Genomics and epidemiology of *Streptococcus pyogenes* and *Streptococcus canis* infections in Scotland

UNIVERSITY
of
GLASGOW

Davide Pagnossin, DVM

School of Biodiversity, One Health & Veterinary Medicine

College of Medical, Veterinary & Life Sciences

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

October 2022

Abstract

Streptococcus pyogenes, also known as Group A *Streptococcus* (GAS), is a strict human pathogen associated with a high burden of disease and millions of deaths per year worldwide. Although asymptomatic carriage and superficial infections are the most common outcomes of GAS colonisation of the human body, invasive infections and post-infectious complications are severe and not rare manifestations of GAS disease. *Streptococcus canis*, conversely, is an opportunistic pathogen that can colonise and cause disease in several mammalian species, particularly dogs and cats. Although rare, *S. canis* infections in humans can lead to death and are likely associated with zoonotic transfer from companion animals. Both *S. pyogenes* and *S. canis* are classified as pyogenic streptococci based on the clinical manifestations of infection and phylogenetic relatedness. The overarching aim of this work was to apply a multi-disciplinary approach to characterise the epidemiology of *S. pyogenes* and *S. canis* infections in humans and animals in Scotland. This was achieved through specific objectives: i) Characterisation of the main epidemiological features of invasive GAS (iGAS) infections in Scotland from 2014 to 2021. Findings confirmed that the annual incidence of iGAS disease in Scotland was comparable to the one reported in recent years in high-income countries. A seasonal pattern characterised by an increased incidence in Winter and Spring compared with Summer and Autumn was highlighted. Neonates, the elderly and people in their 30s were the age groups with the highest incidence of iGAS infections. Many *S. pyogenes* genotypes, defined as *emm* types, were implicated in invasive disease in recent years; some appeared to be consistently common throughout the study period, others were associated with temporary upsurges of disease or sporadic cases. The transition from the first to the second year of COVID-19 pandemic was characterised by a higher-than-expected *emm* type turnaround, suggesting that the restrictions in place had a repercussion on the circulation of specific strains of GAS in Scotland. ii) Investigation of an upsurge of iGAS disease

associated with the genotype *emm*5.23 using whole genome sequencing (WGS) and transcriptomic analyses. Results of phylogenetic, virulence and AMR analyses suggested that acquisition of determinants responsible for a high-virulence phenotype in the *emm*5.23 population circulating in Scotland is unlikely. However, epidemiological connection of the cases appears possible making the *emm*5.23 infection upsurge a potentially undetected outbreak. iii) Characterisation of antimicrobial resistance (AMR), virulence characteristics and population structure of *S. canis* isolates from different species and geographic locations, including Scotland, using WGS analysis. Findings indicated that around 20% of *S. canis* strains (9/39) carried known AMR genes and were resistant to at least one antibiotic class, particularly tetracyclines, macrolides and lincosamides. The majority of virulence genes (17/19) detected in *S. canis* isolates were homologous to *S. pyogenes* genes, suggesting these two bacterial species might share similar virulence mechanisms. Genomic analyses of *S. canis* population structure did not show signs of host adaptation, indicating similar strains circulate and cause disease in different host species. The two genotyping systems currently used to classify *S. canis* strains, multi-locus sequence typing (MLST) and an *S. canis* M-like (SCM) typing scheme, have comparable accuracy in assigning genotypes but lack the discriminatory power of WGS to aid fine resolution such as that needed for outbreak settings. iv) Exploring and optimising the use of data visualisation to communicate the epidemiology of iGAS disease to a cohort of public health and laboratory workers in Scotland. A targeted survey with proposed multiple visualisations of reported results allowed to identify guiding principles that can facilitate the generation of data visualisations to communicate the epidemiology of iGAS disease to a specialised audience.

Contents

List of Tables

List of Figures

Acknowledgements

Scientific Supervision

I am immensely grateful to my PhD supervisors for believing in me since the very beginning and for giving me the opportunity of a lifetime. To Katarína, thank you for guiding me with kindness through this difficult journey. Thank you for listening and making me feel comfortable to open up to you. To Andrew, thank you for always making time for me and for pushing me to overcome my limits. To Willie, thank you for helping me anytime I needed it and for always giving me the wisest advice. Working with you has been a pleasure and I will surely miss being part of such a wholesome team. You have been an inspiration to me, not only as researchers but also as human beings.

Scientific Collaborations

I would like to sincerely thank all the collaborators of this project, in particular Roisin Ure for producing the majority of the WGS analysed in this work. My gratitude goes also to Professor Stephen Beres for giving me invaluable bioinformatic advice, to Dr. Juliana Coelho for providing bacterial isolates and WGS that allowed to expand our database, to Manuel Fuentes for collecting *Streptococcus canis* strains and to Professor Roderic Page for sharing his expertise on phylogenetic analysis. A special thanks to Dr. Helen Purchase for dedicating a great deal of her time to help me conceptualise and develop the chapter on data visualisation in public health.

Financial Support

I would like to acknowledge the University of Glasgow and in particular the School of Biodiversity, One Health & Veterinary Medicine for funding my PhD project with the James Herriot Scholarship Fund, and for awarding me travel and training grants throughout my PhD journey.

Family

Grazie alla mia famiglia per avermi sempre supportato con amore. Mamma e papá, grazie per avermi incoraggiato ad essere me stesso e per avermi dato la certezza che, nonostante tutto, ci sareste sempre stati per me. Grazie per avermi ispirato ad infondere amore e passione in tutto ció che faccio e per avermi mostrato il valore della dedizione, pazienza e tenacia. Ad Alice, grazie per essere al mio fianco quando ne ho bisogno, grazie per la tua bontá e per essere un'amica oltre che una sorella. Grazie a tutto il resto della mia famiglia per essermi sempre vicini nonostante la distanza fisica.

Life partner

Jamie, thank you for being the best partner I could ask for, I cannot imagine how I could have completed this work without your presence in my life. Thank you for making everything better.

Friends

Chiara, you have been such a loyal, loving and inspiring friend since the beginning of my PhD. You have been the one I could always count on, both for the hard and the good times. You helped me to be a better scientist and friend, and I will always be grateful for that. Umu, thank you for being my closest friend in one of the hardest moments of my life. I will always cherish the memories of the time spent with you. Thank you Zofia, Kate and Sujan, for being lovely flatmates and friends. Thanks also to Mousa, Vyome, Hua, Joel, Pippa, Rodrigo and all the other Glasgow friends that have come and left. You have all made my journey happier. Thanks also to my friends from afar, in particular Enrico, Francesco, Giulia, Storm, Eugenio,

Giorgia, Anna, Elisa and Riccardo. Even if I couldn't share my experience in Glasgow with you, I've always kept you in my heart.

General Assistance

Thanks to the amazing OHRBID laboratory group for being such a positive and supportive team to work with.

Everyone else

Thanks to Scotland, which welcomed me and became my second home. Thanks to Marnie, you little rascal drive me crazy but I still love you. Thanks to the part of myself that did not let me drop out when the whole world around me was pushing me to do it. You were right, it has been worth the effort.

Author's declaration

This thesis and the work presented in it was generated by me as the result of my own original research. It is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University and no part has already been, or is concurrently being, submitted for any degree, diploma, or other qualification. It does not exceed 80,000 words, excluding references, bibliography, table of contents and appendices.

Chapter 2: Descriptive epidemiology of invasive Group A *Streptococcus* infections in Scotland from 2014 to 2021. Initial concept developed by D. Pagnossin and A. Smith. Epidemiological data collection was carried out by SMiRL. Data cleaning and analysis was conducted by D. Pagnossin with advice from W. Weir. The chapter was written by D. Pagnossin, and revised by K. Oravcová, A. Smith and W. Weir.

Chapter 3: Genomic characterisation of *Streptococcus pyogenes emm*5.23, a recently emerged strain causing invasive disease in Scotland. Initial conceptualisation by A. Smith and K. Oravcová, with development by D. Pagnossin. Bacterial isolates were provided by SMiRL and Illumina sequences produced by R. Ure (SMiRL). A subset of the isolates and sequences analysed were provided by J. Coelho (UKHSA). Bioinformatic data analysis was carried out by D. Pagnossin with advice from S. B. Beres (Houston Methodist, Texas, USA). The generation of Oxford Nanopore sequences was conducted by D. Pagnossin with advice from C. Crestani (Institute Pasteur, France). Antimicrobial susceptibility data was produced by D. Pagnossin at SMiRL, with advice from S. Rankin and M. Coyne (SMiRL). RNA extraction and transcriptomic bioinformatic analysis was carried out by D. Pagnossin with advice from K. Oravcová. The chapter was written by D. Pagnossin and revised by K. Oravcová, A. Smith and W. Weir.

Chapter 4: Genomic epidemiology, virulence and antimicrobial resistance of the multi-host pathogen *Streptococcus canis*. Initial concept developed by D. Pagnossin. Bacterial isolates were collected by M. Fuentes (University of Glasgow). A subset of the isolates was provided by J. Coelho (UKHSA) and A. Smith. Illumina sequences were produced by R. Ure (SMiRL). Bioinformatic data analysis was conducted by D. Pagnossin with advice from R. Page (University of Glasgow). Antimicrobial susceptibility data was produced by D. Pagnossin. The chapter was written by D. Pagnossin and revised by K. Oravcová, A. Smith and W. Weir.

Chapter 5: Data visualization in the public health sector. Initial conceptualisation of the study by A. Smith and H. Purchase (University of Glasgow), with development by D. Pagnossin. Data were collected and analysed by D. Pagnossin with advice from H. Purchase. The chapter was written by D. Pagnossin and revised by K. Oravcová, A. Smith and W. Weir.

Abbreviations

ABCS	Active Bacterial Core Surveillance
ADS	Arginine Deiminase System
AG	Antigen
AMP	Antimicrobial Peptide
AMR	Antimicrobial Resistance
AMU	Antimicrobial Use
APHA	Animal and Plant Health Agency
APSGN	Acute Post-Streptococcal Glomerulonephritis
ARF	Acute Rheumatic Fever
AST	Antimicrobial Susceptibility Testing
AU	Approximately Unbiased
AW	Adjusted Wallace
BC	Bray Curtis
BLAST	Basic Local Alignment Search Tool
BMD	Broth Microdilution
CARD	Comprehensive Antibiotic Resistance Database
CDC	Centers for Disease Control and Prevention
CDS	Coding Sequences
CI	Confidence Interval
CIA	Cell Invasion Ability
CGS	Core Genome SNP
EARS-Net	European Antimicrobial Resistance Surveillance Network
ECDC	European Centre for Disease Prevention and Control
EG	English Group
ENT	Ear-Nose-Throat

FBP	Fibronectin-binding Protein
GAS	Group A *Streptococcus*
GG&C	Greater Glasgow and Clyde
HPS	Health Protection Scotland
ICE	Integrative and Conjugative Element
IDE	Ig-degrading Enzyme
IG	Immunoglobulin
iGAS	Invasive Group A *Streptococcus*
IQR	Interquartile Range
MGE	Mobile Genetic Element
MIC	Minimum Inhibitory Concentration
MLSB	Macrolide, Lincosamide and Streptogramin B
MLST	Multi-locus Sequence Typing
NCBI	National Centre for Biotechnology Information
NET	Neutrophil Extracellular Trap
NICE	National Institute for Health and Care Excellence
PCA	Principal Component Analysis
PFGE	Pulse Field Gel Electrophoresis
PHS	Public Health Scotland
QRDR	Quinolone Resistance-determining Regions
RAPD	Random Amplified Polymorphic DNA
RALP	RofA-like Protein
RHD	Rheumatic Heart Disease
SAVSNET	Small Animal Veterinary Surveillance Network
SCM	*Streptococcus canis* M-like Protein
SDA	Streptococcal DNase
SG	Scottish Group
SIC	Streptococcal Inhibitor of Complement
SID	Simpson's Index of Diversity
SKA	Streptokinase
SLA	Secreted Phospholipase A
SLO	Streptolysin-O
SLS	Streptolysin-S

SMiRL	Scottish Microbiology Reference Laboratory
SNP	Single Nucleotide Polymorphism
SOF	Serum Opacity Factor
SONAAR	Scottish One Health Antimicrobial Use and Antimicrobial Resistance
SPE	Streptococcal Pyrogenic Exotoxin
SPNA	Streptococcal Nuclease A
spyCEP	*S. pyogenes* cell envelope protease
STSS	Streptococcal Toxic Shock Syndrome
TCS	Two-Component System
THB	Todd-Hewitt Broth
UKHSA	UK Health Security Agency
VFDB	Virulence Factor Database
WGS	Whole Genome Sequence
WHO	World Health Organisation

Chapter 1

Introduction

1.1 Streptococcal taxonomy

Bacteria of the genus *Streptococcus* were first identified by Louis Pasteur in 1879 (Efstratiou, 2000) who described them as "chains of beads" due to their spherical shape and spatial arrangement. The genus *Streptococcus* belongs to the family *Streptococcaceae*, within the order *Lactobacillales* of the class *Bacilli*, phylum *Firmicutes* (Vos et al., 2011). All members of the *Streptococcaceae* are Gram positive, catalase-negative and facultative anaerobic, with some species requiring the presence of 5% CO_2 to grow (Vos et al., 2011). Some streptococci cause haemolysis when growing on blood-rich agar and are thus referred to as haemolytic streptococci. These can be further sub-divided into those that cause partial (alpha(α)) or complete (beta(β)) haemolysis (Vos et al., 2011).

Rebecca Lancefield was one of the first to study the cell-wall polysaccharide composition of streptococcal species (Lancefield, 1933). Based on her findings, a serological grouping scheme for species of the genus was proposed and is still in use (Lancefield, 1933; Schleifer and Kilpper-Bälz, 1987). According to the Lancefield classification system, streptococcal strains are identified using alphabetical letters based on their cell-wall polysaccharide antigen composition (Lancefield, 1933). The Lancefield grouping scheme offers a good representation of all β haemolytic streptococci but it is not accurate in representing non-haemolytic and α-haemolytic species (Vos et al., 2011).

1.2 *Streptococcus pyogenes*

Streptococcus pyogenes, also referred to as Group A *Streptococcus* (GAS) for being the only Lancefield Group A species, is an important human pathogen. *S. pyogenes* together with another eleven phylogenetically related species form the group of pyogenic streptococci (Vos et al., 2011). While members of this group typically display multi-host tropism, GAS has evolved and adapted to a single host (Lefébure et al., 2012). *S. pyogenes* is isolated almost exclusively from humans and is associated with a spectrum of presentations from mild localised disease to serious invasive infection (Walker et al., 2014).

1.2.1 *Streptococcus pyogenes* characterisation

Strains of *S. pyogenes* were historically typed using serological methods (Efstratiou, 2000). This approach, however, can be challenging due to difficulties in producing and maintaining the necessary typing reagents such as strain-specific antisera. In some instances this approach may be impossible due to a lack of target antigens on the *S. pyogenes* strain being typed. For these reasons, non-serological methods based on molecular technologies have been developed to ensure a more reliable approach to GAS classification (Beall et al., 1996; Hookey et al., 1996; Kaufhold et al., 1994; Stanley et al., 1996).

Serological methods

GAS serotyping systems were introduced by Griffith (Griffith, 1934) in 1934 and Lancefield (Lancefield, 1962) in 1962 and are based on the immune-recognition of specific forms of the cell wall proteins T, R and M together with Serum Opacity Factor (SOF). The M protein is a highly specific antigenic marker and an important virulence factor, making it the basis of the most commonly applied typing scheme (Johnson et al., 1996). The identification of M protein type is most often achieved using an immunoprecipitation test, but a more laborious indirect bactericidal test is also available (Johnson et al., 1996). At the beginning of the 2000s, more than 80 different M serotypes had been identified (Efstratiou, 2000; Tyrrell et al., 2002).

Non-serological methods

Molecular technologies for GAS typing have been developed which obviate many of the problems associated with serotyping. These approaches made it possible to characterise serologically 'untypeable' strains and were found to have a higher discriminatory power than traditional serotyping (Johnson et al., 1996).

The molecular typing techniques which have been employed to characterise isolates of *S. pyogenes* are numerous and diverse. These comprise: phage typing (Skjold and Wannamaker, 1976; Skjold et al., 1983); bacteriocin typing (Tagg and Bannister, 1979; Tagg and Martin, 1984); restriction fragment length polymorphism (Cleary et al., 1988, 1992); pulse field gel electrophoresis (PFGE) (Single and Martin, 1992); ribotyping (Bingen et al., 1992a,b); use of oligonucleotide probes targeting the N-terminal region of specific M-protein genes (Kaufhold et al., 1992); multilocus enzyme electrophoresis of cell lysates (Selander et al., 1986; Musser et al., 1992); pyrolytic mass spectrometry (Magee et al., 1989, 1991); random amplification of polymorphic DNA (RAPD) (Welsh and McClelland, 1990; Carapetis et al., 1995); *vir* regulon typing (Gardiner et al., 1995); fluorescent amplified fragment length polymorphisms (Desai et al., 1999); multi-locus sequence typing (MLST) (Enright et al., 2001); *tee* gene sequence typing (Falugi et al., 2008) and *emm* gene sequence typing (Beall et al., 1996; Facklam et al., 1999).

Currently the standard methodology for *S. pyogenes* characterisation is *emm* gene sequence typing (Spellerberg and Brandt, 2016) and this has been adopted for routine use by laboratories around the world. This approach, developed by the Centers for Disease Control and Prevention (CDC), differentiates GAS strains based on the amplification and subsequent nucleotide sequencing of a variable region of the *emm* gene, which encodes the M surface protein (Facklam et al., 1999). An *emm* type is defined as a strain having less than 92% sequence identity to all the other GAS strains over the first 90 nucleotides of coding sequence of the *emm* gene (CDC, 2018). To date, more than 200 GAS *emm* types have been identified (CDC, 2018).

1.2.2 *Streptococcus pyogenes* clinical relevance

Streptococcus pyogenes is a human pathogen that can cause a wide range of diseases (Walker et al., 2014). The most common clinical manifestations of GAS infection are pharyngitis and impetigo, a form of pyoderma. These superficial forms of disease are limited to the skin and mucosal membranes and are generally mild and self limiting (Cunningham, 2000). Occasionally, pharyngitis can be accompanied by another superficial skin disease known as scarlet fever. *S. pyogenes* can also cause invasive and serious forms of disease, which are a consequence of bacterial invasion of normally sterile anatomical sites. Invasive forms of GAS disease can manifest as cellulitis, necrotising fasciitis, bacteraemia and Streptococcal toxic shock syndrome (Walker et al., 2014). *S. pyogenes* infections, both superficial and invasive, can be followed by serious post-infectious complications such as acute rheumatic fever (ARF) and acute post-streptococcal glomerulonephritis (APSGN) (Walker et al., 2014).

Reservoirs of infection and disease transmission

Streptococcus pyogenes is traditionally considered a non-environmental pathogen, although biofilm formation that allows survival in the environment for up to four months has been demonstrated (Marks et al., 2014). The main reservoir of *S. pyogenes* appears to be the human body, in particular the skin and nasopharyngeal mucosa (Bessen, 2009). The colonisation of these sites by *S. pyogenes* generally causes infection and clinical disease, but occasionally it can result in a state of asymptomatic carriage (Kaplan, 1980). Although it tends to localise extracellularly, *S. pyogenes* can enter and survive within cells of the host respiratory tract (Österlund et al., 1997; Cleary et al., 1998), leading to prolonged persistence and disease recurrence (Österlund et al., 1997). Given the restricted ecological niche of *S. pyogenes*, the principal transmission routes for GAS infection are thought to be airborne transmission through respiratory droplets and direct human-to-human contact (Bessen, 2009). These transmission routes have been implicated in many documented cases of GAS disease, for example an *emm*5 outbreak among workers of a digital factory in China (Chen et al., 2017), an *emm*81 outbreak involving Israeli soldiers and an *emm*28.4 outbreak of puerperal sepsis in an Australian hospital (Wasserzug et al., 2009; Ben Zakour et al., 2012). As previously mentioned, environmental contamination and transmission of GAS strains can also

occur, as evidenced by a hospital-acquired *emm*1 outbreak due to contamination of patient curtains (Mahida et al., 2014). Contamination of non-environmental fomites, such as drug paraphernalia, has been described as a likely transmission route for *S. pyogenes* due to the high incidence of GAS disease in intravenous drug users, the frequent isolation of GAS from injection site abscesses and the reports of infection outbreaks in people who inject drugs (Curtis et al., 2007; Lamagni et al., 2008b; Bundle et al., 2017). Foodborne transmission of GAS strains has also been speculated, although evidence to support this hypothesis remains limited (Avire et al., 2021). One outbreak investigation conducted in Minnesota in 2012 found that contaminated pasta at a high school event was the most likely source of a GAS outbreak (Kemble et al., 2013). Interestingly, consumption of cold pasta was also associated with acute GAS pharyngitis during a large-scale outbreak in Denmark (Falkenhorst et al., 2008).

Clinical manifestations

Superficial GAS infections can manifest as pharyngitis, scarlet fever or impetigo. GAS pharyngitis, also referred to as "strep throat", is the most common bacterial cause of pharyngitis in the world, with more than 600 million cases each year (Carapetis et al., 2005). Strep throat is generally a mild but somewhat painful condition. Common symptoms include: sore throat with quick onset, pain when swallowing, fever, petechiae on the palate, cervical lymphadenopathy, headache, nausea and vomiting (Wessels, 2011). GAS pharyngitis is occasionally accompanied by a skin disease known as scarlet fever. Scarlet fever, also known as scarlatina, appears to be caused by streptococcal pyrogenic exotoxins (SPE) (Wessels, 2016). Some of the streptococcal pyrogenic exotoxins commonly linked to scarlet fever are encoded by prophages and therefore not synthesised by all GAS strains, partially explaining why scarlatina does not always complicate strep throat (Weeks and Ferretti, 1984; Bohach et al., 1990). Scarlet fever typically manifests as an erythematous rash accompanied by swollen tongue ("strawberry tongue") and pyrexia (`https://www.gov.uk/government/publications/scarlet-fever-symptoms-diagnosis-treatment`). Streptococcal impetigo, or pyoderma, is a highly contagious skin infection which principally affects children living in tropical countries (Walker et al., 2014). The disease typically presents with the formation of cutaneous vesicles that develop into pustules, which finally burst and

are replaced by thick yellow crusts (Bisno and Stevens, 1996).

Streptococcal cellulitis is an inflammation of the subcutaneous tissue following skin infection by GAS. The condition is painful and characterised by swelling and erythema of the affected area (Bisno and Stevens, 1996). Cellulitis is one of the most common forms of streptococcal invasive disease, representing 20-40% of all invasive cases (Walker et al., 2014). Necrotising fasciitis, formerly known as streptococcal gangrene, is a severe infection followed by necrosis of deep cutaneous tissues and muscles. The occurrence of GAS infection in deep tissues can be the consequence of a blunt traumatic event, a wound or a surgical procedure, and it is facilitated by concomitant diseases and immune system deficiency (Stevens, 1992; Hamilton et al., 2008). In its early stage, necrotising fasciitis can be mistaken for cellulitis due to similar manifestations such as swelling, redness, tenderness and cutaneous pain. As the disease progresses, however, signs of gangrene become apparent (Stevens, 1992). The erythema darkens, eventually turning black, and yellow fluid-filled bullae develop on the affected area, reflecting the progressive death of the infected tissues. Together with the signs described, the disease manifests systemically with fever and mental dullness (Stevens, 1992). Necrotising fasciitis is associated with a high mortality rate, estimated to be around 30% (Lamagni et al., 2008a).

The presence of *S. pyogenes* in the bloodstream is known as GAS bacteremia. This condition can result from a penetrating injury, pre-existing infection or childbirth, the latter case being known as puerperal sepsis. Fever, nausea and vomiting are common symptoms associated with GAS bacteremia (Walker et al., 2014).

Streptococcal toxic shock syndrome (STSS) is another severe form of invasive disease. It can originate from a deep-seated infection such as cellulitis or necrotising fasciitis and it is caused by the production of streptococcal toxins that act as superantigens (Lappin and Ferguson, 2009). STSS is not necessarily associated with bacteremia, although occasionally both conditions occur at the same time (Stevens, 2002). The syndrome is characterised by nausea, vomiting, rapid-onset hypotension and multi-organ failure (Walker et al., 2014; Stevens, 2002).

S. pyogenes infections, both superficial and invasive, may be followed by post-infectious complications in the form of ARF and APSGN. ARF is a delayed complication of untreated GAS pharingitis. Although the pathogenesis of this condition remains unclear, ARF is thought to be caused by an autoimmune response of genetically susceptible hosts (Karthikeyan and Guilherme, 2018). Clinically, ARF manifests as polyarthritis, chorea and valvular disease. The latter can occasionally progress to chronic rheumatic heart disease (RHD) (Karthikeyan and Guilherme, 2018). APSGN is an immune-mediated complication of GAS infections that targets the kidneys. The condition can follow both skin and throat infections (Cunningham, 2000). Oedema, hypertension, haematuria and urinary sediment abnormalities are typical symptoms associated with this syndrome (Cunningham, 2000; Walker et al., 2014).

1.2.3 *Streptococcus pyogenes* virulence mechanisms

During the course of its evolution, *S. pyogenes* has acquired numerous virulence factors as it has adapted to its ecological niche, the human body (Wilkening and Federle, 2017). More than 50 GAS virulence factors have been described to date (Fiedler et al., 2015) and these are implicated in adhesion and colonisation of host tissues, immune response evasion and tissue invasion (Fiedler et al., 2010).

Adhesion and colonisation

Adhesion is the first step in the process of GAS colonisation. Without effective adhesion mechanisms, *S. pyogenes* would be easily removed from the host's skin and pharynx by epithelial exfoliation and salivary flow.

