups. Furthermore, as we have source attribution studies for clonal-complexes, we can also unravel source attribution with NeighbourGroups.

## NeighbourGroups: Expanding to different phylogenetic trees.

Another way to create a robust classification is to try other phylogenetic tree methodologies. Our study used a minimum spanning tree, but the model can also be trained with any different phylogenetic tree algorithm. The only issue with nucleotide-based algorithms is the species' high horizontal gene transfer rate (Golz and Stingl, 2021). Thus, we would have to run tools such as clonalframeML and ptARC to detect recombination in bacterial genomes and account for it in the phylogenetic reconstructed trees (Didelot and Wilson, 2015; Golz and Stingl, 2021).

I have also tried to use a maximum likelihood tree and clonalframeML on top of the isolate sequences. I had to align each cgMLST sequence to each other and then synthetically concatenate the sequences. However, with 10,359 isolates, over 1343 loci, it was not possible to run clonalframeML via HPC clusters (Richards, 2015b). We have tried to select isolates randomly and run clonalframeML, but it was impossible to complete the analysis due to a lack of memory and time on the cluster. In the future, trying different algorithms and carefully selecting representative isolates could potentially give us another NeighbourGroup. However, measuring bacterial evolution remains tricky and challenging, as shown in another paper (Calland *et al.*, 2021).

## NeighbourGroups: Worldwide Samples/Fieldwork

For this thesis, we have used the extensive isolate collections available from the PubMLST database. Although in Chapter 6, we have used over 17,000 isolates from across the globe, 81% of them are UK isolates. In the broader literature, we have had research that investigated antimicrobial resistance across 18 years, but this also focused on the UK and the USA (van Vliet *et al.*, 2022). The broader issue in the surveillance effort is the lack of funding and unavailability of medical infrastructures in low- and middle-income countries.

Thus, this will be a much bigger project, but finding ways to connect to different countries and trying to create a worldwide network of *Campylobacter* community will be a great project to aspire to. Through my studies, I have co-founded a public outreach platform called" Chicken Journal Club" (https://linktr.ee/chickenjclub), which brought people from different countries and by being completely free, we were able to gain more than 300 members who are working on poultry research globally. These initiatives have allowed people working on *Campylobacter* from Peru, Ethiopia, India, Japan, and other countries to come and talk to us about their work. To achieve global surveillance, we need to give every country a platform to speak about their isolates and find a way to help them with the lack of infrastructure. I believe that through new sequencing technologies, such as Oxford Nanopore Technologies, we can support more countries that lack the infrastructure to sequence their bacteria (Marin *et al.*, 2022).

Therefore, using the Chicken Journal Club network as an initiative to explore a more extensive project where we try to gain as many *Campylobacter* isolates from different countries and establish collaborative projects will produce more cohesive epidemiological data. Furthermore, a plan was initially to visit Peru to collect *Campylobacter* isolates from

Lima and Iquitos (Amazon Forest). This would have allowed me to have hands-on experience in the lab work and compare the UK isolates with the Peruvian isolates, known as the Wild West of resistant isolates, that are multi-drug resistant. Although this was impossible due to COVID-19, aiding isolate collection in another country lacking data points will benefit both parties.

## Antimicrobial Resistance

For this thesis, we have used the known knowledge of point mutation in the DNA gyrase (*gyrA*) gene to indicate phenotypic resistance to fluoroquinolones. Logistic regression was used to assess the relationship between fluoroquinolone resistance and time in chapters 2 and 6. However, the study will be enhanced if we have a minimum inhibitory concentration (MIC) level for some of these isolates. This would further allow us to identify whether some variants are more resistant than others and investigate why this could be. Furthermore, lab work on tetracycline and fluoroquinolone resistance would have enhanced Chapter 6 and further identified variant-level characteristics for NeighbourGroups.

For this analysis, besides Chapter 6, we have focused on fluoroquinolones mainly. However, it would be beneficial to understand the other commonly associated antimicrobial resistance, such as macrolides, beta-lactams and aminoglycosides (Wieczorek and Osek, 2013). Although currently macrolides are the most widely prescribed antibiotics, it would be intriguing to identify whether some variants could also be resistant to macrolides. Furthermore, obtaining the same variants for animal sources and comparing them to clinical isolates could be an interesting comparison to see whether any characteristics change in the food chain.

### Predictive and machine learning models:

For this analysis, the CatBoost algorithm, which performs well in categorical values, has been used to establish NeighbourGroups. However, there could have been a further analysis to compare different machine learning models and hyper tunning the parameters to make the model perform better. Furthermore, in terms of predictive analysis, we have touched on it in Chapter 2 but have not dived into the predictive analysis of predicting the antimicrobial resistance of a given isolate. Although I tried to create a resistance prediction based on metadata in my second year, this was not as useful since we have binary outcomes of resistance or susceptibility. Thus, if we can create a model where we can not only predict resistance from metadata but also predict sub-species level and MIC level of resistance, it would be beneficial. Although given the availability of rMLST classification, this model is not as useful if we have cgMLST but could be helpful if we only have MLST loci (Jolley *et al.*, 2012a).

### New Central genotypes

Furthermore, as shown in Chapter 6, now that we have cgMLST, we can also develop the best loci to distinguish the species into variants. We could also add sub-species onto our training data set so that it can determine not only one sub-species but multiple sub-species (ex., *C. jejuni* and *C. coli*) through one NeighbourGroup classification. Identifying the best housekeeping genes or even a single locus that can distinguish them into sub-variants will unravel many new research questions. Furthermore, the findings of these loci will allow us to develop gene knock-out models and studies further to investigate their importance in the survival of these bacteria.

### Generalist vs Specialist Questions: CC-21

The question of generalist and specialist variants comes up often in the literature on *C. jejuni* (Dearlove *et al.*, 2016b; Woodcock *et al.*, 2017; Epping, Antão and Semmler, 2021a). Some variants, such as CC353, CC354 and CC464, are known to be poultry specialists, and CC21 and CC45 are generalists (Sheppard *et al.*, 2014; Epping, Antão and Semmler, 2021b; Mouftah *et al.*, 2022). However, as we have discussed in chapters 5 and 7, the way clonal-complexes have been assigned has not changed since 2001, and it is a mechanistic approach. As we saw in Chapter 5, clonal-complexes are assigned based on a stand-alone list of central genotypes from which new MLSTs are assigned (Maiden, 2006b). The very first housekeeping gene is clonal-complex 21, and as shown in Chapter 5, this central genotype has multiple alleles identical to other central genotypes. It shares four out of 7 central genotypes with CC206 and CC48. This entails that some clonal-complexes are more closely related than others. We can also see the problem in source specialists such as CC353, CC354 and CC464; at least two loci are identical, leading to the fact that they are closely related in the classification system. Therefore, these specialists being known as poultry specialists is not very surprising if they are similar phylogenetically.

Furthermore, in my thesis, CC21 is not a “generalist” but a wrongly classified variant altogether. In Chapter 4, we can see that CC21 is split mainly into NeighbourGroups 15 and 16. Even in Chapter 7, gltA was the only locus that did not align with other majority alleles and changing the locus from allele 1 to allele 12 led to ST50, which split the overall CC21 into two groups. Thus, the way clonal-complex is being set needs to be re-evaluated carefully, and more people using this classification need to understand that within clonal-complex, some of them are more closely related than others.

### Establishment of CC581

Through Chapter 5, we have established CC581, which has been added to the clonal-complex schemes. However, in chapter 7, we have seen that CC581 is, in fact, one of the caution categories that have split into multiple NeighbourGroups. A study has already assigned CC581 to their isolates in Turkey, and adding this clonal-complex could be harming more than helping (Aydin *et al.*, 2023). Specifically, with the application of NeighbourGroups and even identifying potential new housekeeping genes, expanding the clonal-complex is not the best idea.

# Conclusion

In conclusion, this thesis not only deconstructed the current bacterial classification of MLST in *C. jejuni,* but it also established a new classification of NeighbourGroups that will be able to solve many questions across different species. This thesis started by identifying trends in fluoroquinolone resistance using clonal-complex. Then, whether CC353 is misclassified has emerged through the analysis. Through trying to re-classify, CC581 emerged as a new variant through my work, and we established that CC353 cannot be re-classified in the clonal-complex context. Thus, a machine learning approach, NeighbourGroups, was invented to solve the problem. Then, fluoroquinolones and tetracycline resistance were investigated using the newly applied NeighbourGroups on *C. jejuni*. Finally, clonal-complex classification has been re-visited to understand whether we can still use it with high accuracy. The answer is that for half of the clonal-complex, the classification still stands, but for others, such as CC21 and CC353, we should not be using them as classification due to multiple variants being classified into one.

# Bibliography

(Excluding Chapter 2 and 3 which are included in the published format)

1. Abraham, S. *et al.* (2020a) 'Emergence of fluoroquinolone-resistant campylobacter jejuni and campylobacter coli among Australian chickens in the absence of fluoroquinolone use', *Applied and Environmental Microbiology*. American Society for Microbiology, 86(8). doi: 10.1128/AEM.02765-19.
2. Abraham, S. *et al.* (2020b) 'Emergence of Fluoroquinolone-Resistant Campylobacter jejuni and Campylobacter coli among Australian Chickens in the Absence of Fluoroquinolone Use', *Applied and Environmental Microbiology*. American Society for Microbiology (ASM), 86(8). doi: 10.1128/AEM.02765-19.
3. Aksomaitiene, J. *et al.* (2019) 'Overlap of antibiotic resistant Campylobacter jejuni MLST genotypes isolated from humans, broiler products, dairy cattle and wild birds in Lithuania', *Frontiers in Microbiology*. Frontiers Media S.A., 10(JUN), p. 387687. doi: 10.3389/FMICB.2019.01377/BIBTEX.
4. Arning, N. *et al.* (2021a) 'Machine learning to predict the source of campylobacteriosis using whole genome data', *PLOS Genetics*. Public Library of Science, 17(10), p. e1009436. doi: 10.1371/JOURNAL.PGEN.1009436.
5. Arning, N. *et al.* (2021b) 'Machine learning to predict the source of campylobacteriosis using whole genome data', *PLoS genetics*. PLoS Genet, 17(10). doi: 10.1371/JOURNAL.PGEN.1009436.
6. Arning, N. *et al.* (2021c) 'Machine learning to predict the source of campylobacteriosis using whole genome data', *PLoS genetics*. PLoS Genet, 17(10). doi: 10.1371/JOURNAL.PGEN.1009436.

7. Awad, W. A., Hess, C. and Hess, M. (2018) ‘Re-thinking the chicken–Campylobacter jejuni interaction: a review’, *https://doi.org/10.1080/03079457.2018.1475724*. Taylor & Francis, 47(4), pp. 352–363. doi: 10.1080/03079457.2018.1475724.
8. Aydin, F. *et al.* (2023) ‘MLST genotypes and quinolone resistance profiles of Campylobacter jejuni isolates from various sources in Turkey’, *International Journal of Food Microbiology*. Elsevier, 391–393, p. 110137. doi: 10.1016/J.IJFOODMICRO.2023.110137.
9. Beard, L. and Aghassibakes, N. (2021) ‘Tableau (version 2020.3)’, *Journal of the Medical Library Association : JMLA*. Medical Library Association, 109(1), p. 159. doi: 10.5195/JMLA.2021.1135.
10. Bolinger, H. and Kathariou, S. (2017) ‘The current state of macrolide resistance in Campylobacter spp.: Trends and impacts of resistance mechanisms’, *Applied and Environmental Microbiology*. American Society for Microbiology, 83(12). doi: 10.1128/AEM.00416-17/ASSET/53EAD321-296B-4F08-988C-69034EF2D667/ASSETS/GRAPHIC/ZAM9991178940001.JPEG.
11. Calland, J. K. *et al.* (2021) ‘Quantifying bacterial evolution in the wild: A birthday problem for Campylobacter lineages’, *PLOS Genetics*. Public Library of Science, 17(9), p. e1009829. doi: 10.1371/JOURNAL.PGEN.1009829.
12. Cardoso, M. J. *et al.* (2021) ‘Cross-contamination events of Campylobacter spp. in domestic kitchens associated with consumer handling practices of raw poultry’, *International Journal of Food Microbiology*. Elsevier, 338, p. 108984. doi: 10.1016/J.IJFOODMICRO.2020.108984.

13. Cheng, Y. W. and Fischer, M. (2022) ‘Campylobacter’, *Encyclopedia of Gastroenterology, Second Edition*. StatPearls Publishing, pp. 424–427. doi: 10.1016/B978-0-12-801238-3.11068-2.
14. Chopra, I. and Roberts, M. (2001) ‘Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance’, *Microbiology and Molecular Biology Reviews*. American Society for Microbiology (ASM), 65(2), p. 232. doi: 10.1128/MMBR.65.2.232-260.2001.
15. Clarke, A. K. and Ajlouni, S. (2021) ‘Recommended Practices to EliminateCampylobacter from Live Birds andChicken Meat in Japan’, *Food Safety*. Food Safety Commission of Japan, 9(3), p. 57. doi: 10.14252/FOODSAFETYFSCJ.D-20-00021.
16. Clarke, T. N. *et al.* (2021) ‘A systematic review and meta-analysis of Penner serotype prevalence of Campylobacter jejuni in low- and middle-income countries’, *PLOS ONE*. Public Library of Science, 16(5), p. e0251039. doi: 10.1371/JOURNAL.PONE.0251039.
17. Cody, A. J. *et al.* (2010) ‘Ciprofloxacin-resistant campylobacteriosis in the UK’, *The Lancet*. Elsevier B.V., 376(9757), p. 1987. doi: 10.1016/S0140-6736(10)62261-1.
18. Cody, Alison J. *et al.* (2012) ‘A longitudinal 6-year study of the molecular epidemiology of clinical Campylobacter isolates in Oxfordshire, United Kingdom’, *Journal of Clinical Microbiology*. American Society for Microbiology1752 N St., N.W., Washington, DC, 50(10), pp. 3193–3201. doi: 10.1128/JCM.01086-12/SUPPL_FILE/ZJM999091935SO1.XLS.
19. Cody, Alison J *et al.* (2012) ‘A Longitudinal 6-Year Study of the Molecular

Epidemiology of Clinical Campylobacter Isolates in Oxfordshire, United Kingdom'. doi: 10.1128/JCM.01086-12.

20. Cody, A. J. *et al.* (2015) 'Wild bird-associated Campylobacter jejuni isolates are a consistent source of human disease, in Oxfordshire, United Kingdom', *Environmental Microbiology Reports*. Wiley-Blackwell, 7(5), pp. 782–788. doi: 10.1111/1758-2229.12314/FORMAT/PDF.
21. Cody, Alison J. *et al.* (2017) 'Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates', *Journal of Clinical Microbiology*. American Society for Microbiology (ASM), 55(7), p. 2086. doi: 10.1128/JCM.00080-17.
22. Cody, Alison J *et al.* (2017) 'Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates', *Journal of Clinical Microbiology*. Edited by D. J. Diekema, 55(7), pp. 2086–2097. doi: 10.1128/JCM.00080-17.
23. Cody, A. J. *et al.* (2019) 'A systematic review of source attribution of human campylobacteriosis using multilocus sequence typing', *Eurosurveillance*. European Centre for Disease Prevention and Control, 24(43), p. 1. doi: 10.2807/1560-7917.ES.2019.24.43.1800696.
24. Colles, F. M. *et al.* (2003a) 'Genetic Diversity of Campylobacter jejuniIsolates from Farm Animals and the FarmEnvironment', *Applied and Environmental Microbiology*. American Society for Microbiology, 69(12), pp. 7409–7413. doi: 10.1128/AEM.69.12.7409-7413.2003.
25. Colles, F. M. *et al.* (2003b) 'Genetic Diversity of Campylobacter jejuniIsolates from

Farm Animals and the FarmEnvironment', *Applied and Environmental Microbiology*. American Society for Microbiology, 69(12), pp. 7409–7413. doi: 10.1128/AEM.69.12.7409-7413.2003.