Many GAS surface elements have been implicated in the process of adhesion. GAS pili bind to collagen and promote biofilm formation by enhancing bacterial aggregation (Manetti et al., 2007). These pili are encoded in a pathogenicity island known as the fibronectin-binding, collagen-binding, T-antigen region (Manetti et al., 2007). M proteins also facilitate adhesion and achieve this by binding fibronectin and fibrinogen (Schmidt et al., 1993). Proteins of the streptococcal antigen I/II (AgI/II)-family, such as AspA, bind to the salivary protein gp-340, fibronectin and fibrinogen and are also known to play a key role in biofilm formation

(Maddocks et al., 2011). Fibronectin binding proteins, such as PrtF1/SfbI, SOF/SfbII and Fbp54 ligate fibronectin, promoting adhesion (Walker et al., 2014). Additionally, collagen-like proteins, such as Scl1, and laminin binding proteins, such as Lbp and Shr, bind laminin and play a role in GAS adhesion (Caswell et al., 2010). Plasminogen binding proteins, like the streptococcal surface enolase SEN, are involved in adhesion and internalisation by binding plasminogen and thereafter interacting with epithelial integrins (Siemens et al., 2011). The GAS hyaluronic capsule also assists in adhesion, having a role in CD44-mediated tissue invasion (Cywes and Wessels, 2001). Finally, the secreted proteins streptococcal pyrogenic exotoxin B (*speB*) and secreted phospholipase A2 (*slaA*) have the ability to enhance adhesion even though their mechanism of action is still unclear (Brouwer et al., 2016).

Evasion of the host immune response

In order to survive and colonise human skin and mucosae, *S. pyogenes* has developed several mechanisms to resist the host immune response.

One of the main immune barriers to overcome is leukocyte-mediated phagocytosis. The virulence factors used by GAS to escape phagocytosis include inhibitors of complement, leukocidins, immunoglobulin (Ig) binding proteins and Ig-degrading enzymes. The M protein protects GAS against phagocytosis by binding complement-regulating proteins thereby impairing complement deposition (Berggård et al., 2001; Courtney et al., 2006), and by impeding phagosome maturation (Staali et al., 2006). Similarly to the M protein, the hyaluronic capsule surrounding many GAS strains suppresses phagocytosis by impeding complement deposition (Dale et al., 1996). The streptococcal inhibitor of complement (SIC) is a virulence factor secreted only by serotypes M1 and M57 that protects GAS against phagocytosis by inhibiting complement function (Åkesson et al., 1996). Leukocidins are toxins that prevent phagocytosis by killing the host immune cells (Bhakdi et al., 1985; Miyoshi-Akiyama et al., 2005). *S. pyogenes* leukocidins are termed streptolysin O (SLO) and streptolysin S (SLS) and both can damage infected tissues, facilitating tissue invasion (Bricker et al., 2002; Betschel et al., 1998). Proteins belonging to the M family together with the fibronectin binding protein PrtF1/SfbI and the secreted protein SibA can bind different antibodies and thus thwart complement activation and phagocytosis (Carlsson et al., 2003; Medina et al., 2000;

Fagan et al., 2001). GAS strains are able to escape phagocytosis also by releasing proteins that degrade human antibodies. *S. pyogenes* Ig-degrading enzyme (IdeS), also known as Mac-1, along with Sib35 and MspA (Lei et al., 2001; Kawabata et al., 2002), Mac-2, EndoS and SpeB have antibody catalytic activity (Agniswamy et al., 2004; Collin et al., 2002).

GAS strains can survive the damaging action of antimicrobial peptides (AMPs) produced by host cells. *S. pyogenes* is not only able to degrade (Nyberg et al., 2004; Sun et al., 2004) and inactivate (Frick et al., 2003) some AMPs, it is also able to repel them via DltA-mediated esterification of superficial lipoteichoic acid (Kristian et al., 2005).

S. pyogenes is capable of eluding the host immune system by impairing neutrophil intervention. In particular, *S. pyogenes* cell envelope protease (*spyCEP*) and the peptidase ScpA degrade chemotactic molecules, suppressing neutrophilic activity at the site of infection (Edwards et al., 2005; Ji et al., 1996). Moreover, serotype M1T1 can degrade neutrophil extracellular traps (NETs) by producing the DNase enzyme Sda1, which also protects the bacteria against innate immunity responses and macrophage killing (Buchanan et al., 2006; Uchiyama et al., 2012). The integrity of NETs may also be compromised by the action of nuclease A (SpnA) (Chang et al., 2011).

The complex system of virulence factors used by GAS to resist the host immune defences is summarised in Figure 1.1.

Tissue invasion and systemic dissemination

Streptococcus pyogenes can induce tissue damage by interacting with the host plasminogen activation system. GAS strains are able to release the plasminogen activator streptokinase (Ska) and bind plasminogen to surface proteins such as the plasminogen-binding M proteins (Wang et al., 1995; Coleman and Benach, 1999). Following plasminogen recruitment and activation to plasmin, the latter degrades extracellular matrix and basement membranes, facilitating bacterial invasion (Stoppelli, 2013). As previously mentioned, the synthesis and release of the cytotoxins SLO and SLS, as well as the cystein protease SpeB, also facilitate tissue destruction and bacterial invasion (Betschel et al., 1998; Svensson et al., 2000; Bricker et al., 2002).

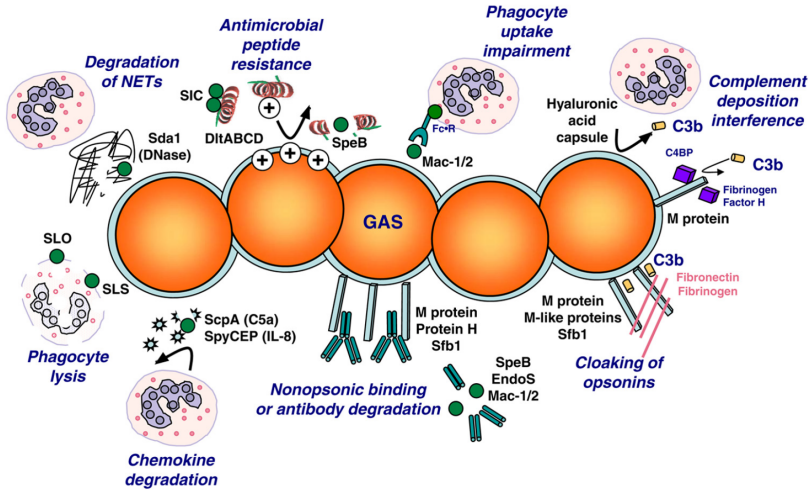


Figure 1.1: Group A *Streptococcus* virulence factors involved in resistance to the host immune response. Source: Walker et al. (2014).

Increased vascular permeability promotes GAS invasion and dissemination in the host body. This can be mediated by two pathways: activation of the coagulation system and/or production of strong pro-inflammatory mediators. The activation of the intrinsic pathway of the coagulation system, which is triggered by the surface M protein, promotes the recruitment of clotting factors such as fibrinogen and kininogen (Loof et al., 2014). Kininogen, which is bound to the M protein, is then cleaved as a consequence of activation of the coagulation system and the potent vasoactive molecule bradykinin is released, increasing vascular permeability at the site of infection (Ben et al., 1997). Similarly, the production of strong pro-inflammatory molecules, called GAS superantigens, promotes vascular permeability and invasive disease (Commons et al., 2014). Known *S. pyogenes* superantigens include the streptococcal pyrogenic exotoxins, such as SpeA, SpeC and SpeG to SpeM, the streptococcal superantigen SSA and the streptococcal mitogenic exotoxin SmeZ (Commons et al., 2014). Most of the genes encoding superantigens are located within prophages, and only *speG* and *smeZ* genes are located in the chromosome (Commons et al., 2014).

Virulence gene regulation systems

More than 100 putative stand-alone and 13 two-component regulatory systems (TCS) have been identified in *S. pyogenes* (Musser and Shelburne, 2009). Some of these can influence either directly or indirectly the expression of virulence factors (Table 1.1).

Table 1.1: Regulators of virulence gene expression in Group A *Streptococcus*. The symbol ↑ indicates promotion of gene expression and ↓ represents suppression of gene expression.

Regulators	Regulatory gene	Maximal activity	Control over gene expression	References
Mga	*mga*	Exponential growth phase	↑ *emm, scl1, scpA, sic, fba, mrp, arp, enn, mga*	(Kreikemeyer et al., 2003)
RALPs	*rofA, nra*	Early stationary phase	↑ *sfbI, rofA* ↓ *sfbII, scpA, sagA, speB, speA, mga*	(Beckert et al., 2001)
Rgg/RopB	*rgg, ropB*	Stationary growth phase	↑ *speB, covRS, ihk/irr, fasBCAX* ↓ *mga, sagA, slo*	(Kreikemeyer et al., 2003)
Ihk/Irr	*ihk, irr*	Exposure to neutrophil reactive oxygen species or AMP	Regulate up to 20% of the whole genome and are involved in neutrophil-mediated killing resistance and invasion	(Voyich et al., 2003)
FasBCAX	*fasA, fasB, fasC, fasX*	End of exponential growth phase	↑ *sagA, ska* ↓ *fbp54, mrp*	(Kreikemeyer et al., 2001)
CovRS	*covR, covS*	Environmental adverse conditions	↑ *speB, sagA, rgg* ↓ *hasA, hasB, hasC, ska, slo, ideS, sda1, sic*	(Musser and Shelburne, 2009)

A well characterised stand-alone regulatory factor for GAS is Mga, which regulates the transcription of genes encoding virulence factors involved in adhesion and immune evasion (Cunningham, 2000). Genes controlled by Mga include: *emm*, which encodes the M protein; *scl1/sclA*, which is the streptococcal collagen-like protein gene; *mrp*, *arp* and *enn*, each of which encode Ig-binding proteins; *scpA*, the gene encoding peptidase ScpA; and *sic*, which encodes the secreted inhibitor of complement (McIver and Scott, 1997; Cunningham, 2000; Lukomski et al., 2001; Terao et al., 2001). Mga also regulates the expression of fibronectin-binding proteins such as Fba, SfbX and SfbII (McIver, 2009). Mga activity peaks during the exponential growth phase and appears to be influenced by sugar availability (Ribardo and McIver, 2006). Conditions of low carbohydrate concentration are thought to induce Mga phosphorylation, which in turn reduces the expression of adhesion virulence factors (Hondorp et al., 2013). This mechanism may be used by *S. pyogenes* to initiate the process of tissue invasion when nutrient availability decreases.

The RofA-like protein (RALP) family represents another stand-alone regulation system of virulence genes involved in adhesion, intracellular migration and host immune evasion (Beckert et al., 2001). The RALP system controls the transcription of *sfbI* and *sfbII* (fibronectin binding proteins 1 and 2 genes), *cpA* (collagen binding protein gene), *sagA* (encoding streptolysin S), *speB*, *speA* and *mga* (Beckert et al., 2001; Podbielski et al., 1999; Molinari et al., 2001). Members of the RALP family can act both as positive and negative regulators of gene expression. For example, RofA is capable of repressing expression of *sagA*, *speB* and *mga*, while another RALP protein, Nra, can also negatively regulate *speA* and *cpa* (Beckert et al., 2001; Molinari et al., 2001). The RALP system is mainly active in the early stage of the stationary growth phase, and appears to promote internalisation and intracellular persistence (Beckert et al., 2001).

The stand-alone regulation system Rgg/RopB regulates the expression of SpeB cysteine protease (Lyon et al., 1998; Ajdic et al., 1999). In some strains, this system interacts with other regulatory pathways such as Mga, CsrRS/CovRS, FasBCAX and Ihk/Irr, influencing the expression of several other virulence genes (Sylva et al., 2002). Rgg/RopB appears to regulate gene expression during the stationary growth phase (Unnikrishnan et al., 1999), and it is thought to be involved in tissue damage and GAS invasion (McIver, 2009). There is also evidence for a role of the Rgg/RopB system in metabolism, as it appears to promote amino acid metabolism in low glucose conditions (Somerville et al., 2003).

The Ihk/Irr TCS is activated following GAS exposure to neutrophil-produced antibacterial molecules (Voyich et al., 2003). It regulates up to 20% of the entire *S. pyogenes* genome and has a role in immune system evasion and tissue invasion (Voyich et al., 2003).

Another TCS implicated in GAS virulence gene control is the FasBCAX system (Kreikemeyer et al., 2001). This system is activated at the end of the exponential growth phase and induces downregulation of adhesion genes, such as *fbp54* and *mrp*. At the same time, the FasBCAX system upregulates the expression of *sagA* and *ska*, which encode SLS and Ska, factors involved in immune system resistance and tissue damage (Kreikemeyer et al., 2001).

The CovR-CovS (or CovRS, control of virulence regulatory system), also known as CsrR-CsrS (or CsrRS, capsule synthesis regulatory system), is a well-characterised TCS of GAS virulence genes. CovRS is triggered by adverse environmental conditions and thus it appears to mediate a stress response (Gryllos et al., 2003; Dalton and Scott, 2004; Churchward, 2007). The system downregulates the expression of the hyaluronic capsule, Ska, SLO, IdeS and Sda1, while it upregulates the expression of SpeB (Sumby et al., 2006). CovRS is mainly activated during the stationary growth phase and it can regulate the transcription of up to 15% of GAS chromosomal genes (Graham et al., 2002).

1.2.4 *Streptococcus pyogenes* antimicrobial resistance

Antimicrobial resistance (AMR) against macrolides, tetracyclines, lincosamides and fluoroquinolones has been documented among GAS strains (Richter et al., 2003; Nielsen et al., 2004; Fay et al., 2021; Tsai et al., 2021). Macrolide resistance is mediated by the *erm* and *mef* genes, which are associated with mobile genetic elements (MGE) such as integrative and conjugative elements (ICE) (Brenciani et al., 2007, 2010). The *mef* genes encode efflux pumps that reduce the concentration of intracellular macrolides (Silva-Costa et al., 2015), whilst the *erm* genes encode methyltransferases that act on the 23S rRNA, conferring resistance also to lincosamides such as clindamycin (Silva-Costa et al., 2015). The *ermB* gene usually confers constitutive resistance to clindamycin. This is a form of resistance that is intrinsically expressed and is not influenced by external factors. The *ermA* and *ermTR* genes, on the other hand, can provide macrolide-mediated clindamycin resistance, also known as inducible clindamycin resistance (Pesola et al., 2015). Tetracycline resistance is mediated by *tet* genes, which are also acquired from MGE (Giovanetti et al., 2003). The frequent co-location of *tet* and macrolide resistance genes within the same MGE can facilitate the acquisition of both forms of resistance in a single horizontal gene transfer event (Giovanetti et al., 2003; Brenciani et al., 2007). Quinolone resistance is known to occur through the acquisition of non-synonymous point mutations in the quinolone resistance-determining regions (QRDR) of the *gyrA*, *gyrB*, *parC* and *parE* genes (Richter et al., 2003; Arai et al., 2011).

Resistance to β-lactams has never been documented in GAS infections, making this class of antibiotic the treatment of choice (Walker et al., 2014). The recent identification of single nucleotide polymorphisms (SNP) conferring reduced sensitivity to ampicillin, amoxicillin and cefotaxime, however, may be the first step towards the future development of full β-lactam resistance (Grebe and Hakenbeck, 1996; Vannice et al., 2019; Musser et al., 2020).

1.2.5 *Streptococcus pyogenes* epidemiology

Superficial disease

Due to the high rate of GAS throat colonisation (with estimated prevalence between 15 and 20% of the population) (Henningham et al., 2012), it is difficult to estimate with accuracy the burden of *S. pyogenes*-associated pharyngitis. The results of some population studies revealed that in developed countries about 15% of school-age children and 4-10% of adults experience at least one episode of GAS pharyngitis per year, with the rate being 5-10 fold higher in developing countries (Carapetis et al., 2005; Ralph and Carapetis, 2012). In a study published in 2005, the annual global burden of GAS pharyngitis was estimated to be higher than 600 million cases, with 550 million cases occurring in developing countries (Carapetis et al., 2005). The majority of the cases of GAS pharyngitis documented in a study carried out in Australia in 2007 were reported in late winter and early spring (Danchin et al., 2007), suggesting fluctuating incidence rates throughout the year.

S. pyogenes pyoderma is more common in tropical and sub-tropical regions, particularly in the summer time (Bessen et al., 2015). Its prevalence is heavily influenced by accessibility to hygiene, and it was estimated to range from 1 to 20% in developing countries, with some regions of the world having prevalence between 40 and 90% (Ralph and Carapetis, 2012).

Scarlet fever incidence has decreased worldwide over the last century (Efstratiou and Lamagni, 2016). Recently, however, there has been an increase in number of scarlet fever cases and large-scale outbreaks (Efstratiou and Lamagni, 2016). In 2014, the incidence of scarlet fever in the UK reached 49 per 100,000 individuals in some parts of the country, while the previously reported incidence was 4 per 100,000 (Guy et al., 2014). The most recent report from Public Health England shows that scarlet fever incidence

in England has decreased since 2014, with a peak value of 22.4 per 100,000 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/800932/hpr1619_gas-sf3.pdf). Risk factors for scarlet fever include age, with children under ten years old being more susceptible, and season, with the majority of the cases occurring in spring (Briko et al., 2003; Guy et al., 2014).

Invasive disease

The incidence of invasive GAS (iGAS) disease has been increasing since the 1980s (Henningham et al., 2012). GAS cellulitis is the most common form of invasive disease. In one study published in 2007, for example, the incidence of GAS cellulitis was as high as 200 cases per 100,000 people in Minnesota (McNamara et al., 2007). More severe forms of iGAS disease are less common, and their reported incidence in 2005 was 1.5-3.9 cases per 100,000 people in the USA, Canada and North Europe; 6.4-10.5 per 100,000 in the Australian non-indigenous population; 13 per 100,000 in Kenyan children; and 82.5 per 100,000 in the Australian indigenous population (Ralph and Carapetis, 2012). Excluding cases of cellulitis, iGAS disease is associated with a high mortality rate of around 20%, a figure which is independent of disease incidence (Sanyahumbi et al., 2016). Known risk factors for iGAS disease are skin lesions, blunt trauma, acute viral infections, influenza, varicella and other non-infectious co-morbidities (Efstratiou and Lamagni, 2016). Between 20 and 30% of all iGAS cases are not associated with known risk factors (Lamagni et al., 2008a).

Post-infectious complications

ARF is still one of the most common causes of acquired heart disease in children worldwide (Carapetis et al., 2005; Jackson et al., 2011). Every year approximately 500,000 new cases of ARF occur (Webb et al., 2015), mostly in poorer areas of the world (Carapetis et al., 2005). The incidence of the disease varies considerably in different countries, with a reported minimum of 0.1 cases per 100,000 people in Greece and a maximum of 826 per 100,000 in Sudan (Jackson et al., 2011). In between 42 and 60% of cases of ARF, RHD occurs (Ralph and Carapetis, 2012). The heaviest burden of RHD is on children and young adults (WHO, 2005), and its incidence can be as high as 850 per 100,000 people, among

the indigenous children of the Australian northern territory (Parnaby and Carapetis, 2010). In 2005 the prevalence of RHD was estimated to be between 15.6 and 19.6 million cases worldwide (Carapetis et al., 2005) with a mortality rate in poor countries of approximately 1.5% (Carapetis et al., 2005).

APSGN is more prevalent in resource-limited settings, being associated with crowding, poor hygiene and poverty (Marshall et al., 2011). The incidence varies accordingly, ranging from 0.04 per 100,000 in an Italian cohort of people under 60 years of age to 239 per 100,000 in indigenous Australians (Jackson et al., 2011). Although the mortality associated with this complication is generally low (0.02-0.4 per 100,000), considerable morbidity has been highlighted in numerous studies (Jackson et al., 2011).

Streptococcus pyogenes emm type distribution

In 2009, a study on the global distribution of GAS *emm* types was published (Steer et al., 2009). Most of the data analysed were collected from high-income countries, with limited data derived from developing regions of the world. The most common *emm* types in Europe, North America, Asia and Latin America were *emm*1 and *emm*12. In Africa, the most common type was *emm*12, followed by *emm*75, while in the Pacific Region, the most prevalent type was *emm*55. The diversity of *emm* types in Africa and the Pacific region appeared to be higher than in the rest of the sampled countries. In high-income countries, the order of prevalence of *emm* types from 1990 to 2009 was relatively stable (Steer et al., 2009).

A more recent study focusing on iGAS in Europe and North America confirms *emm*1 as the most frequently isolated strain (Gherardi et al., 2018). The authors pointed out that the seven most common types circulating in Europe and North America, namely *emm*1, *emm*28, *emm*89, *emm*3, *emm*12, *emm*4 and *emm*6, accounted for 50-70% of all isolates. In a study conducted in Scotland on both invasive and non-invasive GAS isolates, the most frequent *emm* types circulating from 2011 to 2015 were *emm*1, *emm*76 and *emm*89 (Lindsay et al., 2016).

The known genetic determinants of virulence for the five most common *emm* types involved in iGAS disease in Europe and North America are presented in Table 1.2.

Table 1.2: Most common *emm* types involved in invasive *Streptococcus pyogenes* disease in Europe and North America and genetic mechanisms underlying their high pathogenicity.

***emm* type**	**Genetic determinants of pathogenicity**
*emm*1	-Phage encoding an extracellular DNAse (SdaD2) -Phage encoding the SpeA2 toxin (derived from a single nucleotide mutation of the SpeA1 variant of SpeA) -Acquisition, by horizontal gene transfer, of a chromosomal region encoding secreted toxins NAD+-glycohydrolase and SLO (Nasser et al., 2014)
*emm*3	-Phages encoding SpeA, SpeK, SSA and a phospholipase A2 (Sla) (Beres et al., 2002)
*emm*12	-NADase and SLO genes -Demonstrated capacity to acquire CovRS mutations *in vivo* (Feng et al., 2016)
*emm*28	-Demonstrated capacity to acquire CovRS mutations *in vivo* -Demonstrated capacity to acquire single nucleotide indels in the intergenic region between *Spy1336/R28* and *Spy1337* (with possible increased expression of the protein R28 and consequent higher virulence) (Kachroo et al., 2019)
*emm*89 (clade 3)	-Variation in the *nga* promoter region pattern, which is associated with increased production of SPN (*S. pyogenes* NADase) and SLO -HGT mediated loss of *hasABC* genes, with consequent loss of the hyaluronic acid capsule production -Demonstrated capacity to acquire CovR and LiaS mutations *in vivo*, with associated increase in virulence (Turner et al., 2015; Beres et al., 2016)

Surveillance systems for GAS infections

Since currently there is no surveillance network for iGAS disease in Europe, each country undertakes surveillance according to its own criteria (Creti, 2017). In most European countries, *S. pyogenes* infection surveillance is based on voluntary reporting systems, while in a limited number of countries iGAS and/or scarlet fever are notifiable diseases (Lamagni et al., 2005). Across the UK, where iGAS are the only forms of *S. pyogenes*-associated disease notifiable (scarlet fever is also notifiable in England), surveillance is based on isolates submissions to reference laboratories and the collection of enhanced surveillance data (`https://www.gov.uk/guidance/rvpbru-reference-and-diagnostic-services`).

1.2.6 *Streptococcus pyogenes* vaccine development

Despite the globally high burden of GAS disease, there is no commercially available vaccine to prevent *S. pyogenes* infection. Challenges to the development of a GAS vaccine include a high antigenic diversity among GAS strains, cross-reactivity between GAS-targeted antibodies and certain human cells, lack of commercial interest from pharmaceutical companies and the absence of suitable animal models of infection (Dale and Walker, 2020). Despite these barriers, multiple research groups have worked towards the development of a GAS vaccine for a number of decades. Vaccine candidates that have been investigated so far include the M protein, cell-wall carbohydrate and multi-component formulations of antigenic proteins such as the streptococcal superantigens, fibronectin-binding proteins and SLO (Castro and Dorfmueller, 2021). The only candidates to reach clinical trial thus far are M-protein based vaccines such as StreptAvax and StreptAnova (Castro and Dorfmueller, 2021). While it is encouraging to note that the evidence suggests an effective GAS vaccine will be available at some point, years of testing are anticipated before a vaccine candidate is finally approved and produced (Dale and Walker, 2020).

1.2.7 Streptococcal species phylogenetically related to GAS

Streptococcus pyogenes and a number of other streptococcal species have been categorised as pyogenic *Streptococci* based on their associated clinical manifestation. As the name suggests, the pyogenic group encompasses species involved in purulent soft tissue infections

in humans and/or animals. The other pyogenic streptococci are: *S. agalactiae*, *S. canis*, *S. equi*, *S. iniae*, *S. uberis*, *S. parauberis*, *S. phocae*, *S. urinalis*, *S. didelphis* and strains of *S. dysgalactiae* and *S. porcinus* (Vos et al., 2011). Based on a 16S rRNA gene phylogeny, the pyogenic species form a single monophyletic group (Figure 1.2).

Among the species belonging to the pyogenic group, the more relevant from a veterinary and "One Health" standpoint are *S. dysgalactiae* spp., *S. canis*, *S. equi* subsp. *equi* and *zooepidemicus*, *S. uberis* and *S. agalactiae* (Lefébure et al., 2012). Evidence of lateral gene transfer events from these species to *S. pyogenes* suggests that the evolution of GAS has been, and potentially will in the future be, influenced by closely related streptococci (Lefébure et al., 2012). The main biochemical characteristics of *S. pyogenes*, *S. dysgalactiae*, *S. canis*, *S. equi*, *S. uberis* and *S. agalactiae* are reported in Table 1.3, while their major genomic features are shown in Table 1.4.