26. Costa, D. and Iraola, G. (2019) 'Pathogenomics of emerging Campylobacter species', *Clinical Microbiology Reviews*. American Society for Microbiology, 32(4). doi: 10.1128/CMR.00072-18/ASSET/8A262206-7A2F-4596-87CA-BAB0C61E1B29/ASSETS/GRAPHIC/CMR.00072-18-F0002.JPEG.
27. Couturier, B. A., Hale, D. V. C. and Couturier, M. R. (2012) 'Association of Campylobacter upsaliensis with persistent bloody diarrhea', *Journal of Clinical Microbiology*. American Society for Microbiology1752 N St., N.W., Washington, DC, 50(11), pp. 3792–3794. doi: 10.1128/JCM.01807-12/ASSET/7D01BD53-5A7F-4A88-B70D-3B5027E7F2F4/ASSETS/GRAPHIC/ZJM9990920230001.JPEG.
28. Daniel, N. *et al.* (2020) 'The Burden of Foodborne Disease in the UK 2018 Food Standards Agency'.
29. Dearlove, B. L. *et al.* (2016a) 'Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections', *The ISME Journal*. Nature Publishing Group, 10(3), p. 721. doi: 10.1038/ISMEJ.2015.149.
30. Dearlove, B. L. *et al.* (2016b) 'Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections', *ISME Journal*. Nature Publishing Group, 10(3), pp. 721–729. doi: 10.1038/ismej.2015.149.
31. Devleesschauwer, B. *et al.* (2017) 'Health and economic burden of Campylobacter', *Campylobacter: Features, Detection, and Prevention of Foodborne Disease*.

Academic Press, pp. 27–40. doi: 10.1016/B978-0-12-803623-5.00002-2.

32. Didelot, X. and Maiden, M. C. J. (2010) ‘Impact of recombination on bacterial evolution’, *Trends in Microbiology*. Elsevier, 18(7), pp. 315–322. doi: 10.1016/j.tim.2010.04.002.

33. Didelot, X. and Wilson, D. J. (2015) ‘ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes’, *PLoS Computational Biology*. Public Library of Science, 11(2), p. e1004041. doi: 10.1371/journal.pcbi.1004041.

34. Dieckmann, A. L. *et al.* (2021) ‘Genome and Methylome analysis of a phylogenetic novel Campylobacter coli cluster with C. jejuni introgression’, *Microbial Genomics*. Microbiology Society, 7(10), p. 679. doi: 10.1099/MGEN.0.000679.

35. van Dijk, E. L. *et al.* (2018) ‘The Third Revolution in Sequencing Technology’, *Trends in Genetics*. Elsevier Ltd, 34(9), pp. 666–681. doi: 10.1016/j.tig.2018.05.008.

36. Dingle, K. E. *et al.* (2001a) ‘Multilocus sequence typing system for Campylobacter jejuni’, *Journal of Clinical Microbiology*. American Society for Microbiology, 39(1), pp. 14–23. doi: 10.1128/JCM.39.1.14-23.2001/ASSET/661246F4-6684-4ACE-AE4A-AFBE3C8759D9/ASSETS/GRAPHIC/JM0110757003.JPEG.

37. Dingle, K. E. *et al.* (2001b) ‘Multilocus sequence typing system for Campylobacter jejuni’, *Journal of Clinical Microbiology*, 39(1), pp. 14–23. doi: 10.1128/JCM.39.1.14-23.2001/FORMAT/EPUB.

38. Dingle, Kate E. *et al.* (2002) ‘ Molecular Characterization of Campylobacter jejuni Clones: A Basis for Epidemiologic Investigation ’, *Emerging Infectious Diseases*. Centers for Disease Control and Prevention (CDC), 8(9), pp. 949–955. doi:

10.3201/eid0809.02-0122.

39. Dingle, Kate E *et al.* (2002) *Molecular Characterization of Campylobacter jejuni Clones: A Basis for Epidemiologic Investigation*, *Emerging Infectious Diseases •*. Available at: http://campylobacter.mlst.net/.
40. Dingle, K. E. *et al.* (2005) ‘Sequence typing and comparison of population biology of Campylobacter coli and Campylobacter jejuni’, *Journal of Clinical Microbiology*. American Society for Microbiology, 43(1), pp. 340–347. doi: 10.1128/JCM.43.1.340-347.2005/ASSET/116C98E3-DE98-4691-8196-FD7C533D6570/ASSETS/GRAPHIC/ZJM0010549200003.JPEG.
41. Dingle, K. E. *et al.* (2008) ‘Extended Sequence Typing of Campylobacter spp., United Kingdom’, *Emerging Infectious Diseases*. Centers for Disease Control and Prevention, 14(10), p. 1620. doi: 10.3201/EID1410.071109.
42. Elhadidy, M. *et al.* (2020) ‘Antimicrobial resistance patterns and molecular resistance markers of Campylobacter jejuni isolates from human diarrheal cases’, *PLOS ONE*. Edited by I. Karunasagar. Public Library of Science, 15(1), p. e0227833. doi: 10.1371/journal.pone.0227833.
43. Elviss, N. C. *et al.* (2009) ‘Amoxicillin therapy of poultry flocks: effect upon the selection of amoxicillin-resistant commensal Campylobacter spp.’, *Journal of Antimicrobial Chemotherapy*. Oxford Academic, 64(4), pp. 702–711. doi: 10.1093/JAC/DKP277.
44. Emele, M. F. *et al.* (2019) ‘Proteotyping as alternate typing method to differentiate Campylobacter coli clades’, *Scientific Reports 2019 9:1*. Nature Publishing Group, 9(1), pp. 1–11. doi: 10.1038/s41598-019-40842-w.

45. Endtz, H. P. (2020) ‘Campylobacter Infections’, *Hunter’s Tropical Medicine and Emerging Infectious Diseases*. Elsevier, pp. 507–511. doi: 10.1016/B978-0-323-55512-8.00050-8.
46. Epping, L. *et al.* (2021) ‘Genome-wide insights into population structure and host specificity of Campylobacter jejuni’, *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–15. doi: 10.1038/s41598-021-89683-6.
47. Epping, L., Antão, E. M. and Semmler, T. (2021a) ‘Population Biology and Comparative Genomics of Campylobacter Species’, *Current Topics in Microbiology and Immunology*. Springer Science and Business Media Deutschland GmbH, 431, pp. 59–78. doi: 10.1007/978-3-030-65481-8_3/FIGURES/3.
48. Epping, L., Antão, E. M. and Semmler, T. (2021b) ‘Population Biology and Comparative Genomics of Campylobacter Species’, *Current Topics in Microbiology and Immunology*. Springer Science and Business Media Deutschland GmbH, 431, pp. 59–78. doi: 10.1007/978-3-030-65481-8_3/FIGURES/3.
49. Espinoza, N. *et al.* (2020) ‘Validation of the T86I mutation in the gyrA gene as a highly reliable real time PCR target to detect Fluoroquinolone-resistant Campylobacter jejuni’, *BMC Infectious Diseases*. BMC, 20(1). doi: 10.1186/S12879-020-05202-4.
50. Facciolà, A. *et al.* (2017) ‘Campylobacter: from microbiology to prevention’, *Journal of Preventive Medicine and Hygiene*. Pacini Editore, 58(2), p. E79. Available at: /pmc/articles/PMC5584092/ (Accessed: 12 April 2023).

51. Feil, E. J. *et al.* (2004) ‘eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data’, *Journal of Bacteriology*. American Society for Microbiology (ASM), 186(5), p. 1518. doi: 10.1128/JB.186.5.1518-1530.2004.
52. Fitzgerald, C. (2015) ‘Campylobacter’, *Clinics in Laboratory Medicine*. W.B. Saunders, 35(2), pp. 289–298. doi: 10.1016/j.cll.2015.03.001.
53. Food Standards Agency (2015) *A microbiological survey of Campylobacter contamination in fresh whole UK- produced chilled chickens at retail sale- February 2014 to February 2015*, *UK- wide survery of campylobacter contamination on fresh chickens at retail.* Available at: https://www.food.gov.uk/sites/default/files/full-campy-survey-report.pdf (Accessed: 3 October 2020).
54. Frieda, J. *et al.* (2019) *Antimicrobial resistance in Campylobacter jejuni and Campylobacter coli from retail chilled chicken in the UK Year 4:2017-2018*. Available at: https://www.food.gov.uk/research/antimicrobial-resistance/antimicrobial-resistance-in-campylobacter-jejuni-and-campylobacter-coli-from-retail-chilled-chicken-in-the-uk-1 (Accessed: 5 October 2023).
55. Friedrich, A. *et al.* (2016) ‘Seasonality of Campylobacter jejuni isolates associated with human campylobacteriosis in the Manawatu region, New Zealand’, *Epidemiology & Infection*. Cambridge University Press, 144(4), pp. 820–828. doi: 10.1017/S0950268815002009.
56. Di Giannatale, E. *et al.* (2019) ‘Thermotolerant Campylobacter spp. in chicken and bovine meat in Italy: Prevalence, level of contamination and molecular characterization of isolates’, *PLOS ONE*. Public Library of Science, 14(12), p.

e0225957. doi: 10.1371/JOURNAL.PONE.0225957.