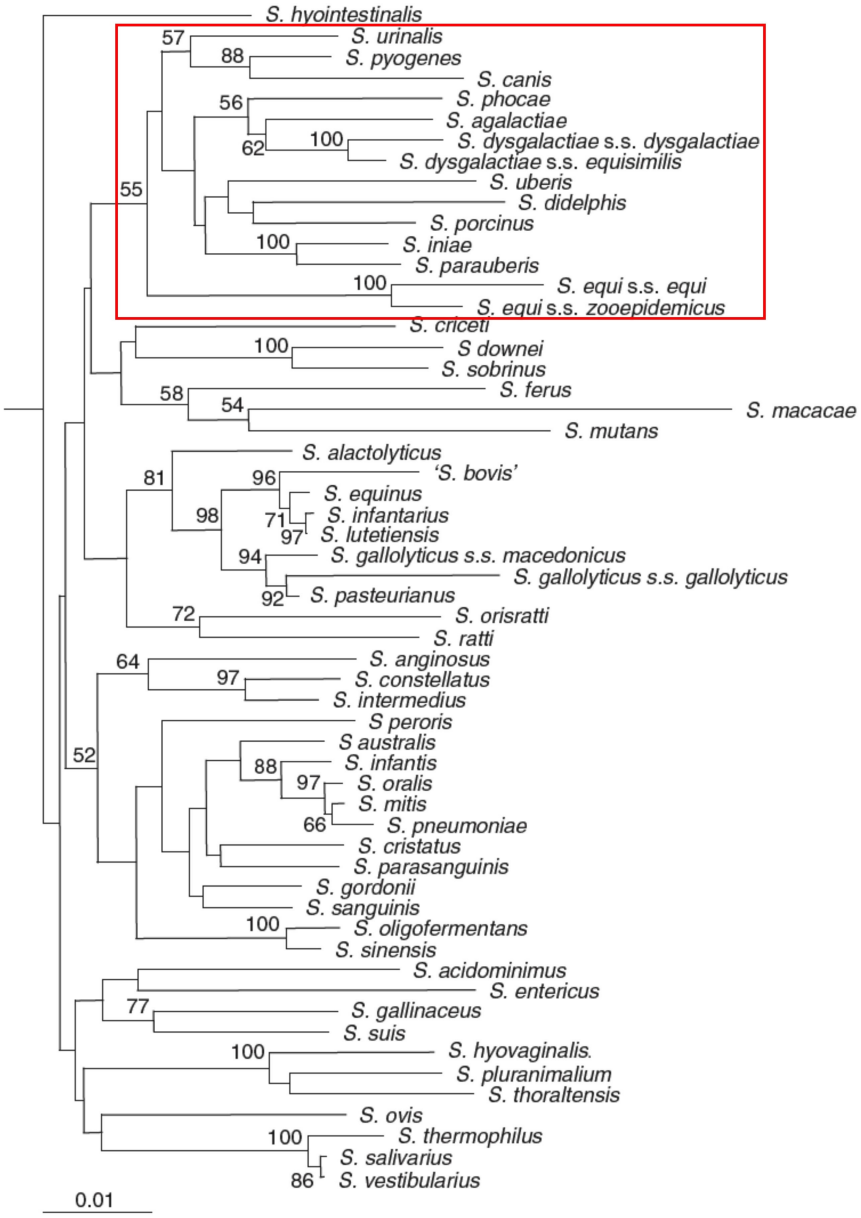


Figure 1.2: Neighbor-joining phylogeny of streptococcal species based on the 16S rRNA gene. Bootstrap values for each node are shown, based on 100 iterations. The red rectangle highlights the species belonging to the pyogenic group. Adapted from Vos et al. (2011).

Table 1.3: Main phenotypic characteristics used to differentiate streptococcal species. Here reported for *Streptococcus pyogenes*, *Streptococcus dysgalactiae*, *Streptococcus canis*, *Streptococcus equi*, *Streptococcus uberis* and *Streptococcus agalactiae*. Symbols: +, >85% positive; d, different strains give different reactions (16-84% positive); -, 0-15% positive; NG, not groupable; ND, not determined. Adapted from (Vos et al., 2011).

Characteristic	***S. pyogenes***	***S. dysgalactiae***	***S. canis***	***S. equi***	***S. uberis***	***S. agalactiae***
Lancefield group antigen	A	C,G,L,A	G	C	NG	B
Catalase	-	-	-	-	-	-
Acid from glycogen	d	d	-	+	d	-
Acid from mannitol	-	d	-	-	+	-
Acid from raffinose	-	-	-	-	-	-
Acid from ribose	-	+	+	- ssp. equi, + ssp. zooepidemicus	+	+
Acid from sorbitol	-	d	-	- ssp. equi, + ssp. zooepidemicus	+	-
Acid from sucrose	+	+	+	+	+	+
Acid from trehalose	+	+	d	d	+	+
Hydrolisys of arginine	+	+	+	+	+	+
Hydrolisys of esculine	d	d	+	d	+	-
Hydrolisys of hippurate	-	-	-	-	+	+
Hydrolisys of starch	-	-	-	+	ND	-
Production of α-D-galactosidase	-	-	d	-	-	d
Production of β-D-galactosidase	-	- ssp. dysgalactiae, + ssp. equisimilis	+	-	-	-
Production of β-D-glucuronidase	d	+	d	+	+	d
Pyrolidonylarylamidase	+	-	-	-	+	-
Acetoin (V–P)	-	-	-	-	+	+
CAMP reaction	-	-	-	-	-	+
Host	Human	Human, horse, cattle, pig, dog	Human, dog, cat, cattle	Human, horse, cattle, sheep, goat, pig, dog, cat	Cattle	Human, cattle, fish

Table 1.4: Major genomic features of *Streptococcus pyogenes*, *Streptococcus dysgalactiae*, *Streptococcus canis*, *Streptococcus equi*, *Streptococcus uberis* and *Streptococcus agalactiae*.

Characteristic	***S. pyogenes***	***S. dysgalactiae***	***S. canis***	***S. equi***	***S. uberis***	***S. agalactiae***
Genome size						
(Mb)	1.8	2.1	2.2	2.1	1.8	2.1
GC %	38.5	39.5	39.7	41.5	36.6	35.6
Protein count	1,693	1941	2212	1874	1762	2127
Gene count	1,801	2128	2297	2025	1871	2279

1.3 *Streptococcus canis*

The name *Streptococcus canis* was first used in 1937 to identify streptococci implicated in infection in dogs (Stafseth et al., 1937). Only in the late 1980s was the name formally ascribed to a bacterial species with defined phenotypic characteristics (Devriese et al., 1986), when it was described as Gram positive, β-haemolytic, Lancefield group G pyogenic coccus that could infect dogs and cattle (Devriese et al., 1986). Numerous biochemical and physiological traits of the newly identified species were reported (Devriese et al., 1986). *S. canis* was initially thought to be solely a canine and bovine pathogen (Devriese et al., 1986) but has since been isolated from a range of mammals including cats, rats, rabbits, minks, foxes, Japanese racoon dogs, kinkajous, seals, sea lions, otters, badgers and humans (Richards et al., 2012; Numberger et al., 2021). It may cause disease in each species, making it one of the streptococcal pathogens with the widest host range (Fulde and Valentin-Weigand, 2012). Despite its broad host tropism, *S. canis* has not been given the same attention as other streptococci (Fulde and Valentin-Weigand, 2012) and this is probably due to the limited number of confirmed cases of infection in humans (Lam et al., 2007). Since most streptococcal isolations in human medicine are not identified to the species level, however, the true disease burden of *S. canis* disease in humans is difficult to estimate (Lam et al., 2007).

1.3.1 *S. canis* epidemiology

Companion animals

In dogs and cats, *S. canis* is regarded as an opportunistic pathogen that can colonise the skin and mucosae of asymptomatic individuals (Lysková et al., 2007b; Timoney et al., 2017). When implicated in disease, *S. canis* is mainly associated with superficial infections (Devriese et al., 1986), with the most common isolation sites being the oral and nasal cavities, the external ear canal, rectum and the genital mucosae (Figure 1.3) (Devriese et al., 1992; Lysková et al., 2007b). However, infection in dogs and cats may sometimes result in severe

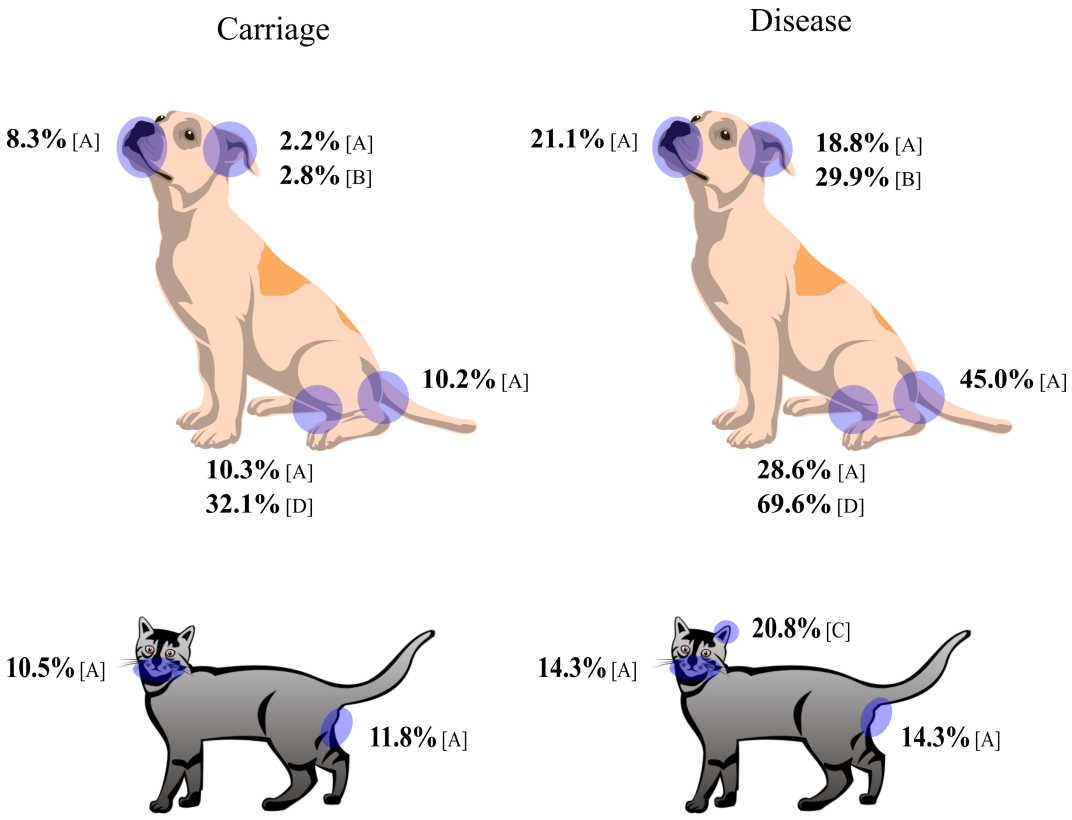


Figure 1.3: Main isolation sites of *Streptococcus canis* in healthy and diseased dogs and cats. The percentages reported indicate the fractions of samples from which *S. canis* was detected in the studies referenced (A: Lysková et al., 2007a; B: Lysková et al., 2007b; C: Guerrero et al., 2018; D: Dégi et al., 2011). Carriage data refer to the detection of *S. canis* in samples collected in body sites that were not showing signs of disease. In dogs, *S. canis* is more frequently isolated from the oral and nasal cavities, the ear canal, the rectum and the genital mucosa. In cats, it is more commonly isolated from the oral and nasal cavities, the ear canal and the rectum.

clinical syndromes such as arthritis (Iglauer et al., 1991), necrotising fasciitis (Prescott et al., 1995; Pesavento et al., 2007), myocarditis (Matsuu et al., 2007), pneumonia (Prescott et al., 1995), meningitis (Pesavento et al., 2007), sepsis (Pesavento et al., 2007) and STSS (Prescott et al., 1995; Pesavento et al., 2007). In a 2007 study, 6.5% of healthy dogs (n=35/539) and 5.9% of healthy cats (n=10/169) tested positive for carriage of *S. canis*, which was isolated principally from the rectum of both species, the praeputium of dogs and the oral cavity of cats (Lysková et al., 2007b). In the same study, it was isolated from various body sites in 22.2% of dogs (n=39/176) and 4.8% of cats (n=2/42) with ongoing infections. Among clinically ill dogs, it was frequently isolated from those with signs of gastrointestinal disease, urogenital infection, otitis externa and rhinitis. In clinically ill cats, *S. canis* was isolated from just two of 42 specimens. However, since co-infection with other pathogens was not considered, it is impossible to determine whether *S. canis* was responsible for the clinical signs reported. As sampling was skewed towards canine samples and external ear canal specimens (Lysková et al., 2007b), this may have contributed to biases in the results reported. Two other studies report a high prevalence of *S. canis*-associated otitis externa in pets. In one, *S. canis* was shown to be the third most common microorganism isolated from dogs with otitis externa (29.9% of the cases) (Lysková et al., 2007a), being found significantly more frequently in the ear canals of dogs with otitis externa than from healthy dogs ($P<0.001$) (Lysková et al., 2007a). Another study revealed a prevalence of 20.83% from the ears of cats with otitis externa, although the sample size was very small (n=24) (Dégi et al., 2011). In a work by Lamm et al., the prevalence of streptococcal isolation from all canine specimens submitted to a diagnostic laboratory was 20.5% (n=499/2432), of which 22.4% (n=106/499) were confirmed as *S. canis* (Lamm et al., 2010). A high proportion of the sampled dogs that tested positive for *Streptococcus* spp. (n=267) showed co-infection with other pathogens, meaning that causative role of streptococci in those disease cases could not be established. The authors found that *S. canis* was the most common streptococcal species isolated from infection sites in dogs and that *S. canis* infection can be associated with dermatitis, septicaemia, placentitis and pneumonia (Lamm et al., 2010). A more recent work by Guerrero et al. suggested an association between vaginal carriage of β-haemolytic streptococci and neonatal death in dogs (Guerrero et al., 2018). No significant difference in the frequency of vaginal isolation of *S. canis*, however, was found between dogs with healthy litters and dogs experiencing neonatal

Table 1.5: Case reports of *Streptococcus canis* infection in companion animals reviewed for this study.

Reference	Host species	Number of cases	Clinical manifestations	Suggested predisposing factors
Iglauer et al., 1991	Cat	6	Arthritis	Possible genetic predisposition due to high inbreeding.
Prescott et al., 1995	Dog	3	Necrotising fasciitis	Trauma.
Matsuu et al., 2007	Cat	1	Myocarditis	Acquired mitral stenosis associated with congenital malformation of the mitral valve complex.
Pesavento et al., 2007	Cat	>150 (3 outbreaks)	Skin ulceration, sinusitis, meningitis, necrotising fasciitis	Indirect contact with dogs and concomitant upper respiratory tract infections.

losses (Guerrero et al., 2018). The role of *S. canis* vaginal colonisation in canine fertility is still unclear and further studies are required. *S. canis* infection outbreaks have also been reported in feline colonies and shelters. An outbreak of contagious arthritis due to *S. canis* in a cat breeding colony over a six-month period has been described (Iglauer et al., 1991). A high level of inbreeding among colony cats was suggested to have contributed to susceptibility to infection (Iglauer et al., 1991), although outbreaks have also been detected among shelter cats. Three outbreaks of *S. canis* infection in cat shelters were reported (Pesavento et al., 2007), two of which were characterised by skin ulceration, sinusitis and meningitis while a third outbreak was associated with necrotising fasciitis and sudden death (Pesavento et al., 2007). Table 1.5 provides a summary of all case reports referenced in this subsection.

Production animals

In cattle, *S. canis* is a recognised cause of mastitis (Chaffer et al., 2005; Hassan et al., 2005; Tikofsky and Zadoks, 2005). Although the prevalence of Group G *Streptococcus* mastitis is thought to be low (Wilson et al., 1997), *S. canis* mastitis outbreaks have been reported, with herd prevalence as high as 38% (Chaffer et al., 2005). In a case of *S. canis* sub-clinical mastitis outbreak that affected 22% (n=11/49) of a dairy herd, PGFE genotyping revealed the isolates were either identical or very closely related, suggesting a clonal spread of *S. canis* that may be explained by cow-to-cow transmission (Hassan et al., 2005). A study by Tikof-

Table 1.6: Case reports of *Streptococcus canis* infection in dairy cattle with subclinical mastitis.

Reference	Number of cases	Herd size	Proportion of herd affected	Suggested risk factors
Chaffer et al., 2005	26	69	38%	Not mentioned
Hassan et al., 2005	11	49	22%	Not mentioned
Tikofsky & Zadoks 2005	46 Group G *Streptococcus cases*. 12 confirmed *S. canis* cases	90	51% Group G *Streptococcus*. 13% confirmed *S. canis*	Direct contact with an infected cat
Król et al., 2015	17	76	22%	Not mentioned
Eibl et al., 2021	9	59	15%	Direct contact with an infected cat

sky and Zadoks described another mastitis outbreak that affected 13% (n=12/90) of lactating cows in a dairy herd (Tikofsky and Zadoks, 2005). The origin of the outbreak was thought to be a cat with chronic sinusitis due to *S. canis* infection. The cat, whose infection predated the outbreak, lived in close contact with the herd. All bovine and feline *S. canis* isolates showed the same ribotype pattern, supporting the hypothesis that the cat was the outbreak source and that infection subsequently spread from cow to cow (Tikofsky and Zadoks, 2005). A similar case of an outbreak of bovine sub-clinical *S. canis* mastitis associated with a cat was reported by Eibl et al. (Eibl et al., 2021). In this instance, strains of the MLST type were isolated from nine cows and one cat living in contact with the herd, but no directionality of transmission could be determined (Eibl et al., 2021). These reports highlight the potential spread of infectious agents from pets to cattle, which should be considered when assessing biosecurity measures on dairy farms. Both reports, however, rely on low discrimination methods to assess the genetic relatedness of bovine and feline isolates (Salipante et al., 2015; Tsang et al., 2017), so should not be considered as conclusive evidence of cats being the source of infection in outbreak scenarios. Importantly, *S. canis* mastitis outbreaks have been documented in dairy herds that were not in contact with dogs and cats, showing alternative routes of herd infection may occur (Chaffer et al., 2005). Król et al. demonstrated the contagious potential of *S. canis* among cows (Król et al., 2015). Relatedness of the outbreak isolates was confirmed by RAPD analysis and PFGE. The authors also showed that *S. canis* was capable of

causing long-term sub-clinical mastitis that persisted for up to 14 months (Król et al., 2015). A summary of *S. canis* mastitis case reports in dairy cattle is shown in Table 1.6.

Role in human health

Streptococcus canis appears to be rarely isolated from humans, although the actual infection burden is hard to estimate (Lam et al., 2007). It shares the same Lancefield classification (group G) with other β-haemolytic streptococci, such as *S. dysgalactiae* and *S. anginosus*, recognised to infect humans. The determination of Lancefield antigenic group is often sufficient for diagnostic and public health purposes and for this reason the prevalence of *S. canis* infection is likely to be underestimated (Lam et al., 2007). In a retrospective study carried out at the University Hospital of Bordeaux from 1997 to 2002, *S. canis* was confirmed in 1% (n=80/6404) of all *Streptococcus*-positive samples submitted for culture (Galpérine et al., 2007). Clinical and microbiological data available for a subset of cases (n=54) revealed that *S. canis* was mainly involved in skin and soft tissue infection (n=35), and occasionally implicated in bacteraemia (n=5), urinary tract infection (n=3), osteoarticular infection (n=2), pneumonia (n=1) and asymptomatic carriage (n=8). Toxic shock was noted in two patients. The majority of the cases for which clinical data was available were confirmed as community acquired (n=39) and mortality attributable to *S. canis* infection was 3.7% (n=2/54). Most patients had comorbidities that predated infection and the majority of *S. canis*-positive samples for which data were available (n=42/54) contained additional bacterial pathogens (Galpérine et al., 2007). It is, therefore, impossible to determine to what extent the presence of *S. canis* contributed to pathology. Figure 1.4 illustrates the common sites of *S. canis* isolation in humans (Galpérine et al., 2007).

Sporadic cases of human infection were described in the literature as case reports, with clinical manifestations such as purulent skin infection (Bert and Lambert-Zechovsky, 1997; Whatmore et al., 2001; Lam et al., 2007), cellulitis (Takeda et al., 2001; Lam et al., 2007), septicaemia (Bert and Lambert-Zechovsky, 1997; Takeda et al., 2001; Whatmore et al., 2001; Ohtaki et al., 2013; Taniyama et al., 2017), endocarditis (Amsallem et al., 2014; Lacave et al., 2016; Mališová et al., 2019), arthritis and bone infection (Tarabichi et al., 2018; McGuire et al., 2021). The majority of case reports of *S. canis* infection involve patients

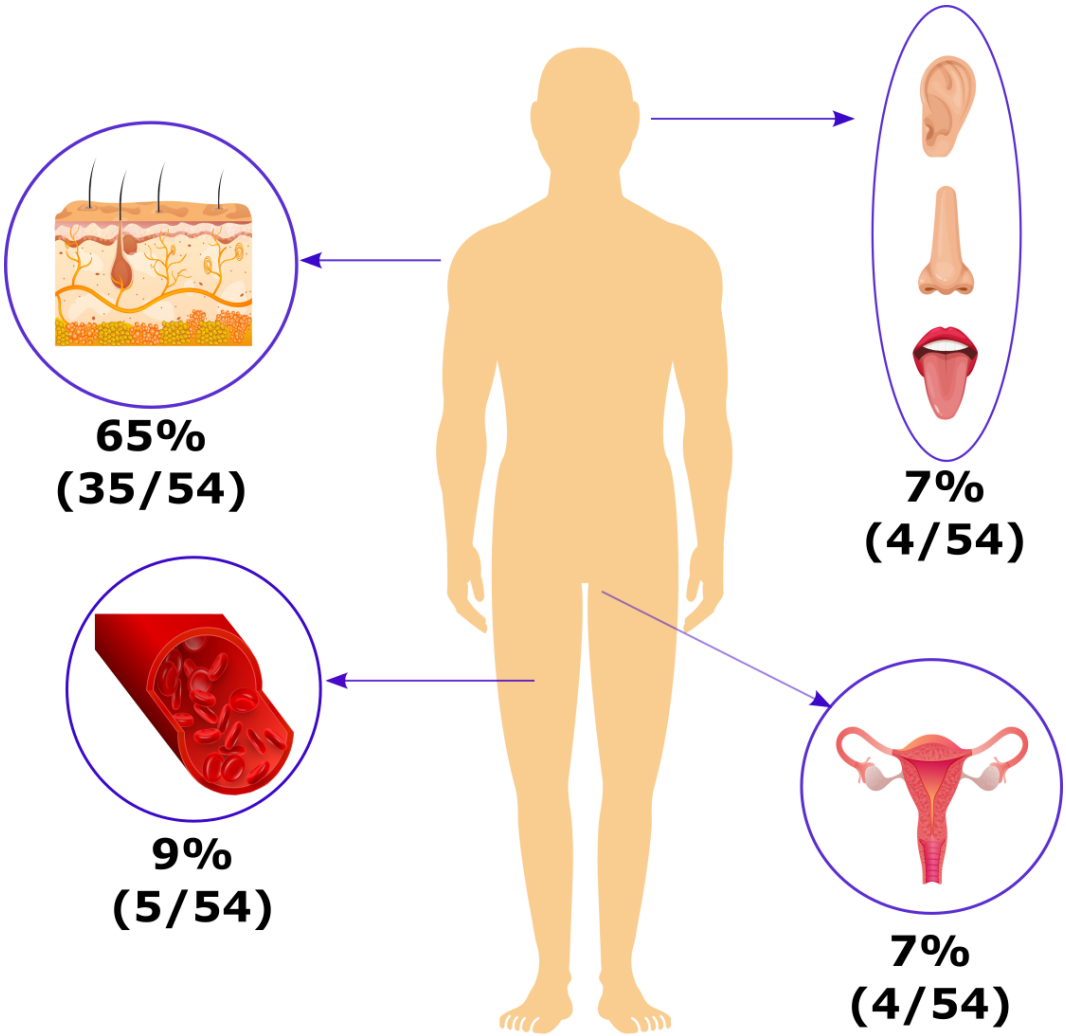


Figure 1.4: Sites of *Streptococcus canis* isolation in humans (Galpérine et al., 2007). This bacterium was isolated principally from the cutaneous tissue, bloodstream, ear-nose-throat (ENT) sphere and vaginal swabs. Percentages refer to the frequency of isolation from the total number of *S. canis*-positive samples.

above 60 years of age, with various comorbidities or previous trauma. Notably, a proportion of cases describe prior interactions with dogs (Takeda et al., 2001; Lam et al., 2007; Ohtaki et al., 2013; Amsallem et al., 2014; Lacave et al., 2016; Taniyama et al., 2017; Tarabichi et al., 2018; Mališová et al., 2019; McGuire et al., 2021), in particular dog bites or scratches (Takeda et al., 2001; Taniyama et al., 2017; Tarabichi et al., 2018). However, more direct evidence to support the contention that dogs may be a source of *S. canis* zoonotic infection was presented in only one report, which described a woman developing *S. canis*

Table 1.7: Reports of *Streptococcus canis* infection in humans.

Reference	Number of cases	Clinical manifestations	Suggested predisposing factors
Bert & Lambert-Zechovsky, 1997	1	Septicemia	Comorbidities, direct contact with a dog, > 60 years of age.
Takeda et al., 2001	1	Cellulitis and septicemia	Comorbidities, dog bite, >60 years of age.
Whatmore et al., 2001	2	Wound infection (first case) and bacteremia (second case)	Not mentioned for the first case, comorbidities and > 60 years of age for the second case.
Ohtaki et al., 2013	1	Septicemia	Trauma, direct contact with a dog, >60 years of age.
Amsallem et al., 2014	1	Endocarditis	Comorbidities, direct contact with a dog, >60 years of age.
Lacave et al., 2016	1	Endocarditis	Comorbidities, direct contact with a dog, > 60 years of age.
Taniyama et al., 2017	1	Cellulitis and bacteremia	Comorbidities, dog bite, >60 years of age.
Tarabichi et al., 2018	1	Periprostetic joint infection and septicemia	Knee prostesis, dog scratch, >60 years of age.
Mališová et al., 2019	1	Endocarditis	Comorbidities, direct contact with a dog, >60 years of age.
McGuire etal., 2021	1	Periprostetic joint infection	Hip surgery, direct contact with a dog, >60 years of age.

septicaemia two weeks after a dog bite (Takeda et al., 2001). *S. canis* was also isolated from the dog's oral cavity and both human and canine strains shared the same PFGE pattern, suggesting a canine-to-human transmission event (Takeda et al., 2001). Although generally reliable, PFGE results are occasionally discordant with higher resolution methods such as whole genome sequencing (Salipante et al., 2015). Further evidence is required to clarify the role dogs play in the transmission of *S. canis* to humans. The reviewed case reports of human *S. canis*-associated disease are summarised in Table 1.7. Based on the epidemiological studies and clinical reports available, a transmission cycle for *S. canis* including the environment, human, canine, feline and bovine hosts is hypothesised and is visually represented in Figure

1.5.

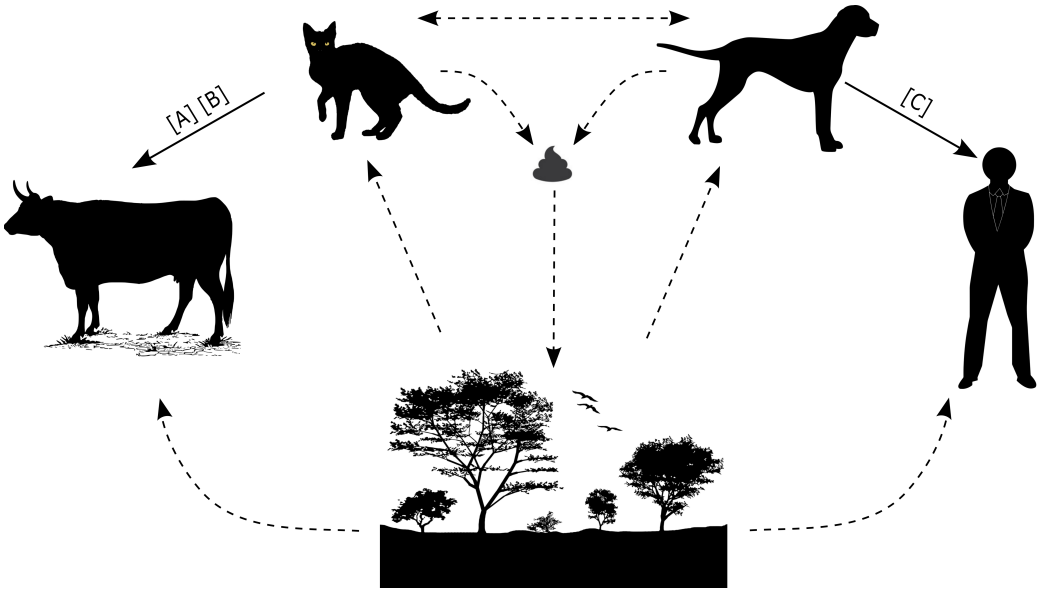


Figure 1.5: Schematic representation of a possible transmission cycle of *Streptococcus canis*. The main host species of *S. canis* appear to be dogs and cats. Dogs and cats have been reported as a potential source of infection for humans and cattle, respectively (A: Tikofsky & Zadoks 2005; B: Eibl et al., 2021; C: Takeda et al., 2001). *S. canis* can be frequently isolated from the rectum of dogs and cats, implying that faecal contamination of the environment, although never demonstrated to our knowledge, may occur. Environmental contamination may be a source of infection not only for dogs and cats but also for other susceptible species, namely wildlife and humans. In the diagram, *S. canis* transmission is represented through solid arrows (direct route) and dashed arrows (indirect route).

1.3.2 *S. canis* virulence mechanisms

The knowledge on pathogenesis and virulence mechanisms of *S. canis* is currently limited. This may be explained by the low prevalence of infection in humans and production animals, together with the fact that it is broadly sensitive to commonly used antibiotics, which may contribute to it being given a low priority (Galpérine et al., 2007; Pinho et al., 2013). However, the health threat represented by *S. canis* should not be underestimated, particularly in light of the severe disease cases reported in humans and the documented acquisition of AMR (Takeda et al., 2001; Galpérine et al., 2007; Lam et al., 2007; Lacave et al., 2016; Tan et al., 2016; Fukushima et al., 2020b; McGuire et al., 2021). Potential virulence determinants of *S. canis* are summarized in Table 1.8. The presence of sequences homologous

Table 1.8: Virulence traits investigated in *Streptococcus canis* in the literature.

Virulence traits	Evidence provided	Reference
Arginine deaminidase system (ADS)	Experimental evidence and bioinformatic analysis	Hitzmann et al., 2013
Christine, Atkins and Munch-Peterson (CAMP) factor	Detection of homologous gene based on WGS bioinformatic analysis	Richards et al., 2012
Intracellular invasion	Experimental evidence	Yoshida et al., 2021
Neuraminidase B	Detection of homologous gene based on WGS bioinformatic analysis	Richards et al., 2012
Resistance to phagocytosis	Experimental evidence – hypothesised role of M protein	DeWinter et al., 1999
Streptococcus canis M-like (SCM) protein	Detection of homologous gene based on Southern hybridisation	DeWinter et al., 1999
	Detection of homologous gene based on WGS bioinformatic analysis	Richards et al., 2012
	Experimental evidence - adherence and tissue invasion, plasminogen-mediated	Fulde at al., 2011a
	Experimental evidence - resistance to phagocytosis	Fulde et al., 2013
	Experimental evidence - overall virulence activity questioned. SCM might facilitate adhesion and persistence in the vaginal environment and biofilm formation	Cornax et al., 2021
Streptolysin O (SLO)	Detection of homologous gene based on Southern hybridisation	DeWinter et al., 1999
	Detection of homologous gene based on WGS bioinformatic analysis	Richards et al., 2012
Streptolysin S (SLS)	Detection of homologous gene based on WGS bioinformatic analysis	Richards et al., 2012
Surface fibrillae	Direct observation through electron microscopy	DeWinter et al., 1999

to well-characterised *S. pyogenes* virulence genes was assessed in the genome of *S. canis*, including 15 isolates from dogs diagnosed with STSS and/or necrotising fasciitis, by Southern hybridisation (DeWinter et al., 1999). Genes homologous to the *S. pyogenes slo* and *emm*, encoding SLO and the M protein, respectively, were detected in the genome of the majority of isolates analysed. However, no matches were found to eight other *S. pyogenes* virulence genes (*speA*, *speB*, *speC*, *speF*, *scpA*, *hasA*, *ska* and *ssa*). Resistance to phagocyto-

sis and presence of surface fibrillae were also observed as *S. canis* virulence characteristics (DeWinter et al., 1999). More recently, genomics has been used to characterise virulence of *S. canis* with 34 candidate virulence genes detected (Richards et al., 2012). Most of these virulence genes constitute part of the *S. pyogenes* pangenome and have been implicated in tissue invasion. The carriage of *slo* and *emm* homologous genes, already described by De Winter et al. (DeWinter et al., 1999), was confirmed in *S. canis*. While an orthologue for *S. pyogenes* exotoxin SLS was identified, no genes encoding pyrogenic exotoxins (i.e. those responsible for *S. pyogenes*-associated toxic shock syndrome) were found, suggesting alternative mechanisms in the pathogenesis of *S. canis*. Some similarity with *S. agalactiae* and *S. pneumoniae* virulence genes, such as those encoding CAMP factor and neuraminidase B, was also found in the *S. canis* genome analysed (Richards et al., 2012). Components of the arginine deiminase system (ADS) have been characterised in the *S. canis* genome, giving insights into a metabolic pathway that could have a role in colonisation and disease (Hitzmann et al., 2013). ADS has been shown to be involved in virulence of *Streptococcus suis* (Fulde et al., 2011b). Three enzymes of the *S. canis* ADS are localised on the cell surface, with possible implications for its virulence, so further investigation is warranted (Hitzmann et al., 2013). The ability of *S. canis* to invade host cells was recently demonstrated (Yoshida et al., 2021). In *S. pyogenes*, cell invasion ability (CIA) is mediated by surface proteins such as fibronectin-binding proteins (FBPs) (Walker et al., 2014). The presence of genes with homology to *S. pyogenes* FBPs in the genome of *S. canis* has been shown together with experimental evidence of CIA in human and animal *S. canis* isolates (Yoshida et al., 2021). All 43 isolates tested showed intracellular invasion, but CIA was highly variable. Due to the lack of required clinical data, no link could be made between levels of CIA and disease severity (Yoshida et al., 2021) and the role of CIA in *S. canis* pathogenesis, thus, remains unknown. The most extensively studied virulence factor of *S. canis* is the M-like protein SCM (Fulde et al., 2011a). Experimental evidence showed that the *S. canis* SCM protein binds to plasminogen of humans, pigs, goats, cats and dogs. Interaction with plasminogen facilitates bacterial adherence and tissue invasion, the latter occurring through fibrinogen and fibrin degradation (Fulde et al., 2011a). SCM was also shown to cooperate in plasminogen recruitment with another surface-expressed virulence factor, enolase, and to have anti-phagocytic activity (Fulde et al., 2013). The *scm* gene has been confirmed as universally present in the

S. canis population, although with substantial allelic variation (Pinho et al., 2019). In particular, some *scm* variants lack the putative IgG binding domain which is thought to contribute to the anti-phagocytic activity of SCM (Bergmann et al., 2017; Pinho et al., 2019). More-

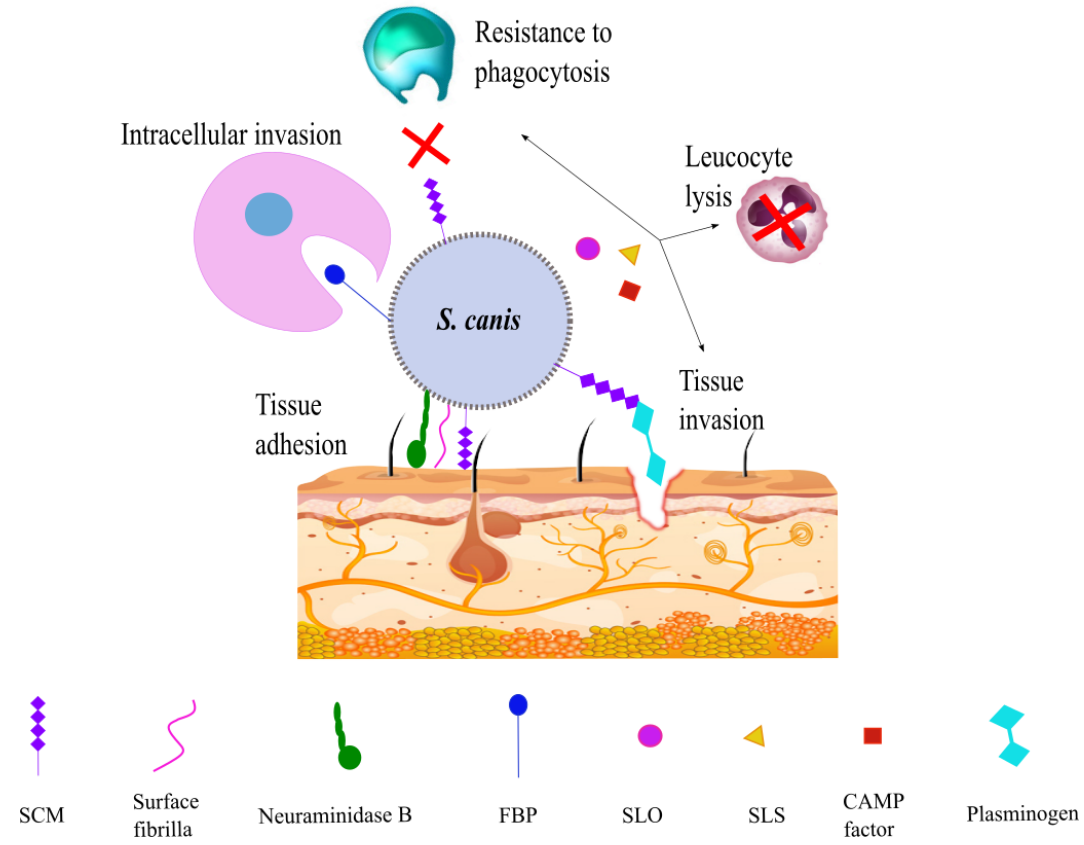


Figure 1.6: Virulence factors of *Streptococcus canis* and their role in pathogenicity. Although the expression and function of the *S. canis* M-like protein (SCM), surface fibrillae and fibronectin-binding protein (FBP) is supported by experimental evidence, the expression and activity of neuraminidase B, streptolysin O (SLO), streptolysin S (SLS) and Christie–Atkins–Munch-Peterson (CAMP) factor is inferred from knowledge of other pathogenic streptococci. Tissue adhesion is understood to be facilitated by SCM, surface fibrillae and, potentially, neuraminidase B. SCM also prevents phagocytosis, a process which may also be impeded by SLO, SLS and CAMP factor. The third main virulence activity of SCM appears to be tissue invasion mediated by plasminogen binding and activation. Tissue invasion may also be facilitated by SLO, SLS and CAMP factor, which are known to possess lytic activity towards leucocytes in other pathogenic streptococci. Finally, experimental evidence shows that FBP can trigger intracellular invasion of *S. canis*, facilitating bacterial survival.

over, some *scm* alleles are associated with lower binding affinity to plasminogen than others (Fulde et al., 2013; Pinho et al., 2019). Although according to some studies SCM appears to

be linked to *S. canis* virulence (Fulde et al., 2011a, 2013), recent findings, based on comparisons between a wildtype strain and an SCM-deficient mutant, questioned the role of SCM in clinical infection (Cornax et al., 2021). The SCM-deficient mutant showed reduced ability to form biofilms compared to the wildtype, but haemolytic activity and survivability in the presence of aminising and oxidising agents were not impacted by the lack of *scm*. There was no effect on survival after exposure to canine macrophages, human neutrophils and human whole blood or the ability to induce an immune response through cytokine production from human monocytes. When tested *in vivo*, the wildtype strain and the mutant were equally virulent in mouse models of dermal and systemic infection. The SCM-deficient strain, however, showed reduced adhesion and persistence in a murine model of vaginal colonisation when compared to the wildtype, suggesting that SCM might confer fitness advantages in particular anatomical sites (Cornax et al., 2021). Overall, the role of SCM as a virulence factor in *S. canis* is unclear, with recent evidence suggesting a marginal involvement in disease progression. A correlation between molecular characteristics of bacterial strains and clinical outcome of infection has not yet been shown for *S. canis*. Evidence based on limited numbers of isolates from dogs with toxic shock syndrome and/or necrotising fasciitis suggested that there was no specific genotype associated with severe disease in dogs (DeWinter and Prescott, 1999). Another study, which included more isolates from dogs and cats also failed to demonstrate a connection (Kruger et al., 2010). Further studies, using larger sample sizes and high-resolution genotyping are required to clarify the association between molecular characteristics of *S. canis* strains and clinical disease. A visual summary of *S. canis* virulence factors is provided in Figure 1.6.

1.3.3 *S. canis* genotyping systems

The MLST scheme developed for *S. canis* is based on allelic variation of seven housekeeping genes, namely *gki*, *gtr*, *murI*, *mutS*, *recP*, *xpt* and *yqiZ* (Pinho et al., 2013). As alluded to in the previous section, there is no evidence that links STs to specific clinical disease manifestation. Only a single study suggested an association between ST-4, 8, 11, 12, 13, 14, 17, 27 and 38, all belonging to clonal complex 13 (CC-13), and canine ulcerative keratitis (Enache et al., 2020) but this is based on a small number of cases and is not statistically supported. With regards to STs and species-specificity, it was shown that isolates sharing

the same ST may be isolated from multiple species, including humans, companion animals and wildlife (Pinho et al., 2013). A genotyping system based on allelic diversity of the *scm* gene has been proposed (Pinho et al., 2019). According to this scheme, 41 allelic variants are grouped into 12 SCM types, forming two major groups. Group I SCM variants (SCM types 1-7) have an IgG binding domain and are most commonly isolated from diseased patients. Group II SCM proteins (SCM type 8-12) lack this domain, which is thought to have anti-phagocytic activity, and the role of group II SCM in pathogenesis is not yet fully understood. MLST has been shown to be a good predictor of SCM type, although the converse is not true (Pinho et al., 2019). Fukushima et al. suggested an alternative SCM-based typing scheme, which currently encompasses 15 types (Fukushima et al., 2018, 2020a). Based on this scheme, SCM types 1-9 are classified as group I (corresponding to group I in the scheme by Pinho et al.) and types 10-15 are classified as group II (group II also for Pinho et al.). The creator of this scheme suggests that SCM group I strains are more commonly isolated in Japan (Fukushima et al., 2020a). As with the scheme of Pinho et al., MLST was shown to be a good predictor of SCM type, although, again, the opposite was not the case. Notably, a significantly higher prevalence of macrolide/lincosamide genetic resistance determinants and fluoroquinolone-resistant phenotype was detected among group I compared to group II strains (Fukushima et al., 2020a). Recently, an association was found between high-frequency CIA and Fukushima SCM types 10 and 11, as well as high-frequency CIA and STs 21 and 41 (Yoshida et al., 2021). It should be noted, however, that a limited number of isolates were tested (n=40) and therefore the resulting low frequency or absence of some SCM types and STs might have been a source of bias. Moreover, the threshold value used to separate low-frequency from high-frequency CIA isolates was arbitrarily chosen with the CIA value for almost one fifth of the isolates tested was just above or just below the threshold value (Yoshida et al., 2021). It remains uncertain, therefore, whether an association exists between CIA and specific strains of *S. canis*. A third SCM-based classification scheme has been described by Timoney et al. (Timoney et al., 2017). Four SCM types were detected among *S. canis* isolates (n=25) from healthy and diseased cats. SCM type 1 strains were most commonly derived from diseased cats, while SCM type 4 strains were almost exclusively isolated from healthy individuals. The authors concluded that type 1 strains were strongly associated with disease and that type 4 strains were avirulent in cats. However, type

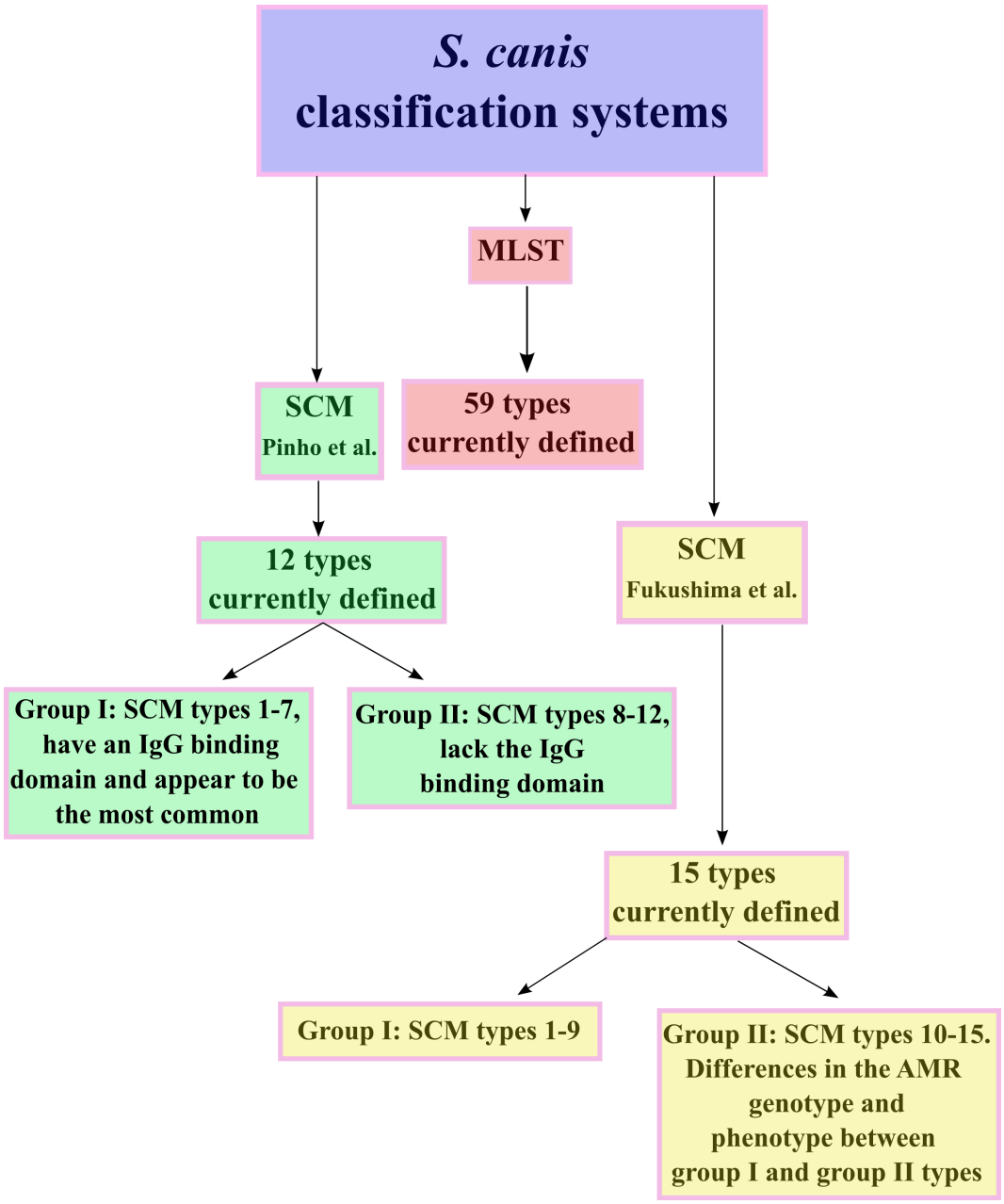


Figure 1.7: Schematic representation of the three main classification systems proposed for *Streptococcus canis* (Pinho et al., 2013; Pinho et al., 2019; Fukushima et al., 2020a).

1 strains were also isolated from healthy cats and one type 4 strain was implicated in a case of bacteraemia, suggesting both types can be associated with either clinical disease or asymptomatic carriage (Timoney et al., 2017). Figure 1.7 summarises the three main genotyping schemes proposed for *S. canis*.

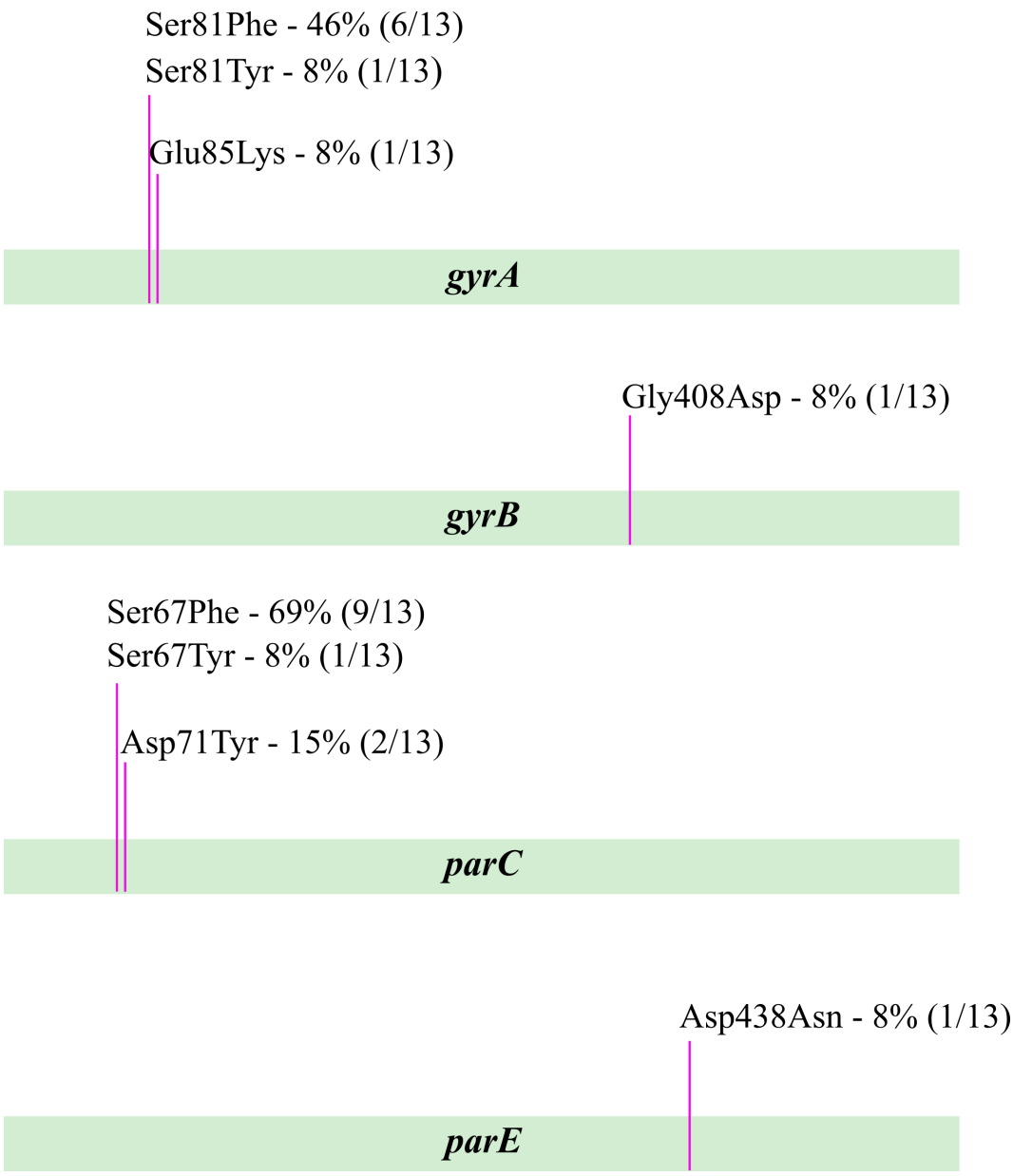


Figure 1.8: Amino acid substitutions observed in the quinolone resistance-determining regions regions of *gyrA*, *gyrB*, *parC* and *parE* in thirteen fluoroquinolone-resistant isolates of *Streptococcus canis*. Percentages and fractions represent the proportion of fluoroquinolone-resistant isolates carrying that mutation. Fluoroquinolone resistance was confirmed when the minimum inhibitory concentration for Levofloxacin by Etest was >1µg/mL (Fukushima et al., 2020b).

1.3.4 *S. canis* antimicrobial susceptibility

Streptococcus canis infections are successfully treated with ampicillin, amoxicillin and clavulanic acid or vancomycin in human medicine and amoxicillin and clavulanic acid or penicillin in veterinary medicine (Takeda et al., 2001; Tikofsky and Zadoks, 2005; Lam et al.,

2007; Lysková et al., 2007b; Pinho et al., 2013; Lacave et al., 2016; Tarabichi et al., 2018). The most commonly encountered AMR phenotype among *S. canis* strains is tetracycline resistance, which is expressed by 30-40% of all the isolates and associated with the carriage of *tet(M)*, *tet(O)*, *tet(S)*, *tet(K)* and *tet(L)* genes (Galpérine et al., 2007; Lysková et al., 2007a; Pinho et al., 2013; Fukushima et al., 2020b; Yoshida et al., 2021). Although less frequent, macrolide, lincosamide and streptogramin B (MLSB) resistance phenotypes have been detected in *S. canis* strains, particularly in association with the presence of the *erm(A)*, *erm(B)*, *mef(A)* and *aadA* genes (Galpérine et al., 2007; Lysková et al., 2007a; Pinho et al., 2013; Fukushima et al., 2020b; Yoshida et al., 2021). Occasional resistance to gentamicin and rifampicin has also been reported in *S. canis* (Galpérine et al., 2007). The occurrence of fluoroquinolone resistance associated with specific amino acid substitutions in the QRDR of the *gyrA*, *gyrB*, *parC* and *parE* genes has recently been documented in a small number of resistant strains (Figure 1.8) (Fukushima et al., 2020b).