57. Goddard, M. R. *et al.* (2022) ‘A restatement of the natural science evidence base regarding the source, spread and control of Campylobacter species causing human disease’, *Proceedings of the Royal Society B: Biological Sciences*. Royal Society Publishing, 289(1977). doi: 10.1098/RSPB.2022.0400/.

58. Golz, J. C. and Stingl, K. (2021) ‘Natural Competence and Horizontal Gene Transfer in Campylobacter’, *Current Topics in Microbiology and Immunology*. Springer Science and Business Media Deutschland GmbH, 431, pp. 265–292. doi: 10.1007/978-3-030-65481-8_10/FIGURES/2.

59. Gripp, E. *et al.* (2011) ‘Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle’, *BMC Genomics*. BioMed Central, 12(1), p. 584. doi: 10.1186/1471-2164-12-584.

60. Grove-White, D. H. *et al.* (2011) ‘Molecular epidemiology and genetic diversity of Campylobacter jejuni in ruminants’, *Epidemiology and infection*. Epidemiol Infect, 139(11), pp. 1661–1671. doi: 10.1017/S0950268810002736.

61. Guerrant, R. L. *et al.* (1978) ‘Campylobacteriosis in man: pathogenic mechanisms and review of 91 bloodstream infections’, *The American journal of medicine*. Am J Med, 65(4), pp. 584–592. doi: 10.1016/0002-9343(78)90845-8.

62. Habib, I., Uyttendaele, M. and De Zutter, L. (2010) ‘Survival of poultry-derived Campylobacter jejuni of multilocus sequence type clonal-complexes 21 and 45 under freeze, chill, oxidative, acid and heat stresses’, *Food microbiology*. Food Microbiol, 27(6), pp. 829–834. doi: 10.1016/J.FM.2010.04.009.

63. Hakeem, M. J. and Lu, X. (2021) ‘Survival and Control of Campylobacter in Poultry

Production Environment', *Frontiers in Cellular and Infection Microbiology*. Frontiers Media S.A., 10, p. 904. doi: 10.3389/FCIMB.2020.615049/BIBTEX.

64. Haldenby, S. *et al.* (2020) 'Increasing prevalence of a fluoroquinolone resistance mutation amongst Campylobacter jejuni isolates from four human infectious intestinal disease studies in the United Kingdom', *PLoS ONE*. Public Library of Science, 15(1). doi: 10.1371/JOURNAL.PONE.0227535.
65. Han, J. *et al.* (2012a) 'A Fluoroquinolone Resistance Associated Mutation in gyrA Affects DNA Supercoiling in Campylobacter jejuni', *Frontiers in Cellular and Infection Microbiology*. Frontiers Media SA, 2, p. 21. doi: 10.3389/FCIMB.2012.00021.
66. Han, J. *et al.* (2012b) 'A Fluoroquinolone Resistance Associated Mutation in gyrA Affects DNA Supercoiling in Campylobacter jejuni', *Frontiers in Cellular and Infection Microbiology*. Frontiers, 2, p. 21. doi: 10.3389/fcimb.2012.00021.
67. Havelaar, A. H. *et al.* (2015) 'World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010', *PLOS Medicine*. Public Library of Science, 12(12), p. e1001923. doi: 10.1371/JOURNAL.PMED.1001923.
68. Hu, T. *et al.* (2021) 'Next-generation sequencing technologies: An overview', *Human Immunology*. Elsevier, 82(11), pp. 801–811. doi: 10.1016/J.HUMIMM.2021.02.012.
69. Hughes, R. A. C. *et al.* (2007) 'Immunotherapy for Guillain-Barré syndrome: a systematic review', *Brain*. Oxford Academic, 130(9), pp. 2245–2257. doi: 10.1093/BRAIN/AWM004.

70. Humphrey, S. *et al.* (2014) ‘Campylobacter jejuni is not merely a commensal in commercial broiler chickens and affects bird welfare’, *mBio*. American Society for Microbiology, 5(4). doi: 10.1128/mBio.01364-14.
71. Humphrey, T. J. *et al.* (2005) ‘Prevalence and Subtypes of Ciprofloxacin-Resistant Campylobacter spp. in Commercial Poultry Flocks before, during, and after Treatment with Fluoroquinolones’, *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology (ASM), 49(2), p. 690. doi: 10.1128/AAC.49.2.690-698.2005.
72. Hur, J. I. *et al.* (2022) ‘Phylogenetic Association and Genetic Factors in Cold Stress Tolerance in Campylobacter jejuni’, *Microbiology Spectrum*. American Society for Microbiology, 10(6). doi: 10.1128/SPECTRUM.02681-22/SUPPL_FILE/SPECTRUM.02681-22-S0001.PDF.
73. Igwaran, A. and Okoh, A. I. (2019) ‘Human campylobacteriosis: A public health concern of global importance’, *Heliyon*. Elsevier Ltd, p. e02814. doi: 10.1016/j.heliyon.2019.e02814.
74. *Information for Health Professionals | Campylobacter | CDC* (2023). Available at: https://www.cdc.gov/campylobacter/technical.html (Accessed: 12 April 2023).
75. Inglis, G. D., Taboada, E. N. and Boras, V. F. (2021) ‘Rates of fluoroquinolone resistance in domestically acquired campylobawecter jejuni are increasing in people living within a model study location in canada’, *Canadian Journal of Microbiology*. Canadian Science Publishing, 67(1), pp. 37–52. doi: 10.1139/CJM-2020-0146/SUPPL_FILE/CJM-2020-0146SUPPLA.PDF.
76. Iovine, Nicole M (2013) ‘Resistance mechanisms in Campylobacter jejuni.’,

*Virulence*. Taylor & Francis, 4(3), pp. 230–40. doi: 10.4161/viru.23753.

77. Iovine, Nicole M. (2013) ‘Resistance mechanisms in Campylobacter jejuni’, *http://dx.doi.org/10.4161/viru.23753*. Taylor & Francis, 4(3), pp. 230–240. doi: 10.4161/VIRU.23753.
78. Islam, Z. *et al.* (2009) ‘Comparative Genotyping of Campylobacter jejuni Strains from Patients with Guillain-Barré Syndrome in Bangladesh’, *PLOS ONE*. Public Library of Science, 4(9), p. e7257. doi: 10.1371/JOURNAL.PONE.0007257.
79. Jain, M. *et al.* (2015) ‘Improved data analysis for the MinION nanopore sequencer’, *Nature Methods 2015 12:4*. Nature Publishing Group, 12(4), pp. 351–356. doi: 10.1038/nmeth.3290.
80. Jehanne, Q. *et al.* (2020) ‘Genome-wide identification of host-segregating SNPs for source attribution of clinical Campylobacter coli isolates’, *Applied and Environmental Microbiology*. American Society for Microbiology, 86(24), pp. e01787-20. doi: 10.1128/AEM.01787-20.
81. Jolley, K. A. *et al.* (2012a) ‘Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain’, *Microbiology*. Microbiology Society, 158(Pt 4), p. 1005. doi: 10.1099/MIC.0.055459-0.
82. Jolley, K. A. *et al.* (2012b) ‘Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain’, *Microbiology*. Microbiology Society, 158(Pt 4), p. 1005. doi: 10.1099/MIC.0.055459-0.
83. Jolley, K. A., Bray, J. E. and Maiden, M. C. . (2018) ‘Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]’, *Wellcome Open Research*. F1000

Research Ltd, 3. doi: 10.12688/wellcomeopenres.14826.1.

84. Jolley, K. A. and Maiden, M. C. J. (2010) 'BIGSdb: Scalable analysis of bacterial genome variation at the population level', *BMC Bioinformatics*. BMC Bioinformatics, 11. doi: 10.1186/1471-2105-11-595.
85. Jorgensen, F. *et al.* (2015) *Antimicrobial Resistance in Campylobacter jejuni and Campylobacter coli from Retail Chilled hicken in the UK Year 1*, *Public Health England*. Available at: https://www.food.gov.uk/research/antimicrobial-resistance/antimicrobial-resistance-in-campylobacter-jejuni-and-campylobacter-coli-from-retail-chilled-chicken-in-the-uk (Accessed: 5 October 2023).
86. Jorgensen, F. *et al.* (2017) *Antimicrobial Resistance in Campylobacter jejuni and Campylobacter coli from Retail Chilled hicken in the UK Year 2*. Available at: https://www.food.gov.uk/research/antimicrobial-resistance/antimicrobial-resistance-in-campylobacter-jejuni-and-campylobacter-coli-from-retail-chilled-chicken-in-the-uk-0 (Accessed: 5 October 2023).
87. Jorgensen, F. *et al.* (2018) *Year 3 Report A microbiological survey of Campylobacter contamination in fresh whole UK-produced chilled chickens at retail sale*. Available at: https://www.food.gov.uk/sites/default/files/media/document/campylobacter-contamination-in-fresh-whole-uk-produced-chilled-chickens-at-retail-sale-year-3-2016-2017_0.pdf (Accessed: 29 May 2019).
88. Jorgensen, F., Corcionivoschi, N., *et al.* (2019) *Antimicrobial resistance in Campylobacter jejuni and Campylobacter coli from retail chilled chicken in the UK (Year 4: 2017-18) FSA Project FS102121 Forming part of the project: A*

*microbiological survey of Campylobacter contamination in fresh whole UK pro*.