1.3.5 *S. canis* zoonotic potential

The ability of *S. canis* to colonise and cause disease in a variety of mammals is well documented (Richards et al., 2012). Human infections are understood to be rare, although there has recently been an increase in reported cases (Takeda et al., 2001; Galpérine et al., 2007; Lam et al., 2007; Lacave et al., 2016; Tan et al., 2016; McGuire et al., 2021), and little is known about epidemiology in humans. Since dogs and cats are recognised as the main host species of *S. canis*, it is likely that human infection can result from direct 'pet-to-people' transmission, making *S. canis* a potentially zoonotic pathogen (Richards et al., 2012). This hypothesis has been supported by reports of human infections following dog bites and other forms of interaction with companion animals (Bert and Lambert-Zechovsky, 1997; Takeda et al., 2001; Lam et al., 2007). It remains unclear, however, whether all *S. canis* strains possess the same multi-species tropism profile or whether adaptation has occurred. From a molecular point of view, preliminary evidence suggesting a lack of host adaptation of *S. canis* strains has been provided through MLST, with strains of the same ST found in both animals and humans (Pinho et al., 2013, 2019), inferring zoonotic potential. However, it may be argued that MLST fails to represent accurately the diversity of bacterial populations when compared to more discriminatory genomic methods (Tsang et al., 2017). Better evidence is

required to aid our understanding of the epidemiology of *S. canis* and provide insight into public health risks.

1.4 Data visualisation in public health

Data visualisation, as the name suggests, is the graphical display of data. Data visualisation is a multidisciplinary field that encompasses areas of graphic design, statistics, psychology and computer science (Aparicio and Costa, 2015). Its purpose is to communicate data in a more direct and more accessible way than a section of text (Sadiku et al., 2016). In order to convey information through a figure, data first need to be encoded in the design of the figure and then the information correctly interpreted by the viewer; this process is also known as graphical perception (Cleveland and McGill, 1986). Cleveland and McGill were among the first to recognise the need to approach data visualisation in a scientific way, studying the accuracy of data interpretation by people presented with different forms of graphs (Cleveland and McGill, 1986). Their experiments allowed them to formulate a paradigm for graphical perception (Cleveland and McGill, 1987). According to the Cleveland and McGill paradigm, ten elementary codes of graphs, which are basic judgements that people make to gather information from a graph, are identified and ranked based on the accuracy with which people judge them (Table 1.9) (Cleveland and McGill, 1987). Cleveland and McGill demonstrated that the choice of graph has critical implications in the way data are shared and interpreted. In other words, data visualisation is not only a matter of graphic design but also, and most importantly, of accurate communication.

Table 1.9: Basic judgements that people make to extract quantitative information from graphs (elementary codes) ranked on the basis of their accuracy in communicating data (Cleveland & McGill 1987).

Elementary codes	Rank
Positions along a common scale	1
Positions along identical, nonaligned scales	2
Lengths	3
Angles	4

Elementary codes	Rank
Slopes	4-10
Areas	6
Volumes	7
Densities	8
Colour saturations	9
Colour hues	10

1.4.1 Historical milestones in data visualisation

The first visual representation of data is found in the history of the ancient Egyptians, who by 200 B.C. were using coordinate systems to draw terrestrial and celestial maps (Friendly, 2008). Other significant examples of some of the earliest recorded data visualisations are Ptolemy's map of a spherical earth (1^{st}-2^{nd} century A.D.) and a series of multiple graphs showing the movement of seven celestial bodies drawn by an anonymous individual in the 10^{th} century (Friendly, 2008). The first attempt to plot a mathematical function was made in the 14^{th} century by Nicole Oresme (Clagett et al., 1968), while one of the earliest graphical representations of sunspots changing over time appeared in the 17^{th} century by Christopher Scheiner (Scheiner, 1630). The 17^{th} century represents the beginning of the age of enlightenment and at this time the first visual representation of statistical data was created in the form of a line graph showing the estimated difference in longitude between Toledo and Rome, by Michael Florent van Langren (Tufte and Robins, 1997). In the same century another important contribution to the data visualisation science was made by Rene Descartes, who introduced the presentation of quantitative data in terms of two-dimensional coordinate scales (Descartes, 1637). The 17^{th} century also saw the creation of the first graph of a continuous distribution function, the first bivariate plot and the first weather map (Friendly, 2008). Contour and topographic maps were introduced by Phillippe Buache and Marcellin du Carla-Boniface the following century (Buache, 1752; du Carla-Boniface, 1782), together with timelines by Jacques Barbeu-Dubourg (Friendly, 2008). Line graphs, bar charts, pie charts and circle graphs were invented by William Playfair at the same time to communicate

socio-economic data (Playfair, 1801a,b). In the 19th century geological maps were first used by William Smith (Smith, 1815) and one of the first modern-style thematic statistical maps was drawn by Charles Dupin, who used continuous shading to indicate different levels of illiteracy throughout France (Friendly, 2008). In 1855 John Snow mapped fatalities linked to the cholera outbreak of 1854 in the Broad Street region of London (Figure 1.9), discovering that the disease was water-borne and originated from the Broad Street water pump (Snow, 1855; Azzam et al., 2013). In 1858 Florence Nightingale published a report on the various causes of death of British soldiers fighting in the Crimean war, using several graphs and tables to show that poor hygiene was responsible for more fatalities than the battlefield (O'Connor et al., 2020). Her graphs helped to convince the British government to improve battlefield hospital hygiene and reduce soldier mortality (O'Connor et al., 2020).

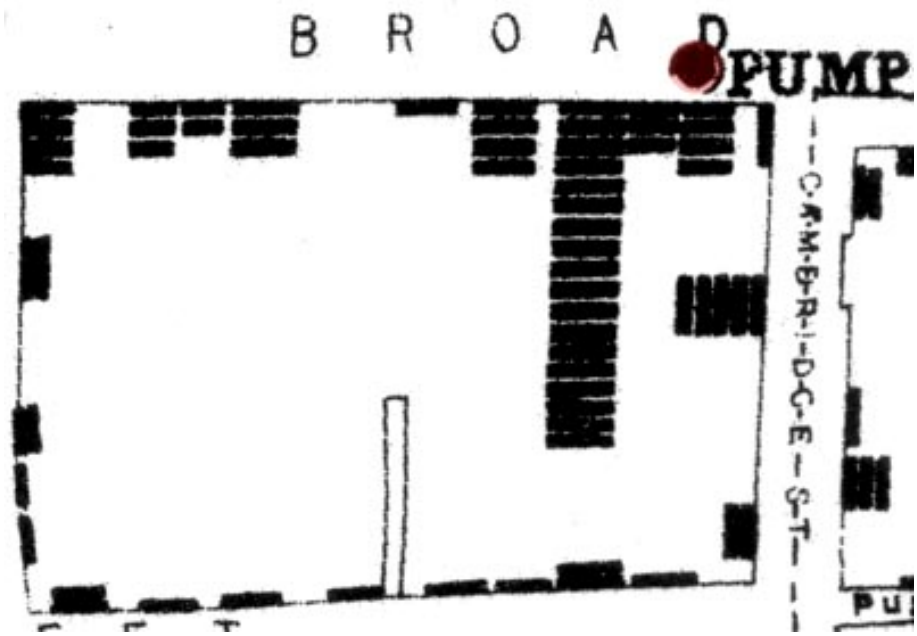


Figure 1.9: John Snow's map of cholera-associated deaths in the Broad Street area of London during the 1854 cholera outbreak. From Snow, 1855.

In the 20th century, statistical graphics started to be widely used in many fields such as education, commerce, science and politics (Friendly, 2008). The rising interest in data visualisation at that time is evidenced by the works of Willard Cope Brinton, John Tukey and Jacques Bertin, who contributed to advancing the field of visual data representation (Brinton, 1939; Tukey, 1962; Bertin and Barbut, 1967). During the second half of the 20th century, attention was given to the theoretical concepts surrounding the visual display of data, as

demonstrated by the work of William Cleveland and Robert McGill (Cleveland and McGill, 1987). At the same time, the development and mass distribution of personal computers started a graphic revolution, offering countless new tools and opportunities to analyse and visualise data (Friendly, 2008).

1.4.2 Common visualisation techniques

Among the countless ways data can be graphically rendered, there are some common and easily recognisable visualisation techniques, exemplars of which are shown in Figure 1.10. Bar charts are used to compare quantities by length, resulting in an accurate form of graphical representation according to the Cleveland and McGill paradigm (Cleveland and McGill, 1987). Simple bar charts are effective for comparing values of a single variable across several items, while stacked bar charts are useful for comparing the contribution of multiple variables across several items (Streit and Gehlenborg, 2014). Four experiments by Talbot et al., building on the previous findings of Cleveland and McGill, revealed some elements that can limit interpretation accuracy when using bar charts (Talbot et al., 2014). Firstly, comparing non-adjacent bars in simple bar charts is more difficult than comparing adjacent bars, particularly if short bars are involved. Secondly, when dealing with stacked bar charts, distractors (such as small dots used to mark the bars) make it more difficult to interpret the graph. Finally, in stacked bar charts, accurate comparison between adjacent bars is more challenging than that between non-adjacent bars.

Line graphs show how one variable changes in relation to another (Sadiku et al., 2016). They are frequently used to display trend changes over time. Line graphs rely on slope judgements, which reveal the rate of change of the variable y as a function of the variable x (Cleveland and McGill, 1987). As highlighted by Cleveland and McGill, the accuracy of slope judgements can be biased by the "shape parameter", which is the slope of the line connecting opposite corners of a graph (Cleveland and McGill, 1987). The angle that this line forms with the lower side of the graph, which can be referred to as "mid-angle", depends on the shape parameter. Based on experimental evidence, the maximum accuracy of slope judgements is achieved with shape parameters associated with a mid-angle of roughly 41° (Cleveland and McGill, 1987).

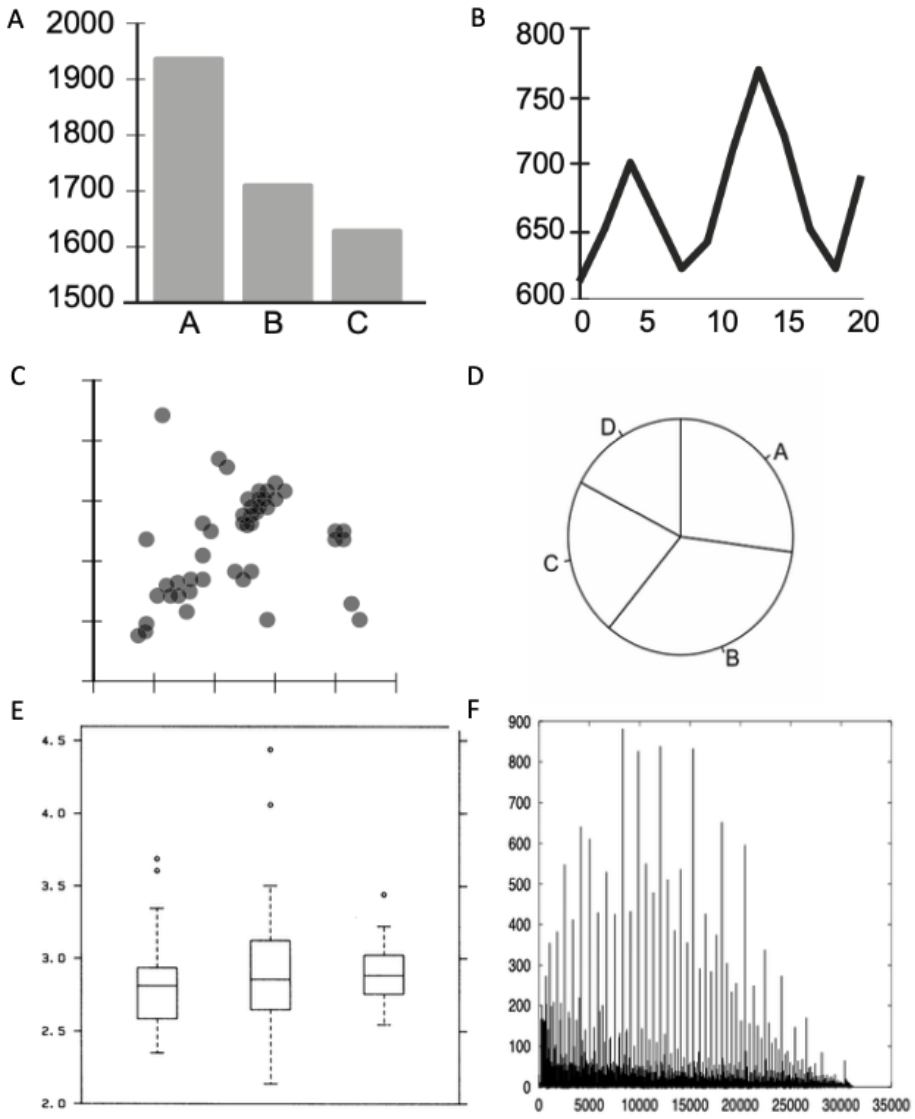


Figure 1.10: Examples of common visualisation techniques. A-bar chart, B-line graph, C-scatter plot (Kelleher & Wagener, 2011), D-pie chart (Siirtola, 2019), E-box plot (Cleveland & McGill, 1985) and F-histogram (Jagadish et al.,1998).

In a scatter plot, each point represents the value of two variables positioned on two continuous, orthogonal dimensions (x- and y-axis) (Sarikaya and Gleicher, 2017). The main purpose of scatter plots is to display the spatial distribution of data, which is plotted in two dimensions. One of the most common problems associated with scatter plots is the overlap of dots in large data-sets, a phenomenon known as overdraw (Sarikaya and Gleicher, 2017). Suggested ways to fix overdrawn scatter plots include downsizing the plotted data set (Bertini and Santucci, 2006; Chen et al., 2014), simplifying the visual representation

(e.g. continuous density estimations) (Tory et al., 2007; Collins et al., 2009; Mayorga and Gleicher, 2013) or modifying the space of the plot (Sarikaya and Gleicher, 2017). The use of different shapes and colours to distinguish two data sets within the same graph can also influence the interpretation of a scatter plot in either a positive or a negative way (Gleicher et al., 2013; Gramazio et al., 2014; Elliott and Rensink, 2015). The distribution of the plotted data should be considered also when drawing a scatter plot (Sarikaya and Gleicher, 2017).

Pie charts are used to represent the component parts of a whole (Sadiku et al., 2016). Although very popular, pie charts have been criticised by several experts for being less accurate than other visualisation techniques (Tufte, 1985; Few and Edge, 2007). The data encoded in a pie chart can be decoded differently based on the element of the chart on which the reader focuses, namely angle, length of arc and area of segment (Siirtola, 2019). Moreover, the interpretation of pie charts requires angle and area judgements, which are known to be less accurate than position and length judgements (Cleveland and McGill, 1987). Experimental evidence supports the notion that pie chart interpretation takes longer and is less accurate than that of stacked bar charts (Siirtola, 2019).

Box plots allow the comparison of the distributions of groups of measurements of a variable (Cleveland and McGill, 1985). Each measurement group is represented as a box with an upper and lower whisker (Streit and Gehlenborg, 2014). The box ranges from the first (Q1) to the third (Q3) quartile and thus represents the interquartile range (IQR). Quartiles are respectively the 25^{th} and 75^{th} percentiles of the sample and cover the central 50% of the data (Streit and Gehlenborg, 2014). A line within each box indicates the median value of the sample (Streit and Gehlenborg, 2014). The box whiskers extend to Q1 – 1.5 × IQR and Q3 + 1.5 × IQR (Streit and Gehlenborg, 2014). Values that fall outside this range are considered outliers and may be plotted as individual marks (Krzywinski and Altman, 2014). As opposed to other types of graphs, box plots are not heavily influenced by the presence of outliers (Krzywinski and Altman, 2014). Although they are considered an accurate representation of sample distribution by experts, box plots are not intuitively accessible and require some degree of familiarity to be interpreted correctly (McGill et al., 1978; Pierce and Chick, 2013).

Histograms are used to visualise the frequency distribution of a variable's measurements, and were thus defined by David W. Scott as "the classical nonparametric density estimator" (Scott, 1979). The histogram of a variable is constructed by dividing the data distribution of that variable into discrete mutually disjointed subsets (buckets) and approximating the frequencies and values in each subset in some common fashion (Ioannidis, 2003). As this definition suggests, histograms rely on the approximation of frequencies and so always contain an error due to the information discarded when the data is summarised (Jagadish et al., 1998). Histograms are then more appropriate for quickly communicating accessible data estimates rather than precise measurements (Jagadish et al., 1998).

1.4.3 Infographics

The term infographics is an abbreviation for information graphics and it refers to a type of "graphic design that combines data visualisations, illustrations, text, and images together into a format that tells a complete story" (Krum, 2013). Scientific evidence suggests that the visual representation of data and other messages facilitates our ability to understand and remember them, as well as to make decisions that involve the newly acquired information (Spiegelhalter et al., 2011; Clark and Mayer, 2016). Because of their combination of images, colours and graphs, infographics are considered to be an engaging and successful method of summarising and communicating complex messages to a wide audience (McCrorie et al., 2016). The versatility and effectiveness of infographics make them suitable for several disciplines that rely on the communication of complex information, such as journalism, education and healthcare (Smiciklas, 2012; Otten et al., 2015; McCrorie et al., 2016).

In the public health sector, infographics are commonly used as a part of awareness campaigns to make otherwise complex messages accessible to the general public. Examples of successful infographics applied to the public health field include those designed by the WHO for the World Antibiotic Awareness Week to educate the population and healthcare workers about antibiotic misuse and the risk of spreading AMR (`https://www.who.int/campaigns/world-antimicrobial-awareness-week`). Figure 1.11, for instance, was used as the main poster for the WHO World Antibiotic Awareness Week in 2018 and summarises, in a concise but memorable way, the need to address the worldwide

AMR crisis currently being faced.

Figure 1.11: Global action plan poster designed by the World Health Organisation for the World Antimicrobial Awareness Week in 2018. From `https://www.who.int/campaigns/world-antimicrobial-awareness-week/2018/advocacy-material`.

1.4.4 Graphic design considerations

The first consideration to make when designing a graphic is its purpose. Graphics used by researchers to investigate the characteristics of a data set, also called exploratory graphics, are

constantly changing and are generally not intended for sharing with a wide audience. Presentation graphics, on the other hand, are made to communicate a message to an audience, often cannot be changed once released and require appropriate design choices in order to be accurately interpreted by the target audience (Chen et al., 2007). Once the purpose and target audience have been clarified, attention should be paid to the data-set itself. The graphical style selected depends mainly on the characteristics of the variable(s) to plot. Single continuous variables, for example, should be visualised with box plots and histograms instead of bar charts and pie charts, the latter two being more appropriate for single categorical variables (Chen et al., 2007). After the visualisation form has been decided, guidelines for implementing the chosen graphic should be consulted (Kelleher and Wagener, 2011). Consideration should be given, for example, to the scales of the axes of a graph. For a categorical variable, defining the axis scale is relatively simple, whereas for a continuous variable, selecting the best scale can be much more challenging (Wilkinson, 2012). Attention to the axes scale is particularly important when two similar graphs need to be compared, since the use of an identical scale for both increases interpretation accuracy (Chen et al., 2007). Annotations, such as sloping and fitted lines, can also facilitate the interpretation of a graph, although they should be added sparingly to avoid a cluttered display (Robbins, 2012). The same considerations apply to the axes labels. The caption should be both informative and concise, to avoid discouraging the reader from engaging with the figure (Robbins, 2012). If a legend is required to decode the graph, it should be added directly to the plot to facilitate interpretation (Tufte, 1985). Other graphic properties that should be considered include size, frame, aspect ratio and colour choice (Chen et al., 2014), the latter potentially being a particularly powerful element in a graph. An appropriate colour scheme should be selected, keeping in mind that some readers may be colour blind and that some colours are culturally associated with specific meanings (Chen et al., 2014). Finally, the overall appearance of a graphic should be kept as simple as possible in order to avoid potential distractions and aid interpretation, according to some experts (Tufte, 1985). Some recent and limited experimental evidence, however, points at a different direction, suggesting that visual embellishments can increase long term memorability of graphics without impacting the accuracy of interpretation of the data displayed (Bateman et al., 2010). It should be noted, though, that adding visual imagery to graphics requires experience and if done incorrectly may hamper the readability of a figure

(Bateman et al., 2010).

1.4.5 Data visualisation in the public health sector

Public health was defined as "the art and science of preventing disease, prolonging life and promoting health through the organised efforts of society" (Acheson et al., 1988). This broad term encompasses several branches of the healthcare system and applies to society as a whole. The three core functions of public health, as defined by the Institute of Medicine in 1988, are assessment, policy development and assurance (Committee for the Study of the Future of Public Health, 1988). All three core functions benefit from the effective visual display of data. In particular, the need to communicate concisely, and sometimes quickly, complex messages to the population or to policy-makers explains the central role that data visualisation has gained in the public health sector (McCrorie et al., 2016). The assessment function of public health, considered as the monitoring, diagnosis and investigation of health threats at a community level, was for instance supported by John Snow's map of cholera-associated deaths in London (Figure 1.9) as previously described. Remarkable visualisations

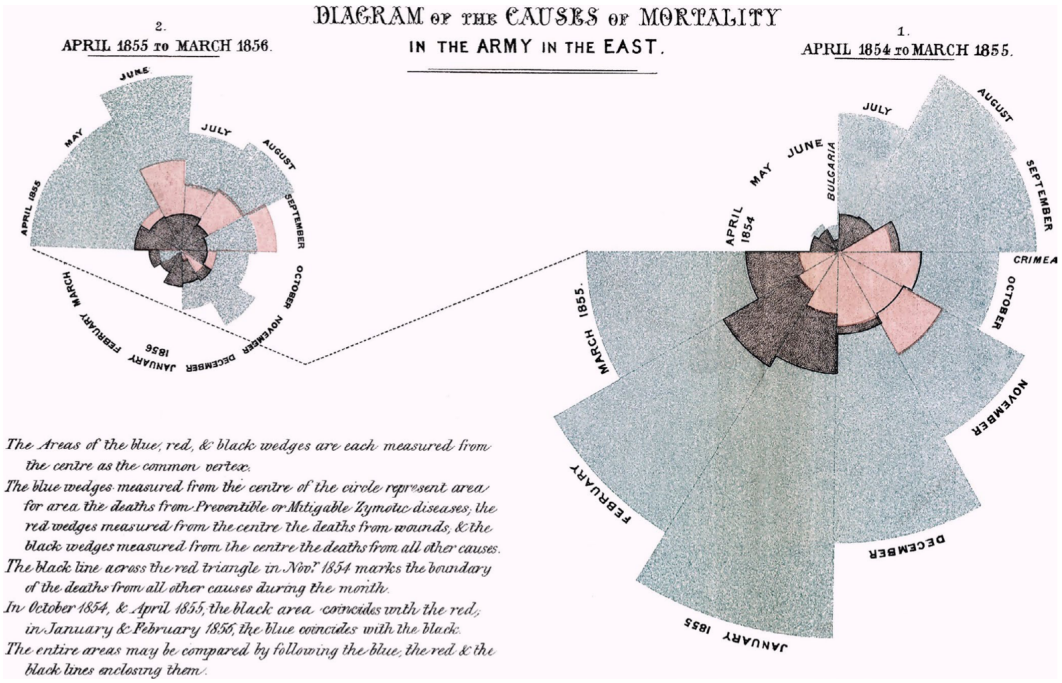


Figure 1.12: A visualisation created by Florence Nightingale to communicate data regarding the causes of death among British soldiers at war. From O'Connor et al. (2020).

that influenced the policy development function of public health in the 19th century were

made by Florence Nightingale, which are also considered one of the first examples of data visualisations applied to public health. The graphs and infographics, published by Nightingale in 1858, were used to inform the British government about the main causes of death for soldiers fighting in the Crimean war (O'Connor et al., 2020). Nightingale's work influenced the government of the day to take decisions that would reduce deaths by preventable causes among soldiers fighting abroad (O'Connor et al., 2020). An example of such visualisations is shown in Figure 1.12. The assurance function of public health, which focuses on support and education of the general public and prevention of health hazards, makes use of data visualisations, for instance, in awareness campaigns such as for annual World Antibiotic Awareness Week, promoted by the WHO (Figure 1.11).

Data visualisations are widely used in the public health sector to inform the public, healthcare workers and policy-makers about infectious disease threats. For infectious disease surveillance and prevention, commonly used approaches include graphs and infographics. These can be used to show the disease burden in the whole population or in different population groups, changing trends in disease incidence (`https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports`), spread of antimicrobial resistance in bacterial populations etc. Other types of visualisations used for infectious disease surveillance are maps, developed with the support of geographic information systems to track disease distribution (AvRuskin et al., 2004; Castronovo et al., 2009; Dominkovics et al., 2011). Dot maps, choropleth maps and isopleth maps, for instance, are commonly used to visualise the spatial distribution of disease cases (Blanton et al., 2006; Reinhardt et al., 2008; Anselin et al., 2010; Porcasi et al., 2012). Both choropleth and isopleth maps make use of colors to display disease density in different geographical areas. In isopleth maps, however, data are not grouped in pre-defined geographical units such as counties, districts, etc. Graphs, infographics and maps are visualisations familiar to every type of audience, hence they can be used to communicate infectious disease data at every societal level. Examples of such visualisations can be seen in WHO prevention and awareness campaigns, such as the Global Malaria Programme (`https://www.who.int/teams/global-malaria-programme`), and in gov-

ernmental websites such as the United States Centre for Disease Control and Prevention (`https://www.cdc.gov`).

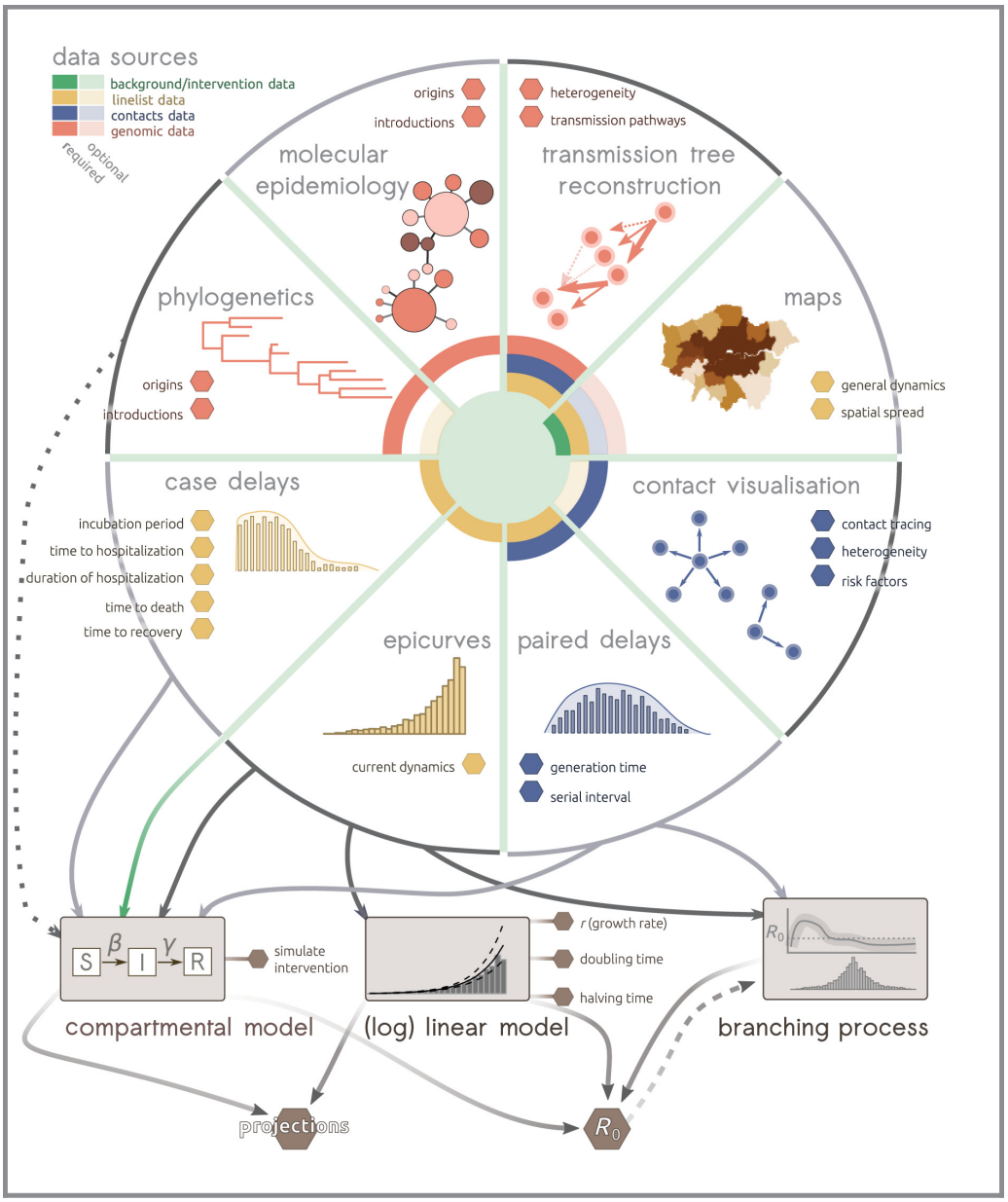


Figure 1.13: Example of outbreak analytics workflow. This schematic represents eight general analyses that can be performed from outbreak data. Outputs containing actionable information for the operations are represented as hexagons. Data needed for each analysis are represented as a different colour in the center, using plain and light shading for mandatory and optional data, respectively. From Polonsky et al. (2019).

Data visualisation also plays a central role during infectious disease outbreaks, serving not only as a communication method but also as an analytical tool. In this case, the target audiences are usually medical specialists and policy makers, and so the type of visualisations used may be more complex and less accessible. Some examples of visualisations involved in outbreak analysis and communication are epicurves, maps, social network diagrams, phylogenetic trees and other forms of pathogen clustering diagrams (Polonsky et al., 2019). The integration of these visualisations in an outbreak analytics workflow is represented in Figure 1.13. Together, these visualisations can provide insights into the aetiological agent involved and help monitor the dynamics of the infection, with a particular focus on where it is spreading and which individuals might contribute to further pathogen transmission (Polonsky et al., 2019).

1.4.6 Examples of the use of data visualisation in infectious disease surveillance and control

One of the earliest examples of visualisations used for infectious disease control is the mapping of cholera cases around Broad Street in London by Jon Snow in 1855 (Azzam et al., 2013). Since then, this has become an essential part of infectious disease surveillance and control, especially during outbreak investigation and management (Polonsky et al., 2019). In this section, some successful examples of the application of data visualisation to infectious disease surveillance and control are presented.

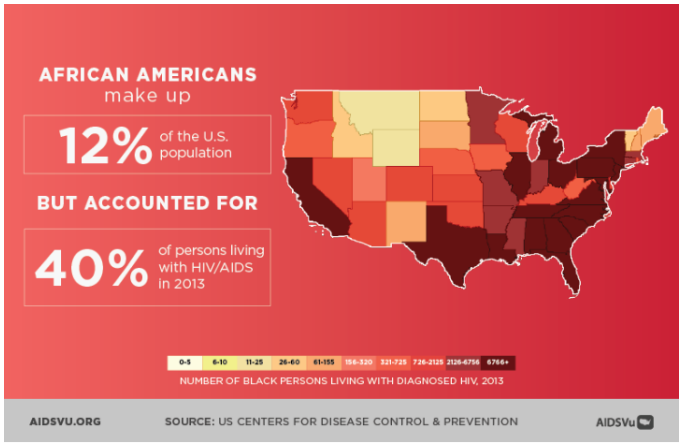


Figure 1.14: Infographic promoting awareness about the HIV/AIDS impact among African American in the United States in 2013. From Valdiserri and Sullivan (2018).

AIDSVu is an interactive map showing the impact of HIV on communities across the United States (https://aidsvu.org). The AIDSVu map, and further visualisations based on the same data have been used not only to raise awareness and educate people about the HIV epidemic in the United States but also to plan effective interventions to protect vulnerable groups of the population (Valdiserri and Sullivan, 2018). For instance, AIDSVu infographics, such as the one shown in Figure 1.14, were successfully used to inform and educate people about the threat of HIV spread when the founder of the Black AIDS Institute, Phill Wilson, shared them in his blog in 2017 (Valdiserri and Sullivan, 2018). From that platform, the infographics received considerable internet visibility, as witnessed by the high number of impressions they received in various social networks (Valdiserri and Sullivan, 2018). The AIDSVu map and data were also used by Rosenberg et al. and Breskin et al., who identified areas of the country where HIV is hyperendemic among men who have sex with men and where it is more prevalent in women than men (Rosenberg et al., 2016; Breskin et al., 2017). This kind of information can assist in the development of targeted interventions to support infected people and to reduce further disease spread.

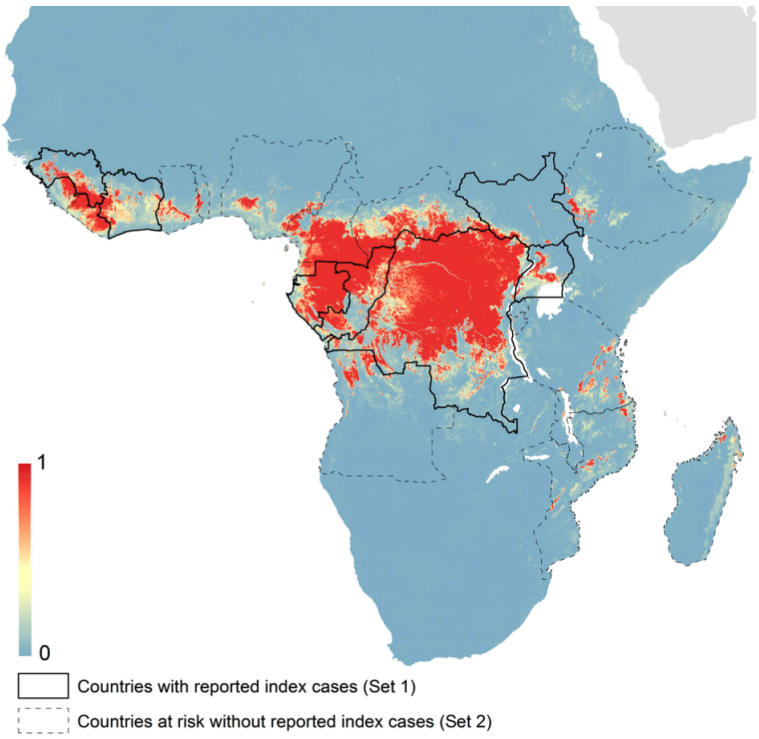


Figure 1.15: Predicted zoonotic transmission niche for Ebola virus. The scale represents the probability that transmission of Ebola virus from animals to humans occur at these locations. From Pigott et al. (2014).

An example of data visualisation applied to exploring the dynamics of infection spread can be found in the work of Pigott et al., who mapped the zoonotic transmission cases of Ebola virus in Africa from 1976 to 2014 (Pigott et al., 2014). Geographical data, together with reservoir species distribution models and environmental data, allowed the prediction of a zoonotic transmission niche for Ebola virus in Africa, as shown in Figure 1.15 (Pigott et al., 2014). In another work carried out by the same authors, the pandemic potential of viral haemorrhagic fever in Africa was assessed, mapping the index-case potential, the outbreak potential, the local epidemic potential and the global epidemic potential of Crimean-Congo haemorrhagic fever, Ebola virus disease, Lassa fever and Marburg virus disease (Pigott et al., 2017).

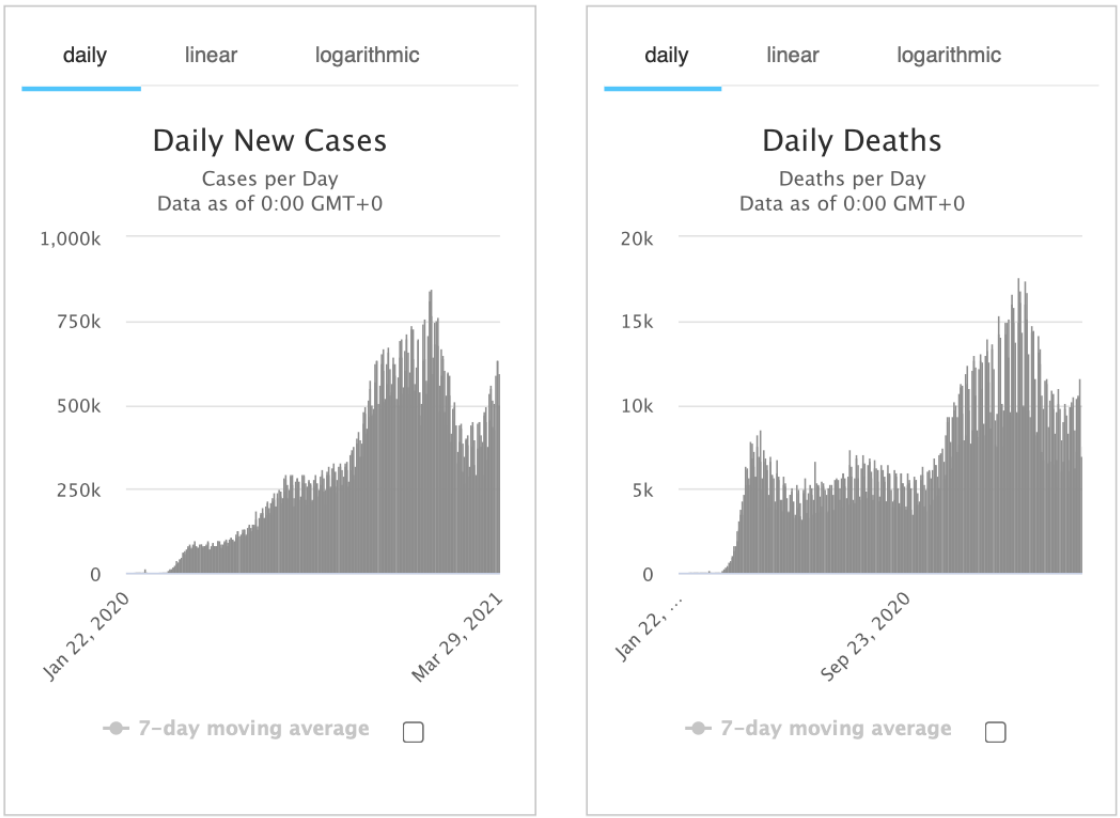


Figure 1.16: Histograms showing the daily new cases of COVID-19 and daily deaths associated with SARS-CoV-2 infection from January 2020 to March 2021 across the world. From the Worldometer website consulted on the 30/03/21 (`https://www.worldometers.info/coronavirus/`).

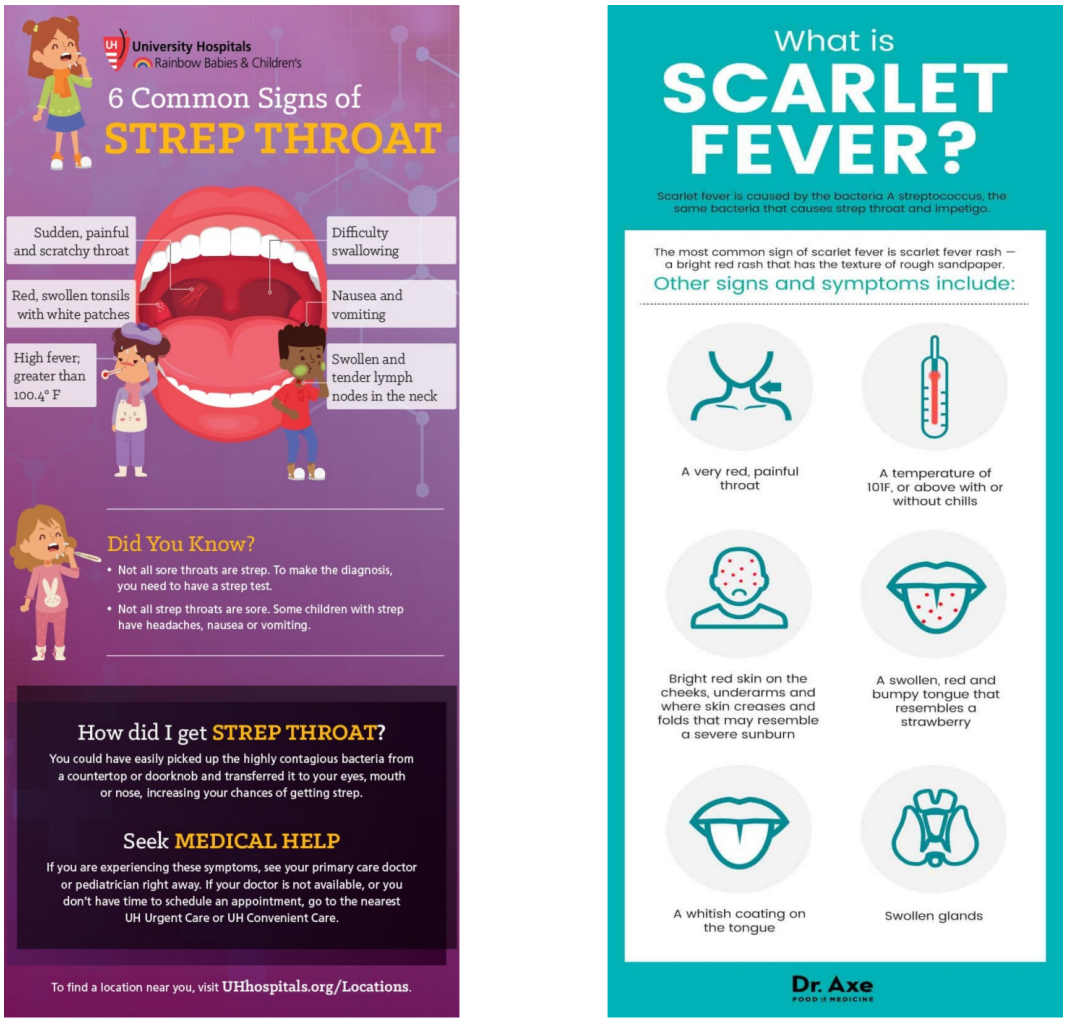
Data visualisation has played a central role in the COVID-19 pandemic. At an early stage of the outbreak, visualisations forming part of an exploratory data analysis on the

spread of SARS-CoV-2 inside and outside China were presented by Dey et al. (Dey et al., 2020). These visualisations offered a reliable source of information for the scientific community and policy makers allowing them to monitor the viral spread and take decisions on how to contend with the health crisis (Dey et al., 2020). Furthermore, the Worldometer website has shared epidemiological data on the global impact of COVID-19 in the form of different graphs and tables (e.g. Figure 1.16), becoming a popular platform to share information in near real-time (`https://www.worldometers.info/coronavirus/`). A similar website, offering real-time interactive visualisations on the spread of COVID-19 in Scotland, is travellingtabby (`https://www.travellingtabby.com/scotland-coronavirus-tracker/`).

1.4.7 Data visualisation in the epidemiology of Group A *Streptococcus*

Graphical visualisations are commonly used to communicate data relating to the epidemiology of GAS. The vast majority of these visualisations are designed for specialist audiences, such as the scientific research community, medical experts and policy makers. The lack of visualisations targeting the general public might be explained by the lower incidence of severe GAS disease compared to other common infections (Efstratiou and Lamagni, 2016). Another possible explanation is that the risk factors for GAS disease are relatively non-specific and associated with other bacterial infections (Lamagni et al., 2008a). GAS disease data visualisations designed for the general public are mainly in the form of infographics which aim to inform about the most common clinical manifestations associated with GAS infection, namely GAS tonsillitis (strep throat) and scarlet fever (Figure 1.17).

GAS infection rates over time and across different age groups are often displayed as line graphs and bar charts for specialist audiences (Lamagni et al., 2008a; Lepoutre et al., 2011; Efstratiou and Lamagni, 2016; Imöhl et al., 2017; Hammond-Collins et al., 2019). Bar charts are also used to show the distribution of GAS strains. In a publication by Steer et al., for example, the global disease burden attributable to different *emm* types is represented as a series of bar charts (Steer et al., 2009). In another work on the epidemiology of invasive GAS disease in Denmark during 2003 and 2004 the *emm* type and T type distribution are shown as



A B

Figure 1.17: Visualisations relating to Group A *Streptococcus* disease designed for the general public. A - from `https://www.uhhospitals.org/Healthy-at-UH/articles/2018/11/is-it-strep-throat-how-to-tell`; B - from `https://pl.pinterest.com/pin/516999232230804583/`.

bar charts and pie charts, respectively (Luca-Harari et al., 2008). Similarly, in a study on the distribution of macrolide resistance among invasive GAS isolates in Iceland from 1995 to 2016, the annual incidence of iGAS infection attributable to different erythromycin-resistant *emm* types was depicted as a stacked bar chart while the overall proportion of erythromycin-resistant *emm* types was represented as a pie chart (Figure 1.18) (Southon et al., 2020). In a study by Zachariadou et al., the total number and relative proportion of invasive disease cases caused by *emm* types 1 and 12 over a 5 year time-period in 17 Greek hospitals were plotted as a combined bar chart and line graph (Zachariadou et al., 2014). In GAS epidemi-

ological studies, maps are also utilised as seen, for example, in studies by Lamagni et al. and You et al., where gradient maps were used to display the different incidence of GAS infections in European countries and Chinese regions, respectively (Lamagni et al., 2008a; You et al., 2018). Although less common, network diagrams may be found in the literature of GAS epidemiology. For example, this may be observed in a work by Bubba et al., where all the cases of a GAS invasive disease outbreak were shown as being connected based on interactions among patients before symptoms appeared (Bubba et al., 2019).

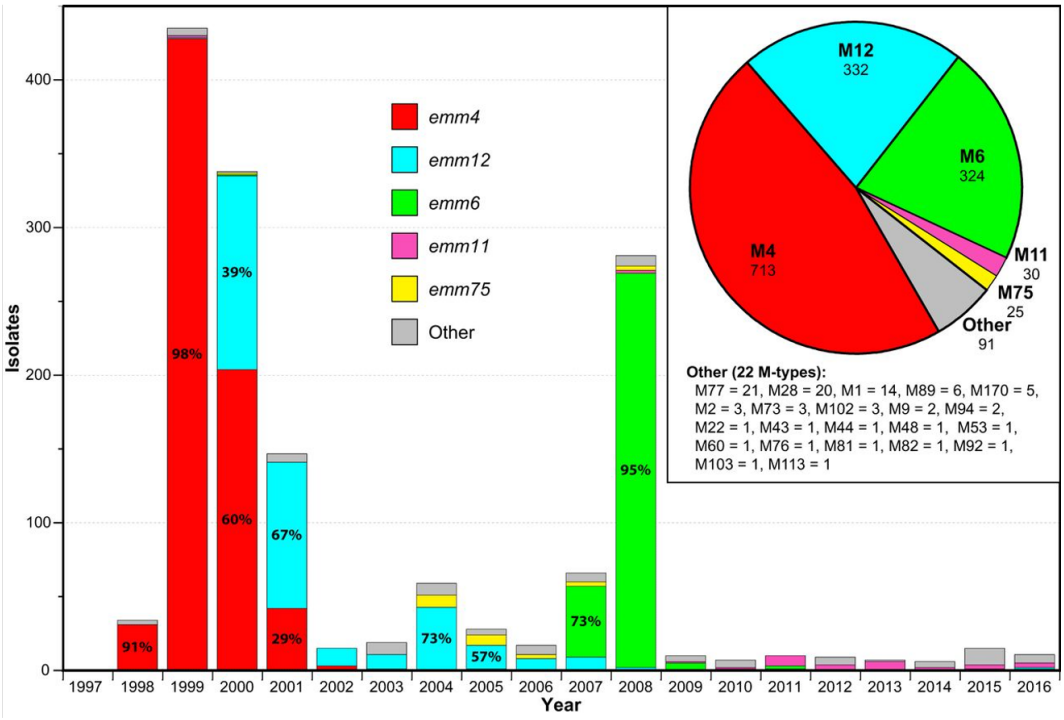


Figure 1.18: Annual invasive disease incidence and proportion of *Streptococcus pyogenes emm* types of erythromycin-resistant isolates in Iceland from 1995 to 2016. Isolate proportions are coloured by *emm* type. From Southon et al. (2020).

In the context of GAS molecular epidemiology, phylogenetic trees are often used to display the genetic relationships among invasive isolates. These trees can appear in different formats, such as linear rooted (Engelthaler et al., 2016), circular rooted (Davies et al., 2019) or unrooted trees (Beres et al., 2016), with colour occasionally being used to identify groups of isolates sharing specific features (Beres et al., 2016; Southon et al., 2020). Rooted phylogenetic trees, an example of which can be seen in Figure 1.19, may be familiar to many medically-informed viewers and appear easy to interpret but they imply the existence of a

single common ancestor (i.e. root of the tree), which is theoretical in nature and does not necessarily relate to any of the bacterial isolates being analysed. For this reason they can be more misleading than unrooted trees, which are less intuitive at first (Figure 1.20). Some GAS phylogenies utilise different colours and shapes for the leaves to display more than one attribute of the isolates examined (Kachroo et al., 2019). Combinations of phylogenetic trees and heat maps have been used to show the genetic relatedness and the distribution of specific traits (e.g. mutations at specific sites) among GAS strains (Le Breton et al., 2015; Kachroo et al., 2019). Histograms have also been used to communicate aspects of the molecular epidemiology of GAS, in particular to display the frequency of mutations occurring across specific genomic regions (Davies et al., 2019; Kachroo et al., 2019; Nasser et al., 2014). Scatter plots and dot matrix plots have also been employed to visualise molecular characteristics of GAS isolates (Nakagawa et al., 2003; Beres et al., 2017; Kachroo et al., 2019). More complex visualisations, such as three dimensional principal component analysis (PCA) plots, are less frequently utilised in the molecular epidemiology of GAS (Kachroo et al., 2019).

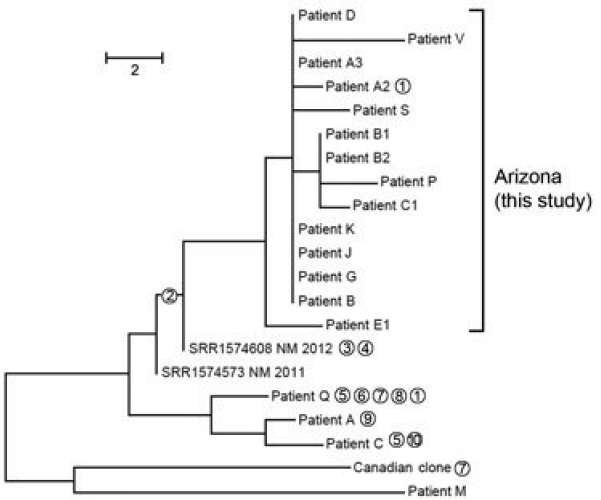


Figure 1.19: Example of a rooted tree used to display genomic relatedness of Group A *Streptococcus* isolates. From Engelthaler et al. (2016).

1.4.8 Data visualisation in the epidemiology of bacterial infections in animals

Streptococcal infections in animals are mainly caused by the species *S. agalactiae*, *S. canis*, *S. dysgalactiae*, *S. equi*, *S. suis* and *S. uberis* (Vos et al., 2011). Although they can

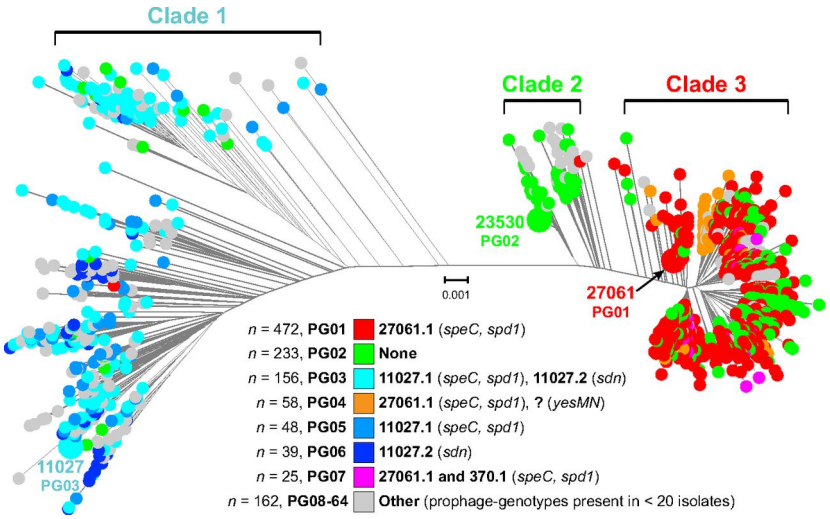


Figure 1.20: Example of an unrooted tree used to display the genomic relatedness of Group A *Streptococcus* isolates. From Beres et al. (2016).

infect and cause disease in humans, these species are generally considered a threat to animal health, particularly for livestock, with the exception of *S. agalactiae*, which is an important threat also for human health (Vos et al., 2011). Graphical visualisations are less commonly used to investigate and communicate aspects of the epidemiology of veterinary streptococcal species compared to human infections. An explanation for this might be the lack of systematic surveillance on these animal infections, which are frequently perceived as a minor zoonotic threat (Fulde and Valentin-Weigand, 2012). Despite the under-representation of data visualisation in this field, some studies using graphical displays of data to present the epidemiology of streptococcal species in animals can be found in the literature. These studies tend to focus on livestock and racehorses, suggesting a financial imperative linked to interest in these pathogens.