89. Jorgensen, F., Rodgers, J., *et al.* (2019) ‘Levels and trends of antimicrobial resistance in Campylobacter spp. from chicken in the UK’. doi: 10.46756/SCI.FSA.DUD728.
90. Jorgensen, F., Kesby, M., *et al.* (2021) ‘A microbiological survey of campylobacter contamination in fresh whole UK-produced chilled chickens at retail sale (Y6)’. doi: 10.46756/SCI.FSA.XXZ973.
91. Jorgensen, F., Charlett, A., *et al.* (2021) ‘A survey of the levels of Campylobacter spp. contamination and prevalence of selected antimicrobial resistance determinants in fresh whole UK-produced chilled chickens at retail sale (Y5)’. doi: 10.46756/SCI.FSA.XLS618.
92. Kaakoush, N. O. *et al.* (2015a) ‘Global Epidemiology of Campylobacter Infection.’, *Clinical microbiology reviews*. American Society for Microbiology Journals, 28(3), pp. 687–720. doi: 10.1128/CMR.00006-15.
93. Kaakoush, N. O. *et al.* (2015b) ‘Global Epidemiology of Campylobacter Infection’. doi: 10.1128/CMR.00006-15.
94. Kang, C. R., Bang, J. H. and Cho, S. Il (2019) ‘Campylobacter jejuni Foodborne Infection Associated with Cross-contamination: Outbreak in Seoul in 2017’, *Infection & Chemotherapy*. The Korean Society of Infectious Diseases and Korean Society for Chemotherapy, 51(1), pp. 21–27. doi: 10.3947/IC.2019.51.1.21.
95. Katoh, K. and Standley, D. M. (2013) ‘MAFFT multiple sequence alignment software version 7: improvements in performance and usability’, *Molecular biology and evolution*. Mol Biol Evol, 30(4), pp. 772–780. doi: 10.1093/MOLBEV/MST010.

96. Kenyon, J. *et al.* (2020) ‘Campylobacter outbreak associated with raw drinking milk, North West England, 2016’, *Epidemiology & Infection*. Cambridge University Press, 148, p. e13. doi: 10.1017/S0950268820000096.
97. Kittl, S. *et al.* (2013a) ‘Source Attribution of Human Campylobacter Isolates by MLST and Fla-Typing and Association of Genotypes with Quinolone Resistance’, *PLOS ONE*. Public Library of Science, 8(11), p. e81796. doi: 10.1371/JOURNAL.PONE.0081796.
98. Kittl, S. *et al.* (2013b) ‘Source Attribution of Human Campylobacter Isolates by MLST and Fla-Typing and Association of Genotypes with Quinolone Resistance’, *PLOS ONE*. Public Library of Science, 8(11), p. e81796. doi: 10.1371/JOURNAL.PONE.0081796.
99. Kovač, J. *et al.* (2014a) ‘The evidence for clonal spreading of quinolone resistance with a particular clonal-complex of Campylobacter jejuni’, *Epidemiology and Infection*. Cambridge University Press, 142(12), pp. 2595–2603. doi: 10.1017/S0950268813003245.
100. Kovač, J. *et al.* (2014b) ‘The evidence for clonal spreading of quinolone resistance with a particular clonal-complex of Campylobacter jejuni’, *Epidemiology & Infection*. Cambridge University Press, 142(12), pp. 2595–2603. doi: 10.1017/S0950268813003245.
101. Kwan, P. S. L. *et al.* (2008) ‘Longitudinal study of the molecular epidemiology of Campylobacter jejuni in cattle on dairy farms’, *Applied and Environmental Microbiology*. American Society for Microbiology, 74(12), pp. 3626–3633. doi: 10.1128/AEM.01669-07/ASSET/9BEB3C42-E303-47F5-96CE-

4BE08226B330/ASSETS/GRAPHIC/ZAM0120889460002.JPEG.

102. Lai, H. *et al.* (2021) ‘An investigation into the critical factors influencing the spread of campylobacter during chicken handling in commercial kitchens in China’, *Microorganisms*. MDPI AG, 9(6). doi: 10.3390/MICROORGANISMS9061164/S1.

103. Lai, H. *et al.* (2022) ‘Handling practice as a critical point influencing the transmission route of campylobacter throughout a commercial restaurant kitchen in China’, *Food Control*. Elsevier, 139, p. 109056. doi: 10.1016/J.FOODCONT.2022.109056.

104. Lampis, G. *et al.* (2004) ‘Culture and isolation of Campylobacter species by the Linearcount 3MA system’, *Anaerobe*. Academic Press, 10(4), pp. 213–215. doi: 10.1016/J.ANAEROBE.2004.05.004.

105. Letunic, I. and Bork, P. (2021) ‘Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation’, *Nucleic Acids Research*. Oxford Academic, 49(W1), pp. W293–W296. doi: 10.1093/NAR/GKAB301.

106. Lin, J. (2017) ‘Antimicrobial resistance: from basic science to translational innovation’, *Animal Health Research Reviews*. Cambridge University Press, 18(2), pp. 85–86. doi: 10.1017/S1466252317000172.

107. Llarena, A. K., Taboada, E. and Rossi, M. (2017) ‘Whole-Genome Sequencing in Epidemiology of Campylobacter jejuni Infections’, *Journal of Clinical Microbiology*. American Society for Microbiology1752 N St., N.W., Washington, DC, 55(5), pp. 1269–1275. doi: 10.1128/JCM.00017-17.

108. Longenberger, A. H. *et al.* (2013) ‘Campylobacter jejuni Infections Associated With Unpasteurized Milk—Multiple States, 2012’, *Clinical Infectious*

*Diseases*. Oxford Academic, 57(2), pp. 263–266. doi: 10.1093/CID/CIT231.

109. Lopes, B. S. *et al.* (2019) 'Nationwide Stepwise Emergence and Evolution of Multidrug-Resistant Campylobacter jejuni Sequence Type 5136, United Kingdom', *Emerging Infectious Diseases*. Centers for Disease Control and Prevention, 25(7), p. 1320. doi: 10.3201/EID2507.181572.

110. Lorenzo, J. M. *et al.* (2018) 'Main Groups of Microorganisms of Relevance for Food Safety and Stability: General Aspects and Overall Description', *Innovative technologies for food preservation: Inactivation of spoilage and pathogenic microorganisms*. Academic Press, pp. 53–107. doi: 10.1016/B978-0-12-811031-7.00003-0.

111. Luo, N. *et al.* (2005) 'Enhanced in vivo fitness of fluoroquinolone-resistant Campylobacter jejuni in the absence of antibiotic selection pressure', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 102(3), pp. 541–546. doi: 10.1073/pnas.0408966102.

112. Madeira, F. *et al.* (2019) 'The EMBL-EBI search and sequence analysis tools APIs in 2019.', *Nucleic Acids Research*. Oxford University Press, 47(W1), pp. W636–W641. doi: 10.1093/NAR/GKZ268.

113. Maiden, M. C. J. *et al.* (1998a) 'Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 95(6), pp. 3140–3145. doi: 10.1073/PNAS.95.6.3140.

114. Maiden, M. C. J. *et al.* (1998b) 'Multilocus sequence typing: A portable

approach to the identification of clones within populations of pathogenic microorganisms’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 95(6), pp. 3140–3145. doi: 10.1073/pnas.95.6.3140.

115. Maiden, M. C. J. (2006a) ‘Multilocus Sequence Typing of Bacteria’, *https://doi.org/10.1146/annurev.micro.59.030804.121325*. Annual Reviews, 60, pp. 561–588. doi: 10.1146/ANNUREV.MICRO.59.030804.121325.

116. Maiden, M. C. J. (2006b) ‘Multilocus Sequence Typing of Bacteria’, *Annual Review of Microbiology*. Annual Reviews, 60(1), pp. 561–588. doi: 10.1146/annurev.micro.59.030804.121325.

117. Maiden, Martin C J *et al.* (2013) ‘MLST revisited: the gene-by-gene approach to bacterial genomics.’, *Nature reviews. Microbiology*. Europe PMC Funders, 11(10), pp. 728–36. doi: 10.1038/nrmicro3093.

118. Maiden, Martin C.J. *et al.* (2013) ‘MLST revisited: The gene-by-gene approach to bacterial genomics’, *Nature Reviews Microbiology*. Nature Publishing Group, pp. 728–736. doi: 10.1038/nrmicro3093.

119. Manyi-Loh, C. *et al.* (2018) ‘Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications’, *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry*. Multidisciplinary Digital Publishing Institute (MDPI), 23(4). doi: 10.3390/MOLECULES23040795.

120. Marin, C. *et al.* (2022) ‘Rapid Oxford Nanopore Technologies MinION Sequencing Workflow for Campylobacter jejuni Identification in Broilers on Site—

A Proof-of-Concept Study', *Animals*. MDPI, 12(16), p. 2065. doi: 10.3390/ANI12162065/S1.

121. McCarthy, N. D. *et al.* (2012) 'Molecular epidemiology of human Campylobacter jejuni shows association between seasonal and international patterns of disease', *Epidemiology & Infection*. Cambridge University Press, 140(12), pp. 2247–2255. doi: 10.1017/S0950268812000192.

122. Misawa, N. *et al.* (2002) 'Isolation and characterization of Campylobacter, Helicobacter, and Anaerobiospirillum strains from a puppy with bloody diarrhea', *Veterinary Microbiology*. Elsevier, 87(4), pp. 353–364. doi: 10.1016/S0378-1135(02)00086-X.

123. Montgomery, M. P. *et al.* (2018) 'Multidrug-Resistant Campylobacter jejuni Outbreak Linked to Puppy Exposure — United States, 2016–2018', *Morbidity and Mortality Weekly Report*. Centers for Disease Control and Prevention, 67(37), p. 1032. doi: 10.15585/MMWR.MM6737A3.

124. Mouftah, S. F. *et al.* (2022) 'Local accessory gene sharing among Egyptian Campylobacter potentially promotes the spread of antimicrobial resistance', *Microbial Genomics*. Microbiology Society, 8(6), p. 000834. doi: 10.1099/MGEN.0.000834/CITE/REFWORKS.

125. Mourkas, E., Bayliss, S., *et al.* (2019) 'Comparative pangenomics of Campylobacter species', *Access Microbiology*. Microbiology Society, 1(1A), p. 810. doi: 10.1099/ACMI.AC2019.PO0520.

126. Mourkas, E., Florez-Cuadrado, D., *et al.* (2019) 'Gene pool transmission of multidrug resistance among Campylobacter from livestock, sewage and human

disease', *Environmental Microbiology*. Blackwell Publishing Ltd, 21(12), pp. 4597–4613. doi: 10.1111/1462-2920.14760/FORMAT/PDF.

127. Munk, P. *et al.* (2022) 'Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance', *Nature Communications 2022 13:1*. Nature Publishing Group, 13(1), pp. 1–16. doi: 10.1038/s41467-022-34312-7.

128. Neal-McKinney, J. M. *et al.* (2021) 'Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of Campylobacter jejuni', *Scientific Reports 2021 11:1*. Nature Publishing Group, 11(1), pp. 1–10. doi: 10.1038/s41598-021-84956-6.

129. Nennig, M. *et al.* (2021) 'Investigating Major Recurring Campylobacter jejuni Lineages in Luxembourg Using Four Core or Whole Genome Sequencing Typing Schemes', *Frontiers in Cellular and Infection Microbiology*. Frontiers Media S.A., 10, p. 608020. doi: 10.3389/FCIMB.2020.608020/BIBTEX.

130. Nielsen, L. N. *et al.* (2010a) 'MLST clustering of Campylobacter jejuni isolates from patients with gastroenteritis, reactive arthritis and Guillain–Barré syndrome', *Journal of applied microbiology*. Europe PMC Funders, 108(2), p. 591. doi: 10.1111/J.1365-2672.2009.04444.X.

131. Nielsen, L. N. *et al.* (2010b) 'MLST clustering of Campylobacter jejuni isolates from patients with gastroenteritis, reactive arthritis and Guillain–Barré syndrome', *Journal of Applied Microbiology*. Oxford Academic, 108(2), pp. 591–599. doi: 10.1111/J.1365-2672.2009.04444.X.

132. O'Neill, J. (2014) *Antimicrobial Resistance: Tackling a crisis for the health*

*and wealth of nations*. Available at: https://amr-review.org/sites/default/files/AMR Review Paper - Tackling a crisis for the health and wealth of nations_1.pdf (Accessed: 28 May 2019).

133. Ocejo, M. *et al.* (2021) 'Whole genome-based characterisation of antimicrobial resistance and genetic diversity in Campylobacter jejuni and Campylobacter coli from ruminants', *Scientific Reports*. Nature Publishing Group, 11(1), p. 8998. doi: 10.1038/S41598-021-88318-0.

134. Olkkola, S. *et al.* (2016) 'Antimicrobial Resistance and Multilocus Sequence Types of Finnish Campylobacter jejuni Isolates from Multiple Sources', *Zoonoses and Public Health*. John Wiley & Sons, Ltd, 63(1), pp. 10–19. doi: 10.1111/ZPH.12198.

135. Painset, A. *et al.* (2020) 'Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of Campylobacter jejuni and Campylobacter coli isolated from cases of diarrhoeal disease in England and Wales, 2015–16', *Journal of Antimicrobial Chemotherapy*. Oxford Academic, 75(4), pp. 883–889. doi: 10.1093/JAC/DKZ539.

136. Park, M. *et al.* (2022) 'Development of Fluoroquinolone Resistance through Antibiotic Tolerance in Campylobacter jejuni', *Microbiology Spectrum*. American Society for Microbiology, 10(5). doi: 10.1128/SPECTRUM.01667-22/SUPPL_FILE/SPECTRUM.01667-22-S0001.PDF.

137. Pascoe, B. I. *et al.* (2020) 'Genomic epidemiology of Campylobacter jejuni associated with asymptomatic pediatric infection in the Peruvian Amazon', *PLOS Neglected Tropical Diseases*. doi: 10.1371/journal.pntd.0008533.

138. Piddock, L. J. V. (2012) 'The crisis of no new antibiotics-what is the way forward?', *The Lancet Infectious Diseases*. Elsevier, pp. 249–253. doi: 10.1016/S1473-3099(11)70316-4.

139. Poropatich, K. O., Walker, C. L. F. and Black, R. E. (2010) 'Quantifying the Association between Campylobacter Infection and Guillain-Barré Syndrome: A Systematic Review', *Journal of Health, Population, and Nutrition*. BioMed Central, 28(6), p. 545. doi: 10.3329/JHPN.V28I6.6602.

140. Prestinaci, F., Pezzotti, P. and Pantosti, A. (2015) 'Antimicrobial resistance: a global multifaceted phenomenon', *Pathogens and Global Health*. Taylor & Francis, 109(7), p. 309. doi: 10.1179/2047773215Y.0000000030.

141. Ragimbeau, C. *et al.* (2008) 'Multilocus sequence typing, pulsed-field gel electrophoresis, and fla short variable region typing of clonal-complexes of Campylobacter jejuni strains of human, bovine, and poultry origins in Luxembourg', *Applied and Environmental Microbiology*. American Society for Microbiology, 74(24), pp. 7715–7722. doi: 10.1128/AEM.00865-08/SUPPL_FILE/FIGS1S2.ZIP.

142. Ramonaite, S. *et al.* (2014) 'Prevalence and genotypes of Campylobacter jejuni from urban environmental sources in comparison with clinical isolates from children', *Journal of Medical Microbiology*. Microbiology Society, 63(PART 9), pp. 1205–1213. doi: 10.1099/JMM.0.072892-0/CITE/REFWORKS.

143. Ramonaite, S. *et al.* (2017a) 'MLST genotypes of Campylobacter jejuni isolated from broiler products, dairy cattle and human campylobacteriosis cases in Lithuania', *BMC Infectious Diseases*. BioMed Central Ltd., 17(1), pp. 1–10. doi: 10.1186/S12879-017-2535-1/TABLES/3.