Line graphs have been used to show the changing proportion of animals infected by *S. uberis* and *S. suis* over time (Jayarao et al., 1999; Cloutier et al., 2003). Bar charts can also be found, for example when the proportion of isolates belonging to a specific strain needs to be shown (Mahmmod et al., 2015; Parkinson et al., 2011). The use of maps to indicate the spatial distribution of streptococcal disease cases is also occasionally encountered (Ivens et al., 2011; Parkinson et al., 2011). Phylogenetic trees appear in studies that investigate the molecular epidemiology of pathogens such as *S. agalactiae*, *S. dysgalactiae* and *S. equi* (Ivens et al., 2011; Parkinson et al., 2011; Lefébure et al., 2012; Rato et al., 2013; Mahmmod

et al., 2015; Carvalho-Castro et al., 2017).

Unlike streptococcal infections, which are considered a minor human threat and occasionally an economic concern for the food and race-horse industries, zoonotic pathogens and AMR are high in the public consciousness and aspects of their epidemiology are often translated into figures. For zoonotic bacteria, the most common visualisations encountered are infographics displaying the routes of transmission of a pathogen from animals to people (García et al., 2010; Lim et al., 2019; Shin and Park, 2018). These infographics, which are generally easy to interpret, are included both in scientific publications and on mainstream platforms (`https://www.ecdc.europa.eu/en/zoonoses`) for the need to educate the public about potential health threats.

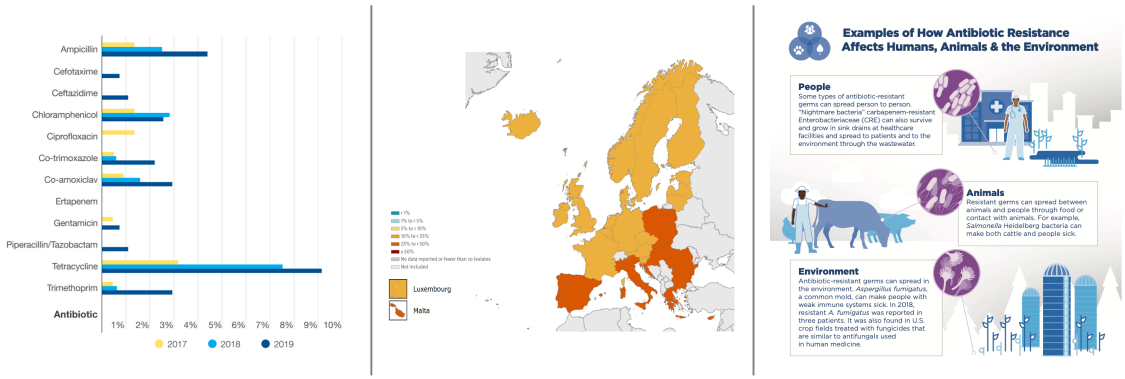


Figure 1.21: Examples of the three most common visualisations in, respectively, the Scottish One Health Antimicrobial Use and Antimicrobial Resistance (SONAAR) report (left), the European Antimicrobial Resistance Surveillance Network (EARS-Net) report (centre) the Antibiotic Resistance (AR) Threats report (right). Left panel, bar chart showing the proportion of *Escherichia coli* isolates non-susceptible to selected antimicrobials in healthy cattle in Scotland, from 2017 to 2019. Central panel, percentage of *E. coli* invasive isolates resistant to fluoroquinolones by country in Europe in 2019. Right panel, infographic about the role that people, animals and the environment play in antimicrobial resistance occurrence and spread.

The spread of AMR, which is influenced by human-animal interactions and by the use of antimicrobials in pets and farm animals, is communicated through research papers, official surveillance reports and awareness campaigns. Surveillance reports are widely accessible to the general public, although they are generally targeted to spe-

cialised audiences. The type of visualisations found in these reports is varied (Figure 1.21). In the Scottish One Health Antimicrobial Use and Antimicrobial Resistance (SONAAR) report, the use of antimicrobials and the spread of AMR among pathogenic bacteria is displayed through graphs, in particular line graphs, bar charts and pie charts (https://www.hps.scot.nhs.uk). In the ECDC European Antimicrobial Resistance Surveillance Network (EARS-Net) report, the prevalence of bacterial isolates resistant to specific antibiotics throughout Europe is displayed using colour-coded maps (https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019). The CDC Antibiotic Resistance Threats (AR Threats) report makes use of infographics to communicate data regarding AMR spread in the US (https://www.cdc.gov/drugresistance/biggest-threats.html). These three different approaches each have their advantages and disadvantages. The visualisations found in the SONAAR report are intuitive and easy to interpret but they are not visually attractive and memorable, and one would predict they be of very limited interest to a non-specialist audience. Health Protection Scotland provides a summary of the SONAAR report findings in the form of infographics to inform the general public, but these visualisations are not included in the main report and could be easily overlooked (https://www.hps.scot.nhs.uk). The maps in the EARS-Net report are effective in summarising the prevalence of AMR in specific bacterial species across Europe, but they fail to communicate other relevant data, such as changing trends over time. The infographics that are part of the AR Threats report are a good choice to inform the general public about the AMR spread and threats, but they require time to be interpreted and are less intuitive than other visualisations. A good balance of graphs, maps and infographics could make an AMR report an effective means to communicate data to both a specialist audience and the general public.

1.5 Aim and objectives

The overall aim of this PhD project was to gain insights into the epidemiology of *Streptococcus pyogenes* and *Streptococcus canis* infections in Scotland using a multi-disciplinary approach. Consequently, the following objectives were formulated:

1. To characterise the main epidemiological features of iGAS infections in Scotland from 2014 to 2021 using data collected by the Scottish Microbiology Reference Laboratory (SMiRL).

2. To investigate the upsurge of cases of invasive disease associated with the uncommon and recently emerged genotype *emm*5.23 of *S. pyogenes* circulating in Scotland using whole genome sequencing (WGS) and transcriptomic analyses.

3. To characterise AMR prevalence, virulence gene carriage and population structure of *S. canis* strains isolated from different host species in multiple geographic locations, including Scotland, using WGS analyses and antimicrobial susceptibility testing (AST).

4. To explore and optimise the use of data visualisation to communicate the epidemiology of iGAS disease to a small cohort of public health and laboratory workers in Scotland.

Chapter 2

Descriptive epidemiology of invasive Group A *Streptococcus* infections in Scotland from 2014 to 2021

2.1 Introduction

Invasive GAS disease can manifest in different ways such as cellulitis, necrotising fasciitis, bacteremia, puerperal sepsis, meningitis, septic arthritis, peritonitis, endocarditis and osteomyelitis (Walker et al., 2014; Sanyahumbi et al., 2016). In Scotland, necrotising fasciitis is a notifiable disease and *S. pyogenes* isolated from blood, cerebrospinal fluid or other normally sterile site is a notifiable pathogen (`https://www.legislation.gov.uk/asp/2008/5/schedule/1/2010-01-01`). Data on the global epidemiology of iGAS infections may suffer from bias for at least two reasons. Firstly, most of the available data are collected in high-income countries (Carapetis et al., 2005) and secondly, the lack of an international surveillance network to monitor iGAS infections prevents the establishment of standardised surveillance methods and a universally accepted case definition (Efstratiou and Lamagni, 2016). This chapter presents and describes epidemiological data on iGAS infections in Scotland from 2014 to 2021 and data on GAS-positive specimens submitted to diagnostic laboratories of the Greater Glasgow & Clyde (GG&C) Health Board from 2018 to early 2022.

Incidence of iGAS disease The global incidence of iGAS disease appears to have increased since the 1980s (Henningham et al., 2012), with periodic upsurges due to the emergence of highly virulent strains, such as *emm*1, 3, 12, 28 and 89 (Beres et al., 2002; Nasser et al., 2014; Turner et al., 2015; Beres et al., 2016; Feng et al., 2016; Kachroo et al., 2019), together with occasional outbreaks associated with uncommon *emm* types in at-risk communities, such as the upsurge of invasive *emm*59 cases in Western Canada from 2006 to 2009 (Tyrrell et al., 2010). In the last 20 years, the iGAS infection rate in high income countries has ranged from 2.45 to 9.8 cases per 100,000 people (Carapetis et al., 2005; Lamagni et al., 2008a; Lepoutre et al., 2011; Stockmann et al., 2012), depending on the geographical area, the surveillance system in use and the occurrence of large scale outbreaks (Efstratiou and Lamagni, 2016). In low income countries and among indigenous populations of Australia and the USA, the incidence is estimated to be relatively high, with recorded rates of 12, 46 and even 83 cases per 100,000 people (Carapetis et al., 2005; Efstratiou and Lamagni, 2016). In 2019, the rate of GAS bacteraemia in people older than 75 years was 15 per 100,000 people in England and 7 per 100,000 people in Scotland. In the same year, the rate of GAS bacteraemia in neonates (< 1 year) was 6.3 and 7.8 per 100,000 people in England and Scotland, respectively (`https://webarchive.nationalarchives.gov.uk/ukgwa/20210216063045/https://www.gov.uk/government/publications`; `https://www.hps.scot.nhs.uk/data/`).

Risk factors for iGAS disease Recognised risk factors for iGAS disease include co-morbidities, seasonality, age and socio-economic factors. A variety of cutaneous lesions, including blunt force trauma, have been associated with an increased risk of developing iGAS disease (Lamagni et al., 2008a; Lamb et al., 2015). Other clinical conditions regarded as predisposing factors include diabetes mellitus, which is associated with an increased risk of developing pneumonia, urinary tract infections and skin infections (Benfield et al., 2007), and acute viral respiratory infections (Efstratiou and Lamagni, 2016). Among the latter, influenza in adults and varicella in children are the most frequently associated with iGAS infections (Lamagni et al., 2008a; Morens et al., 2008). Pregnancy is another suggested risk factor for iGAS disease (Leonard et al., 2019), which has been associated with an incidence of 12 cases per 100,000 livebirths in pregnant women in high-income countries (Sherwood

et al., 2022).

Epidemiological data collected in North America and Europe suggest a seasonal pattern to iGAS infection rates. Although cases are diagnosed throughout the year, the majority occur in the winter and spring, with summer and autumn being characterised by a lower incidence of disease (Lamagni et al., 2008a, 2009). It has been hypothesised that this pattern may be explained, in part, by lower environmental humidity and solar exposure in the winter and spring months, which would influence the local and systemic immune response (Dowell et al., 2003; Molesworth et al., 2003). The same seasonal pattern is also observed for influenza and other acute respiratory viral infections (Monto, 2002; Lagace-Wiens et al., 2010), which are recognised risk factors for iGAS disease and which may in turn have an additive effect.

Age represents another recognised risk factor for iGAS disease. Published data report the highest incidence among the elderly and infants, and the lowest rates in teenagers and young adults (Lepoutre et al., 2011; Efstratiou and Lamagni, 2016). In France in 2007, more than 30% of all iGAS cases were diagnosed in people above 80 years of age, 6% in infants (0-4 years old) and less than 3% in individuals between 10 and 29 (Lepoutre et al., 2011). In a recently published systematic review and meta-analysis, the estimated incidence of iGAS disease in neonates was 4 cases per 100,000 livebirths worldwide, 12 cases per 100,000 livebirths in low and middle-income countries and 2 cases per 100,000 livebirths in high-income countries (Sherwood et al., 2022).

Poverty and poor hygiene conditions not only facilitate occurrence and transmission of iGAS disease, but also limit the efficacy of prevention and treatment measures (Carapetis et al., 2005; WHO, 2005). Intravenous drug use and, in particular, sharing of injecting equipment, are other recognised socio-economic risk factors for iGAS infection due to the use of contaminated injection paraphernalia (Navarro et al., 1993; Lamagni et al., 2008a,c).

A substantial proportion of all iGAS cases, around 20-30%, occur in individuals with no identifiable risk factors, suggesting there may be additional predisposing conditions or

routes of transmission yet to be described (Lamagni et al., 2008a; O'Loughlin et al., 2007). Around 20% of all the people who develop iGAS disease die within the first seven days of infection (Sanyahumbi et al., 2016).

Antimicrobial treatment Penicillin is the treatment of choice for GAS pharyngitis due to the universal susceptibility of *S. pyogenes* strains to β-lactams (Bisno et al., 2002). The National Institute for Health and Care Excellence (NICE) recommends the following penicillin treatment regimens for GAS pharyngitis: 500 mg four times a day or 1,000 mg twice a day for 5 to 10 days to treat patients aged 18 years and over; 125 mg twice a day for 5 to 10 days for children and young people under 18 years (`https://www.nice.org.uk/guidance/ng84/chapter/Recommendations#choice-of-antibiotic`). When traditional treatment cannot be undertaken due to penicillin allergy, patients can be successfully treated with macrolides, such as erythromycin and azithromycin, or lincosamides, such as clindamycin (Bisno et al., 2002). Dose and duration of macrolide and lincosamide treatment for GAS pharyngitis are reported on the NICE guidance (`https://www.nice.org.uk/guidance/ng84/chapter/Recommendations#choice-of-antibiotic`). Similarly, the Infectious Diseases Society of America recommends the use of erythromycin, clindamycin or amoxicyllin-clavulanate to treat GAS impetigo (Stevens et al., 2014). Invasive infections, conversely, should be managed with the administration of both penicillin and clindamycin or with vancomycin, linezolid, quinupristin/dalfopristin or daptomycin in case of penicillin allergy (Stevens et al., 2014). As previously mentioned, resistance to penicillin has never been observed in GAS strains, although recent evidence suggests that mutation and recombination events responsible for reduced susceptibility to β-lactams can occur in *S. pyogenes* isolates (Beres et al., 2022). Resistance to other antibiotic classes used to treat GAS infections, however, is well documented. In a study on AMR among iGAS strains isolated in the US, the prevalence of erythromycin and clyndamycin resistance in 2017 was 22.8% and 22.0%, respectively (Fay et al., 2021). The results presented in the same study revealed that the prevalence of iGAS resistance to both antibiotics had been increasing from 2006 to 2017 (Fay et al., 2021). Other reported rates of GAS resistance to macrolides ranged from 27.5% to 61.0% (Ubukata et al., 2020; Tsai et al., 2021), while clyndamycin resistance appears to

be slightly lower, ranging from 3% to 23% of all isolates (Pesola et al., 2015). Given the reported increase in macrolide resistance rates in recent years, the combined use of macrolides and lincosamides is recommended in patients allergic to penicillin (Fay et al., 2021). No vancomycin resistance has been documented to date (Johnson and LaRock, 2021).

Epidemiology of iGAS *emm* types The epidemiology of iGAS infections is influenced by the virulence characteristics of the bacteria involved, which can be associated with specific *emm* types. Certain *emm* types cause invasive disease more frequently than others, mostly due to the acquisition of virulence traits such as increased production of toxins or loss of the hyaluronic capsule (Beres et al., 2002, 2016; Feng et al., 2016; Kachroo et al., 2019; Nasser et al., 2014). Sporadic iGAS outbreaks can also be associated with the introduction of novel *emm* types in a "virgin" population for which herd immunity has not yet developed (Southon et al., 2020). In low-income countries, iGAS *emm* type diversity appears to be broader than that of high-income countries, with no significant predominance of specific *emm* types in the overall disease burden (Steer et al., 2009). An association between *emm* types and certain clinical manifestations has been described. STSS, for example, was significantly associated with *emm* 1 and 3, puerperal sepsis with *emm*28 and cellulitis with *emm*87 or *emm*83 in a study conducted in Europe (Luca-Harari et al., 2009). Some *emm* types have been also associated with specific age groups. *Emm*1 and 4, for example, appear to be significantly more common in children (Lepoutre et al., 2011; Olafsdottir et al., 2014; Tamayo et al., 2014), while *emm*89, *emm*77 and *emm*28 have been associated with disease in the elderly (Lepoutre et al., 2011; Tamayo et al., 2014).

Surveillance systems for iGAS disease The epidemiology of GAS infections and impact on patients is not systematically monitored at an international level. Surveillance, when practiced, is implemented only at a national level and it is frequently limited to monitoring only the invasive cases of disease. The most recent WHO report on the worldwide epidemiology of GAS infections was published in 2005, summarising data from all published studies on GAS disease burden until that year (WHO, 2005). The median incidence rate of iGAS disease was estimated to be 2.45 per 100,000 people in high-income countries, while in indigenous populations in the USA and Australia the estimated incidence

was as high as 46 per 100,000 people (WHO, 2005). However, iGAS epidemiological data regarding low-income countries was found to be limited due to the paucity of good quality population-based studies conducted in these areas (WHO, 2005). Nevertheless, one such study revealed that GAS was the third most common cause of neonatal bacteraemia in Kenya (Berkley et al., 2005) while another suggested that the overall incidence of GAS bacteraemia in Fiji was 8.7 per 100.000 people (WHO, 2005). The WHO report goes on to estimate the mortality rate for iGAS disease expressed as cases per 100,000 people per year to be 0.15 in high-income countries and 0.25 in low-income countries and indigenous populations of the USA and Australia (WHO, 2005).

In the US, the CDC implements surveillance of iGAS infections through an active laboratory and population-based system known as Active Bacterial Core surveillance (ABCs) (CDC, 2018). ABCs covers ten areas within the US and represents more than 34 million people (CDC, 2020). ABCs defines a case of iGAS as the "isolation of Group A *Streptococcus* from a normally sterile site or from a wound culture accompanied by necrotising fasciitis or streptococcal toxic shock syndrome in a resident of a surveillance area" (CDC, 2020). Based on a recently published ABCs report on iGAS, the incidence rate in the US was estimated to be as high as 7.69 per 100,000 people with a fatality rate of 0.71 per 100,000 people (CDC, 2020).

In Europe, prospective population-based surveillance on iGAS infections was implemented across eleven countries between 2003 and 2004 through the surveillance programme "Strep-EURO" (Lamagni et al., 2008a). Strep-EURO surveillance revealed an overall crude infection rate of 2.79 per 100,000 people across the countries involved (Lamagni et al., 2008a). Based on Strep-EURO data, the risk of infection was highest among the elderly and the overall seven-day case fatality rate was calculated to be 19% (Lamagni et al., 2008a). Except for the Strep-EURO project, there have been no international surveillance efforts for GAS-associated disease in Europe. The European Centre for Disease Prevention and Control (ECDC) does not currently consider GAS associated disease as a priority for inclusion in any surveillance network (`https://www.ecdc.europa.eu/en/publications-data?f%5B0%5D=output_types%3A1244`).

In England, where cases of scarlet fever and iGAS infection are notifiable, surveillance data on GAS epidemiology are published regularly. The UK Health Security Agency (UKHSA) reports the seasonal activity of scarlet fever and iGAS infections in England three times per year (GOV.UK, 2020). In their report regarding the disease season immediately prior to the pandemic, from week 37 of 2019 to week 11 of 2020, the highest incidence of iGAS disease recorded in England was 3.7 per 100,000 people, in the Yorkshire and Humber region (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/875173/hpr0620_GAS-SF_fin5.pdf). In the same report, the highest rate of infection per age category was 9 per 100,000 population in people above 75 years of age. In a more recent report, the highest rates from week 37 of 2021 to week 10 of 2022 was 1.7 per 100,000 people in the North East of England (https://www.gov.uk/government/publications). In Scotland, where only invasive cases of GAS infection are notifiable, epidemiological data on iGAS disease used to be published once a year by Health Protection Scotland (HPS) until 2019. Diagnostic laboratories in Scotland are responsible for the submission of iGAS isolates to SMiRL and, additionally, laboratory positive reports to Public Health Scotland (PHS), previously known as HPS, via the Electronic Communication of Surveillance in Scotland (ECOSS) system (https://www.hps.scot.nhs.uk/a-to-z-of-topics/streptococcal-infections/). For each confirmed case of iGAS disease, HPS collects enhanced surveillance data. Although scarlet fever is no longer a notifiable disease in Scotland, laboratory confirmed cases of GAS from the upper respiratory tract are used as a proxy for this condition (https://www.hps.scot.nhs.uk/a-to-z-of-topics/streptococcal-infections/).

Aims and objectives The primary aim of this chapter was to describe the distribution and characteristics of iGAS infections in the Scottish population. The secondary aim of the chapter was to assess the impact of the COVID-19 pandemic on epidemiological patterns of iGAS disease in Scotland and on all GAS isolations in the GG&C Health Board. The following objectives were defined to address these aims:

- To present the incidence of iGAS disease in the Scottish population before and during the COVID-19 pandemic.
- To consider patterns of seasonal and age-specific incidence of iGAS disease before and during the COVID-19 pandemic.
- To characterise the *emm*-specific iGAS disease burden before and during the COVID-19 pandemic.
- To present the anatomical isolation sites of iGAS cases before and during the COVID-19 pandemic.
- To describe differences in number and characteristics of all GAS isolations in the GG&C Health Board before and during the COVID-19 pandemic.

2.2 Methods

2.2.1 Case definition and data collection

In Scotland, a case of iGAS disease is confirmed when GAS is isolated from a normally sterile body site or from non-sterile body sites of patients with severe clinical manifestations, such as Streptococcal Toxic Shock Syndrome or necrotising Fasciitis (`https://www.nhsggc.org.uk/media/239650/shlmprl-user-manual.pdf`). Invasive isolates are submitted to the Bacterial Respiratory Infection Service, SMiRL, at the Glasgow Royal Infirmary, for MLST and *emm* typing (`https://www.nhsggc.org.uk/media/239650/shlmprl-user-manual.pdf`). Each GAS isolate sent to SMiRL is accompanied by a submission form in which information regarding date and Health Board of isolation, as well as patient's age and specimen type, are collected (`https://www.nhsggc.org.uk/media/262227/smirl-shlmprl-request-form-rf-4.pdf`). For the purposes of this study, data on iGAS epidemiology were accessed and collected from the SMiRL database. Approval for data access and analysis was sought and received through a Public Benefit and Privacy Panel (PBPP) for Health and Social Care application (Appendix A.2). All iGAS cases and associated metadata collected by SMiRL from 2014 to 2021 were considered. The common

pattern for referral of iGAS isolates to SMiRL is to send one invasive isolate per patient. In cases of multiple specimens collected from the same patient, only isolates of different *emm* types were classified as multiple cases of infection. Data de-duplication was carried out by authorised staff at SMiRL before data was transferred to the University of Glasgow. Limited data was extracted from the SMiRL database and used for analysis as per the PBPP process. Briefly, data fields analysed in this thesis were patient age, gender, week, month and year isolate collected, site of isolate, *emm* type and whole genome sequence (where available).

Data regarding all GAS-positive specimens submitted to diagnostic laboratories of the GG&C Health Board from 2018 to April 2022 were also considered in this work. Although unlikely to be representative of the real disease burden of GAS infections, these data provide some insight into the frequency and characteristics of GAS isolations in this Health Board. Data on GAS-positive throat swabs were accompanied by antimicrobial susceptibility testing (AST) data. AST was performed by diagnostic laboratories on nearly all GAS isolates derived from throat swabs using Vitek 2 technology (Ligozzi et al., 2002). GG&C data were collected and provided by SMiRL.

2.2.2 Data analysis and visualisation

Data were cleaned and analysed using both Microsoft Excel v16.46 (Microsoft Corporation, 2021) and RStudio v1.4.1103 (RStudio Team, 2020). Data visualisations were generated with RStudio v1.4.1103 (RStudio Team, 2020).

2.2.3 Incidence of iGAS in Scotland

The incidence of invasive infections in the Scottish population was expressed as number of cases per 100,000 people per year. Population estimates for each of the years considered in this study were collected from the National Records of Scotland website (`https://www.nrscotland.gov.uk`). Given the association between GAS disease and influenza (Lamagni et al., 2008a), annual iGAS incidence rates were compared to those of severe influenza, expressed as cases that required hospitalisation per 10,000 population. Influenza data for the Scottish population was provided by PHS. The monthly incidence of iGAS

infections was expressed as number of confirmed cases per calendar month. The incidence of iGAS disease in different age groups was presented as the number of cases per 100,000 population and population estimates for each age group considered were collected from from the National Records of Scotland website (`https://www.nrscotland.gov.uk`).

2.2.4 Scottish iGAS *emm* types

The overall diversity of invasive *emm* types isolated each year in Scotland was measured using the Simpson's Index of Diversity (SID) with 95% confidence intervals (CI) (Simpson, 1949). The SID represents the probability that two randomly chosen strains isolated in the same year belonged to separate *emm* types. The SID takes into account both the number of different *emm* types recorded each year and the relative frequency of isolation of each *emm* type. The Bray Curtis (BC) dissimilarity index was then used to estimate the *emm* type diversity from year to year. The BC dissimilarity index ranges from 0 for identical samples to 1 for completely distinct samples. The BC dissimilarity index is influenced not only by the sample composition (in this case the *emm* types recorded each year) but also by the total sample size. Consequently, since the number of iGAS cases per year from 2014 to 2021 was highly variable, and since the main focus of this analysis was to quantify *emm* type turnover from year to year, the absolute count of each *emm* type per year was converted to its relative frequency. Practically, the number of cases of each *emm* type in a specific year was divided by the total number of iGAS cases for that year and multiplied by 100. The BC dissimilarity index was then calculated for the transformed values, giving a measure of the different *emm* type composition from year to year. Finally, the relative frequency of isolation of each *emm* type involved in invasive disease in Scotland for each year from 2014 to 2021 was plotted and visually explored.

2.2.5 iGAS isolation site

The frequency of isolation of iGAS strains from different anatomical sites was calculated for the years immediately preceding the pandemic and for the pandemic years. Data on iGAS isolation site were reported following the recommendations provided by the Statistical Disclosure Control guidelines of the National Services of

Scotland (https://www.isdscotland.org/About-ISD/confidentiality/disclosure_protocol_v3.pdf). Specifically, as outlined in Subsection 4.3 of the disclosure protocol, specimen sources associated with fewer than 5 isolations in the whole Scotland were grouped under one single category "other".