144. Ramonaite, S. *et al.* (2017b) ‘MLST genotypes of Campylobacter jejuni isolated from broiler products, dairy cattle and human campylobacteriosis cases in Lithuania’, *BMC Infectious Diseases*. BioMed Central Ltd., 17(1), p. 430. doi: 10.1186/s12879-017-2535-1.

145. Rawson, T. *et al.* (2020) ‘A Mathematical Modeling Approach to Uncover Factors Influencing the Spread of Campylobacter in a Flock of Broiler-Breeder Chickens’, *Frontiers in Microbiology*. Frontiers Media S.A., 11, p. 576646. doi: 10.3389/FMICB.2020.576646/BIBTEX.

146. Rawson, T., Dawkins, M. S. and Bonsall, M. B. (2019a) ‘A mathematical model of campylobacter dynamics within a broiler flock’, *Frontiers in Microbiology*. Frontiers Media S.A., 10(AUG), p. 465419. doi: 10.3389/FMICB.2019.01940/BIBTEX.

147. Rawson, T., Dawkins, M. S. and Bonsall, M. B. (2019b) ‘A Mathematical Model of Campylobacter Dynamics Within a Broiler Flock’, *Frontiers in Microbiology*. Frontiers Media S.A., 10(AUG), p. 1940. doi: 10.3389/fmicb.2019.01940.

148. Read, D. S. *et al.* (2013) ‘Evidence for phenotypic plasticity among multihost Campylobacter jejuni and C. coli lineages, obtained using ribosomal multilocus sequence typing and raman spectroscopy’, *Applied and Environmental Microbiology*. American Society for Microbiology, 79(3), pp. 965–973. doi: 10.1128/AEM.02521-12/SUPPL_FILE/ZAM999104079SO4.PDF.

149. Rello, J., Bunsow, E. and Perez, A. (2016) ‘What if there were no new antibiotics? A look at alternatives’, *Expert Review of Clinical Pharmacology*. Taylor

& Francis, 9(12), pp. 1547–1555. doi: 10.1080/17512433.2016.1241141.

150. Richards, A. (2015a) ‘University of Oxford Advanced Research Computing’, *Zenodo*.

151. Richards, A. (2015b) ‘University of Oxford Advanced Research Computing’. doi: 10.5281/ZENODO.22558.

152. Van Rossum, G. , Drake, F. . (2009) ‘Python 3 Reference Manual’, *CA: CreateSpace.*

153. Saif, N. A. *et al.* (2022) ‘A pilot study revealing host-associated genetic signatures for source attribution of sporadic Campylobacter jejuni infection in Egypt’, *Transboundary and Emerging Diseases*. John Wiley & Sons, Ltd, 69(4), pp. 1847–1861. doi: 10.1111/TBED.14165.

154. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) ‘DNA sequencing with chain-terminating inhibitors’, *Proceedings of the National Academy of Sciences*. Proceedings of the National Academy of Sciences, 74(12), pp. 5463–5467. doi: 10.1073/PNAS.74.12.5463.

155. Santos-Ferreira, N. *et al.* (2021) ‘Cross-contamination of lettuce with Campylobacter spp. via cooking salt during handling raw poultry’, *PLOS ONE*. Public Library of Science, 16(5), p. e0250980. doi: 10.1371/JOURNAL.PONE.0250980.

156. Sheppard, S. K. *et al.* (2009) ‘Campylobacter genotyping to determine the source of human infection’, *Clinical Infectious Diseases*. Oxford Academic, 48(8), pp. 1072–1078. doi: 10.1086/597402/2/48-8-1072-FIG003.GIF.

157. Sheppard, S. K. *et al.* (2010) ‘Host association of campylobacter genotypes

transcends geographic variations', *Applied and Environmental Microbiology*, 76(15), pp. 5269–5277. doi: 10.1128/AEM.00124-10/FORMAT/EPUB.

158. Sheppard, S. K. *et al.* (2011) 'Introgression in the genus Campylobacter: generation and spread of mosaic alleles', *Microbiology*. Microbiology Society, 157(Pt 4), p. 1066. doi: 10.1099/MIC.0.045153-0.

159. Sheppard, S. K. *et al.* (2013) 'Progressive genome-wide introgression in agricultural Campylobacter coli', *Molecular Ecology*. John Wiley & Sons, Ltd, 22(4), pp. 1051–1064. doi: 10.1111/MEC.12162.

160. Sheppard, S. K. *et al.* (2014) 'Cryptic ecology among host generalist Campylobacter jejuni in domestic animals', *Molecular Ecology*. Blackwell Publishing Ltd, 23(10), pp. 2442–2451. doi: 10.1111/MEC.12742/FORMAT/PDF.

161. Skarp-de Haan, C. P. A. *et al.* (2014) 'Comparative genomics of unintrogressed Campylobacter coli clades 2 and 3', *BMC Genomics*. BioMed Central, 15(1), pp. 1–14. doi: 10.1186/1471-2164-15-129/FIGURES/6.

162. Skarp, C. P. A. *et al.* (2017) 'Accessory genetic content in Campylobacter jejuni ST21CC isolates from feces and blood', *International Journal of Medical Microbiology*. Urban & Fischer, 307(4–5), pp. 233–240. doi: 10.1016/J.IJMM.2017.04.001.

163. Sproston, E. L., Wimalarathna, H. M. . and Sheppard, S. K. (2018) 'Trends in fluoroquinolone resistance in campylobacter', *Microbial Genomics*, 4(8). doi: 10.1099/mgen.0.000198.

164. 'Tableau (version. 9.1)' (2016) *Journal of the Medical Library Association : JMLA*. Medical Library Association, 104(2), p. 182. doi: 10.3163/1536-

5050.104.2.022.

165. Tam, C. C. *et al.* (2012) 'Longitudinal study of infectious intestinal disease in the UK (IID2 study): Incidence in the community and presenting to general practice', *Gut*. BMJ Publishing Group, 61(1), pp. 69–77. doi: 10.1136/gut.2011.238386.

166. Team, R. C. (2020) 'R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.' Available at: https://www.r-project.org/ (Accessed: 19 April 2022).

167. Tegner, C. *et al.* (2019) 'Detection, Identification, and Antimicrobial Susceptibility of Campylobacter spp. and Salmonella spp. from Free-ranging Nonhuman Primates in Sri Lanka', *Journal of Wildlife Diseases*. Allen Press, 55(4), pp. 879–884. doi: 10.7589/2018-08-199.

168. Theobald Smith, B. and Taylor, M. S. (1918) 'SOME MORPHOLOGICAL AND BIOLOGICAL CHARACTERS OF THE SPIRILLA (VIBRIO FETUS, N. SP.) ASSOCIATED WITH DISEASE OF THE FETAl, MEMBRANES IN CATTLE', *J. Exp. Med*, p. 701. Available at: http://rupress.org/jem/article-pdf/30/4/299/1175399/299.pdf (Accessed: 12 April 2023).

169. Thépault, A. *et al.* (2017a) 'Genome-wide identification of host-segregating epidemiological markers for source attribution in Campylobacter jejuni', *Applied and Environmental Microbiology*. American Society for Microbiology, 83(7). doi: 10.1128/AEM.03085-16/SUPPL_FILE/ZAM999117733S1.PDF.

170. Thépault, A. *et al.* (2017b) 'Genome-wide identification of host-segregating epidemiological markers for source attribution in Campylobacter jejuni', *Applied*

*and Environmental Microbiology*. American Society for Microbiology, 83(7), pp. 3085–3101. doi: 10.1128/AEM.03085-16/SUPPL_FILE/ZAM999117733S1.PDF.

171. Truccollo, B. *et al.* (2021) 'Genetic characterisation of a subset of Campylobacter jejuni isolates from clinical and poultry sources in Ireland', *PLOS ONE*. Public Library of Science, 16(3), p. e0246843. doi: 10.1371/JOURNAL.PONE.0246843.

172. Urwin, R. and Maiden, M. C. J. (2003) 'Multilocus sequence typing: a tool for global epidemiology', *Trends in Microbiology*. Elsevier Current Trends, 11(10), pp. 479–487. doi: 10.1016/J.TIM.2003.08.006.

173. Veltcheva, D. *et al.* (2022) 'Emerging patterns of fluoroquinolone resistance in Campylobacter jejuni in the UK [1998-2018', *Microbial genomics*. NLM (Medline), 8(9), p. 000875. doi: 10.1099/MGEN.0.000875/CITE/REFWORKS.

174. Veltcheva, D. *et al.* (2023) 'NeighbourGroups: a machine learning classication tool that assigns microbial Multilocus genotypes to clusters'. doi: 10.21203/rs.3.rs-2666125/v1.

175. VÉRON, M. and CHATELAIN, R. (1973) 'Taxonomic Study of the Genus Campylobacter Sebald and Véron and Designation of the Neotype Strain for the Type Species, Campylobacter fetus (Smith and Taylor) Sebald and Véron', *International Journal of Systematic and Evolutionary Microbiology*. Microbiology Society, 23(2), pp. 122–134. doi: 10.1099/00207713-23-2-122.

176. Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature Methods 2020 17:3*. Nature Publishing Group, 17(3), pp. 261–272. doi: 10.1038/s41592-019-0686-2.

177. van Vliet, A. H. M. *et al.* (2022) ‘Genomic Screening of Antimicrobial Resistance Markers in UK and US Campylobacter Isolates Highlights Stability of Resistance over an 18-Year Period’, *Antimicrobial agents and chemotherapy*. Antimicrob Agents Chemother, 66(5). doi: 10.1128/AAC.01687-21.

178. Wang, X. *et al.* (2022) ‘Correlation between Multilocus Sequence Typing and Antimicrobial resistance, Virulence Potential of Campylobacter jejuni Isolates from Poultry Meat’, *Foods 2022, Vol. 11, Page 1768*. Multidisciplinary Digital Publishing Institute, 11(12), p. 1768. doi: 10.3390/FOODS11121768.

179. Wareing, D. R. A. *et al.* (2003) ‘Reference isolates for the clonal-complexes of Campylobacter jejuni’, *Letters in Applied Microbiology*. John Wiley & Sons, Ltd, 36(2), pp. 106–110. doi: 10.1046/j.1472-765X.2003.01270.x.

180. Wei, B., Kang, M. and Jang, H. K. (2019) ‘Genetic characterization and epidemiological implications of Campylobacter isolates from wild birds in South Korea’, *Transboundary and Emerging Diseases*. John Wiley & Sons, Ltd, 66(1), pp. 56–65. doi: 10.1111/TBED.12931.

181. Whelan, M. V. X. *et al.* (2019) ‘Acquisition of fluoroquinolone resistance leads to increased biofilm formation and pathogenicity in Campylobacter jejuni’, *Scientific Reports*. Nature Research, 9(1), pp. 1–13. doi: 10.1038/s41598-019-54620-1.

182. Wieczorek, K. and Osek, J. (2013) ‘Antimicrobial resistance mechanisms among campylobacter’, *BioMed Research International*. Hindawi, 2013, p. 340605. doi: 10.1155/2013/340605.

183. Williams, L. K., Fonseca, B. B. and Humphrey, T. J. (2016a) ‘Campylobacter

jejuni in poultry: A commensal or a pathogen?', *Campylobacter Spp. and Related Organisms in Poultry: Pathogen-Host Interactions, Diagnosis and Epidemiology*. Springer International Publishing, pp. 75–87. doi: 10.1007/978-3-319-29907-5_5/FIGURES/2.

184. Williams, L. K., Fonseca, B. B. and Humphrey, T. J. (2016b) 'Campylobacter jejuni in poultry: A commensal or a pathogen?', in *Campylobacter Spp. and Related Organisms in Poultry: Pathogen-Host Interactions, Diagnosis and Epidemiology*. Springer International Publishing, pp. 75–87. doi: 10.1007/978-3-319-29907-5_5.

185. Wimalarathna, H. M. *et al.* (2013) 'Widespread acquisition of antimicrobial resistance among Campylobacter isolates from UK retail poultry and evidence for clonal expansion of resistant lineages', *BMC Microbiology*. BioMed Central, 13(1), pp. 1–9. doi: 10.1186/1471-2180-13-160/FIGURES/3.

186. Woese, C. R., Kandler, O. and Wheelis, M. L. (1990) 'Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 87(12), p. 4576. doi: 10.1073/PNAS.87.12.4576.

187. Woodcock, D. J. *et al.* (2017) 'Genomic plasticity and rapid host switching can promote the evolution of generalism: a case study in the zoonotic pathogen Campylobacter', *Scientific Reports 2017 7:1*. Nature Publishing Group, 7(1), pp. 1–13. doi: 10.1038/s41598-017-09483-9.

188. Yabe, S. *et al.* (2010) 'Molecular typing of Campylobacter jejuni and C. coli from chickens and patients with gastritis or Guillain-Barré syndrome based on

multilocus sequence types and pulsed-field gel electrophoresis patterns', *Microbiology and Immunology*. John Wiley & Sons, Ltd, 54(6), pp. 362–367. doi: 10.1111/J.1348-0421.2010.00222.X.

189. Yahara, K., Méric, G., Taylor, A. J., de Vries, Stefan P.W., *et al.* (2017) 'Genome-wide association of functional traits linked with Campylobacter jejuni survival from farm to fork', *Environmental Microbiology*. John Wiley & Sons, Ltd, 19(1), pp. 361–380. doi: 10.1111/1462-2920.13628.

190. Yahara, K., Méric, G., Taylor, A. J., de Vries, Stefan P. W., *et al.* (2017) 'Genome-wide association of functional traits linked with <scp> *C* </scp> *ampylobacter jejuni* survival from farm to fork', *Environmental Microbiology*. Blackwell Publishing Ltd, 19(1), pp. 361–380. doi: 10.1111/1462-2920.13628.

191. Yamada, K. *et al.* (2019) 'Long-term observation of antimicrobial susceptibility and molecular characterisation of Campylobacter jejuni isolated in a Japanese general hospital 2000–2017', *Journal of Global Antimicrobial Resistance*. Elsevier, 18, pp. 59–63. doi: 10.1016/J.JGAR.2019.02.001.

192. Yang, Y. *et al.* (2019) 'A Historical Review on Antimicrobial resistance of Foodborne Campylobacter', *Frontiers in Microbiology*. Frontiers, 10, p. 1509. doi: 10.3389/fmicb.2019.01509.

193. Yeh, H. Y. and Awad, A. (2020) 'Genotyping of Campylobacter jejuni Isolates from Poultry by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)', *Current Microbiology*. Springer, 77(8), pp. 1647–1652. doi: 10.1007/S00284-020-01965-W/FIGURES/2.

194. Yu, H. *et al.* (2020) 'Epidemiological and Genomic Characterization of

Campylobacter jejuni Isolates from a Foodborne Outbreak at Hangzhou, China', *International Journal of Molecular Sciences 2020, Vol. 21, Page 3001*. Multidisciplinary Digital Publishing Institute, 21(8), p. 3001. doi: 10.3390/IJMS21083001.

195. Yuki, N. and Hartung, H.-P. (2012) 'Guillain–Barré Syndrome', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 366(24), pp. 2294–2304. doi: 10.1056/NEJMRA1114525/SUPPL_FILE/NEJMRA1114525_DISCLOSURES.PDF.

196. Zawack, K. *et al.* (2016) 'Monitoring Antimicrobial Resistance in the Food Supply Chain and Its Implications for FDA Policy Initiatives', *Antimicrobial agents and chemotherapy*. Antimicrob Agents Chemother, 60(9), pp. 5302–5311. doi: 10.1128/AAC.00688-16.

197. Zhang, A. *et al.* (2016) 'Molecular subtyping and erythromycin resistance of Campylobacter in China', *Journal of Applied Microbiology*. Oxford Academic, 121(1), pp. 287–293. doi: 10.1111/JAM.13135.

198. Zhang, P. *et al.* (2020) 'Multilocus Sequence Types and Antimicrobial Resistance of Campylobacter jejuni and C. coli Isolates of Human Patients From Beijing, China, 2017–2018', *Frontiers in Microbiology*. Frontiers Media S.A., 11, p. 554784. doi: 10.3389/FMICB.2020.554784/BIBTEX.

199. Zhang, P. *et al.* (2022) 'Genomic Insights into the Increased Occurrence of Campylobacteriosis Caused by Antimicrobial-Resistant Campylobacter coli', *mBio*. American Society for Microbiology, 13(6). doi: 10.1128/MBIO.02835-

22/SUPPL_FILE/MBIO.02835-22-S0005.PDF.

200. Zhou, Z. *et al.* (2018) ‘GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens.’, *Genome research*. Cold Spring Harbor Laboratory Press, 28(9), pp. 1395–1404. doi: 10.1101/gr.232397.117.