2.2.6 GAS isolates from the Greater Glasgow & Clyde Health Board

The monthly count of all GAS-positive specimens (both invasive and non-invasive) submitted to diagnostic laboratories of the GG&C Health Board from 2018 to April 2022 was initially assessed. Since one of the main carriage and colonisation sites of GAS is the upper respiratory tract, particular attention was given to throat swabs data and the relative frequency of positive throat swabs from specific age groups from 2018 to 2021 calculated. The majority of isolates from throat swabs had been tested for sensitivity towards Clarithromycin, Clindamycin, Doxycycline and Penicillin. Resistance rates for the isolates tested were available for each year from 2018 to 2021. Finally, the gender-specific proportion of GAS-positive throat swabs was calculated for the total number of specimens and for the underage ($<$ 18 years) vs adult population ($\geq$ 18 years) for each year. Previous studies hypothesised that men tend to be less likely than women to promptly seek help in case of disease (Galdas et al., 2005). In order to investigate whether this behavioural factor could have been responsible for the higher number of female derived specimens in the available dataset, the gender-specific proportion of throat swabs in the underage population (less likely to make independent decisions regarding their health status) was compared to that of the adult population using a z-test for equality of proportions.

2.3 Results

All supplementary material for this chapter can be found in Appendix A.

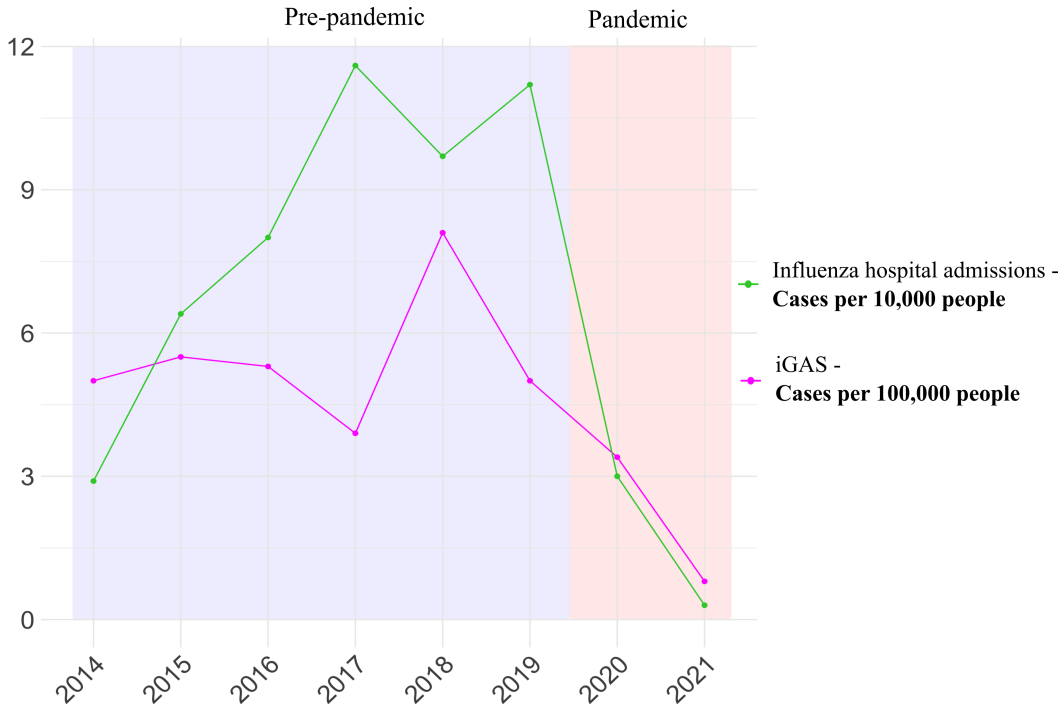


Figure 2.1: Annual incidence of invasive Group A *Streptococcus* disease in Scotland from 2014 to 2021 expressed as number of cases per 100,000 people. An estimation of the Scottish annual incidence of severe influenza cases, represented by the number of influenza hospital admissions per 10,000 people, is also provided for comparison purposes. Influenza data from 2015 to 2021 were collected and provided by Public Health Scotland.

2.3.1 Incidence of iGAS in Scotland

The overall incidence of iGAS disease in Scotland before the COVID-19 pandemic (2014 - 2019) fluctuated between 3.9 and 8.1 cases per 100,000 people per year (Figure 2.1). In 2020 and 2021, the annual iGAS incidence rate dropped to 3.4 and 0.8 per 100,000 people, respectively. The incidence of severe influenza, expressed as hospital admissions per 10,000 people, followed a trend that did not resemble that of iGAS across the years preceding the pandemic.

The monthly incidence of iGAS infection was variable throughout the year, but with a higher incidence in winter and spring compared to summer and autumn (Figure 2.2). Before the COVID-19 pandemic, iGAS incidence in Scotland ranged from less than 10 to more than 70 cases per month. From the beginning of the pandemic (i.e. the second half of March 2020) to the end of December 2021, iGAS monthly cases dropped to between 1 (November

2021) and 18 (April 2020) (Figure 2.2).

The incidence of iGAS disease in the Scottish population showed substantial differences among different age groups, as shown in Figure 2.3. Before the COVID-19 pandemic (2014-2019), iGAS incidence across age groups resembled a W shaped curve, with the highest rates for very young (0-4 years) and very old (≥80) individuals, together with a smaller peak coinciding with the 30-39 age group (Figure 2.3). A very similar pattern was evident throughout 2020, although at a marginally lower level than that during the pre-pandemic period. During 2021, however, the difference in iGAS incidence between age groups was less pronounced and four minor peaks, corresponding to age groups 0-4, 20-29, 40-49 and ≥80, were noted (Figure 2.3).

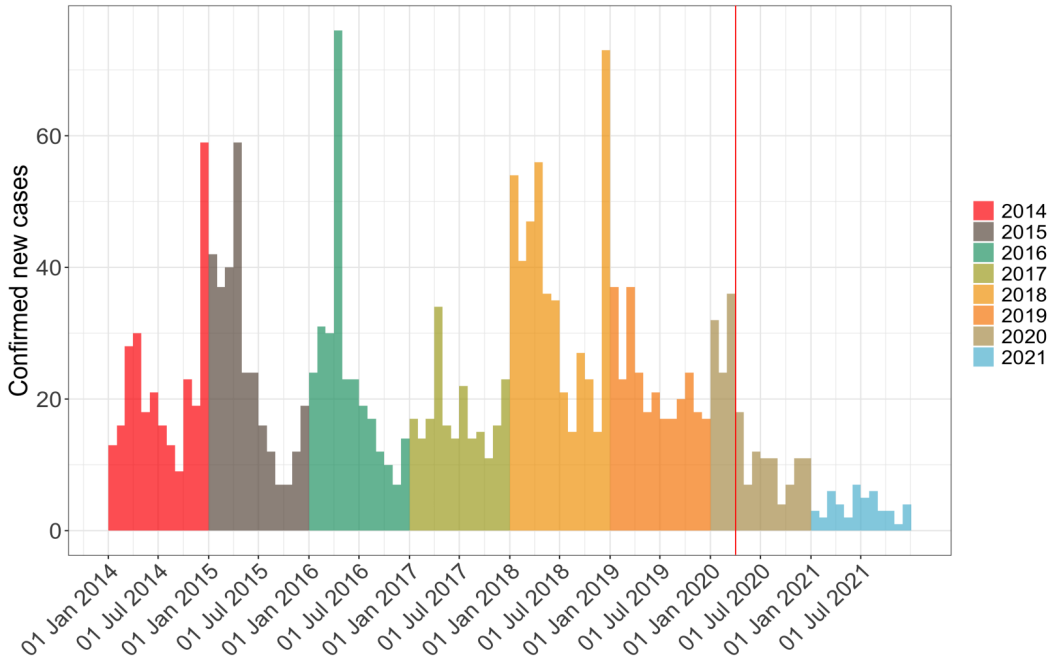


Figure 2.2: Monthly incidence, expressed as number of confirmed cases per month, of invasive Group A *Streptococcus* disease in Scotland from 2014 to 2021. A red line between March and April 2020 highlights the beginning of the first Scottish national lockdown in response to the COVID-19 pandemic.

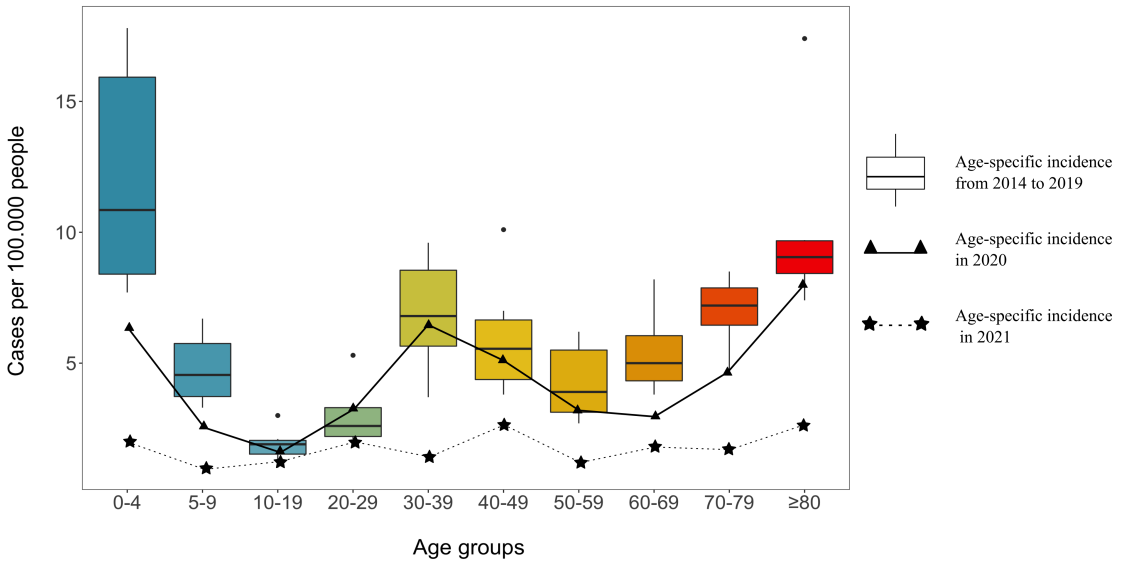


Figure 2.3: Age-specific incidence of invasive Group A *Streptococcus* disease in the Scottish population recorded each year from 2014 to 2021. Data points referring to the pre-pandemic period (2014-2019) are combined in a box plot, while those for 2020 and 2021 are depicted as line graphs.

2.3.2 Scottish iGAS *emm* types

A total of 2,020 iGAS isolations of 125 different *emm* types and sub-types were reported in Scotland from 2014 to 2021. A breakdown of all invasive strains recorded each year is provided in Table A.1.

In order to quantify the overall *emm* type diversity of invasive isolates, the SID with 95% CI was calculated for each of the years considered in this study (Figure 2.4). The SID in this case represents the probability that two randomly selected invasive strains from the same year belong to different *emm* types. As shown in Figure 2.4, the SIDs for 2014 to 2017 (inclusive) are similar, whereas increased diversity was measured from 2018 to 2021. It can be observed that the SID for the years of the COVID-19 pandemic are not dissimilar from those of 2018 and 2019.

The BC dissimilarity index was used to assess the year-to-year diversity in invasive *emm* type composition in Scotland (Figure 2.5). The higher the BC dissimilarity value for two consecutive years, the higher the *emm* type turnover from one year to another. From 2014 to 2017 the year-on-year dissimilarity index ranged from 0.20 to 0.27. An increase in the

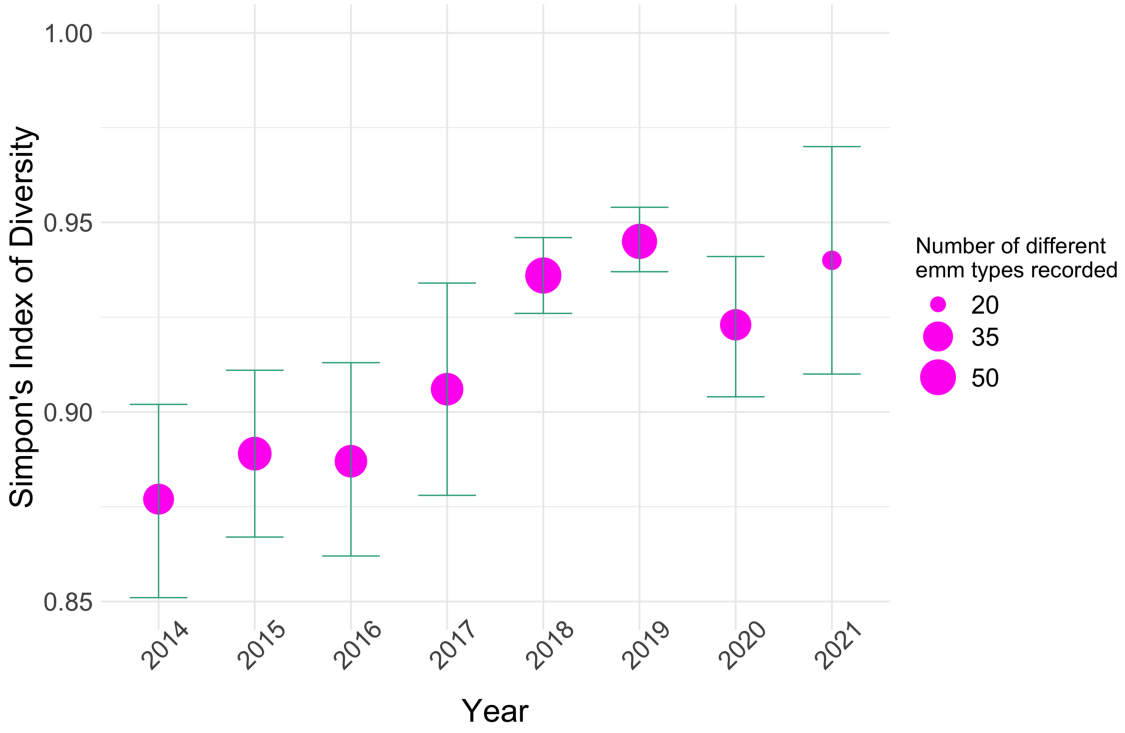


Figure 2.4: Simpson's Index of Diversity with 95% CI for invasive Group A *Streptococcus* disease in Scotland for each year considered in this study.

BC dissimilarity index was noticed after 2017, with values of 0.36 and 0.37 for 2017-2018 and 2018-2019, respectively. The BC index for 2019-2020, which indicates the transition from the pre-pandemic era to the first year of the COVID-19 pandemic, is 0.33, suggesting an *emm* type turnover similar to that observed in previous years. The BC dissimilarity index for 2020-2021 (0.57), however, was notably higher than every previous year studied. This implies that the *emm* type composition of 2021 was substantially different to that of the previous year (i.e. the first year of the pandemic). When compared to previous year-on-year differences, a strikingly high level of *emm* type turnover can be observed between 2020 and 2021.

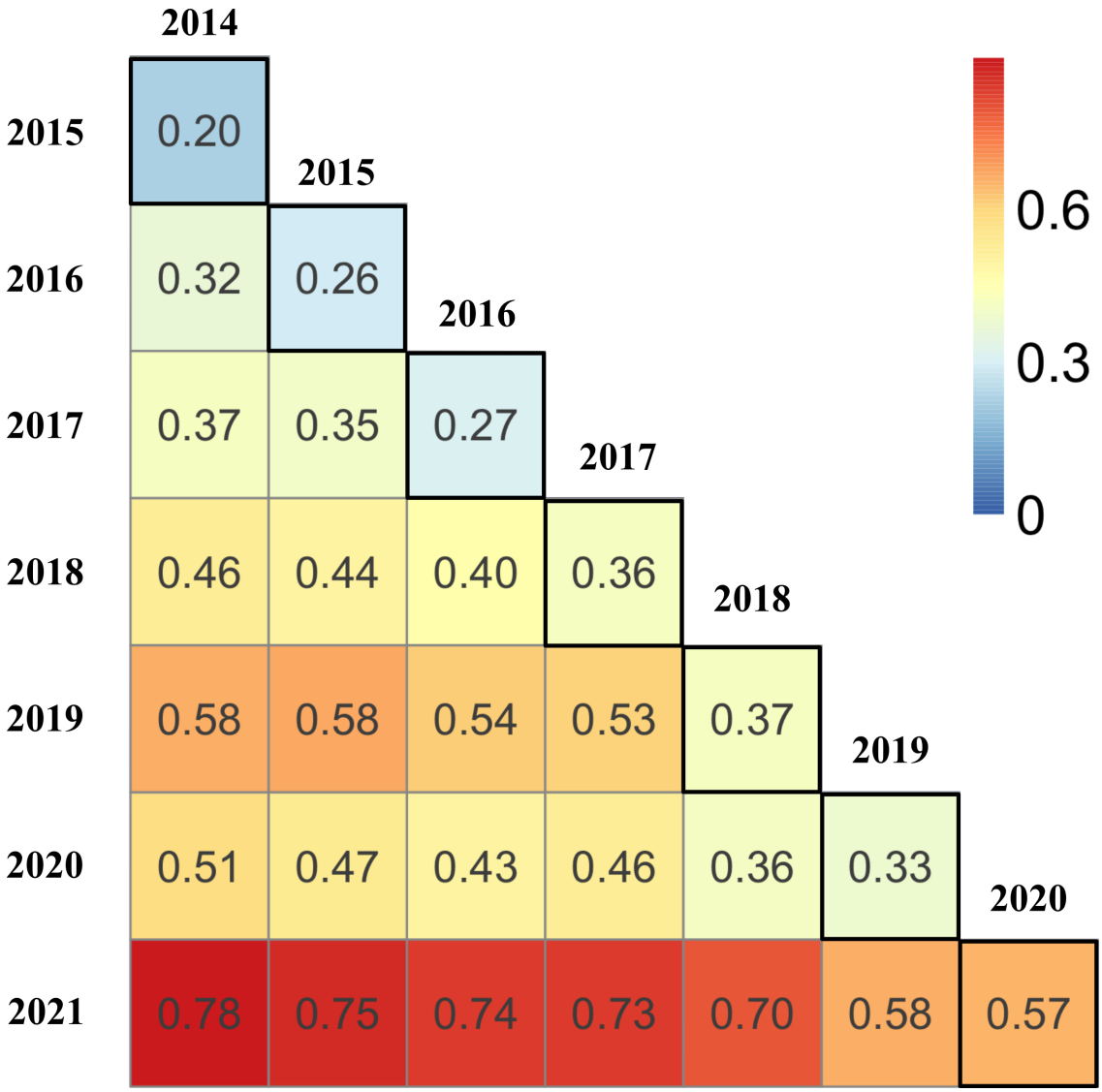


Figure 2.5: BC dissimilarity index for each pair of years from 2014 to 2021. BC index was calculated based on the *emm* types causing invasive disease in Scotland each year. Indices referring to consecutive years are highlighted with a thicker border and can be considered a reflection of the year-on-year turnover of invasive *emm* types.

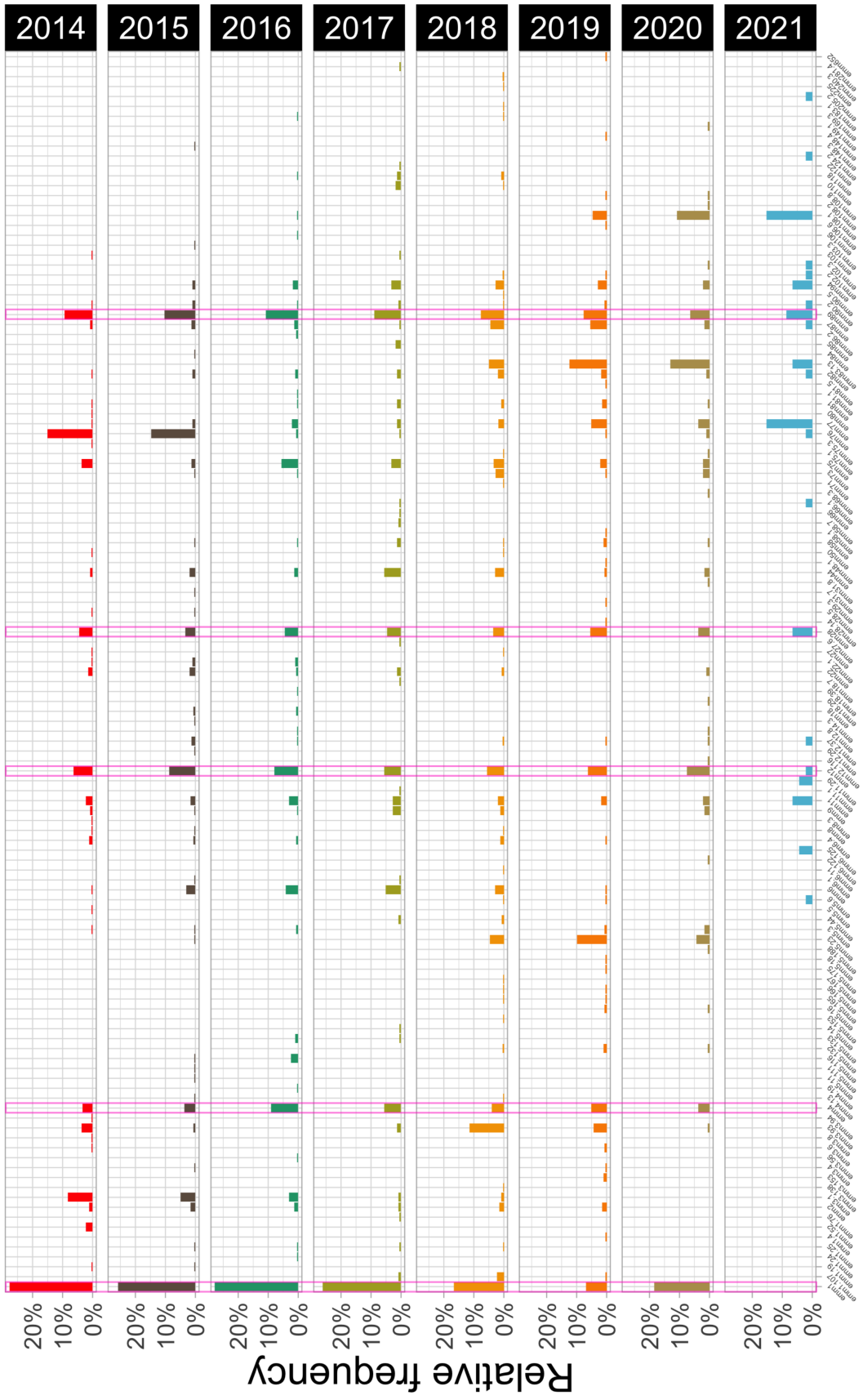


Figure 2.6: Disease burden of all *emm* types implicated in invasive infections in Scotland from 2014 to 2021. Burden of disease of the Scottish predominant *emm* types (*emm*1, *emm*4, *emm*12, *emm*28 and *emm*89) is highlighted.

The *emm* type-specific invasive disease burden, expressed as proportion of isolates of a specific *emm* type over the total number of isolates per year, was calculated and is shown in Figure 2.6. In agreement with the BC indices previously presented, a variation in the proportion of isolates of different *emm* types was noted from year to year. Only five *emm* types, namely *emm*1, 4, 12, 28 and 89, were consistently among the ten most frequently isolated strains across each of the pre-pandemic years. These five strains will henceforth be referred to as predominant *emm* types. Certain other *emm* types, such as *emm*76, *emm*83.13 and *emm*5.23, were responsible for a high proportion of invasive disease cases only for a few years. The majority of the *emm* types recorded were associated with a very low infection burden, suggesting a sporadic involvement in invasive disease. The first year of the pandemic was characterised by a reduction in absolute number of iGAS cases, but the relative frequency of *emm* types was not strikingly different from those of the pre-pandemic years. In 2021, however, there was a reduction not only in the total number of cases but also in the number and frequency of individual *emm* types. In particular, *emm*1 and *emm*4, which were two of the predominant types until 2020, were not responsible for any case of invasive disease in 2021. In contrast, the disease burden of *emm*28 and *emm*89 in 2021 was very similar to that recorded in pre-pandemic years. While invasive cases of *emm*12 still took place in 2021, the frequency of this strain was noticeably reduced compared to previous years. The change of invasive disease burden of the five predominant *emm* types across the years is shown in Appendix A.1.

2.3.3 iGAS isolation site

Invasive strains of *S. pyogenes* were isolated from at least 14 different body sites. As shown in Table 2.1, however, the majority of these isolations (92.8%) were from blood, cutaneous tissue or the respiratory tract. There were no striking differences in the frequency of iGAS isolation sites recorded in pre-pandemic years and during the pandemic.

Table 2.1: Count and proportion of invasive Group A *Streptococcus* strains according to specimen origin, before and during the COVID-19 pandemic.

Specimen source	Pre-pandemic (2014-2019)	Pandemic (2020-2021)	Total
Blood	955 (53.4%)	134 (58.0%)	1089 (53.9%)
Cutaneous tissue	496 (27.7%)	59 (25.5%)	555 (27.5%)
Respiratory tract	210 (11.7%)	17 (7.4%)	227 (11.2%)
Reproductive tract (Female)	66 (3.7%)	5 (2.2%)	71 (3.5%)
Joint	22 (1.2%)	9 (3.9%)	31 (1.5%)
Unspecified	16 (0.9%)	0 (0.0%)	16 (0.8%)
Abdominal cavity	6 (0.3%)	0 (0.0%)	6 (0.3%)
Urine	6 (0.3%)	0 (0.0%)	6 (0.3%)
Other	12 (0.7%)	7 (3.0%)	19 (0.9%)
Total	1789 (100%)	231 (100%)	2020 (100%)

2.3.4 GAS isolates from the GG&C Health Board

In this section, data on the epidemiology on all GAS infections in the GG&C Health Board area are presented. This data refers to all GAS-positive specimens submitted to diagnostic laboratories (including invasive ones) and is not a precise representation of the real GAS disease burden. Since non-invasive cases tend to be mild, self-limiting and/or easy to treat empirically, only a proportion of patients seek medical advice or have samples collected. The data described here is, nevertheless, useful for observing changing trends in GAS diagnostic submissions in the GG&C area over time. The total number of GAS-positive specimens confirmed each month from January 2018 to April 2022 is shown in Figure 2.7. From January 2018 to March 2020 the monthly count ranged from 138 (September 2018) to 683 (March

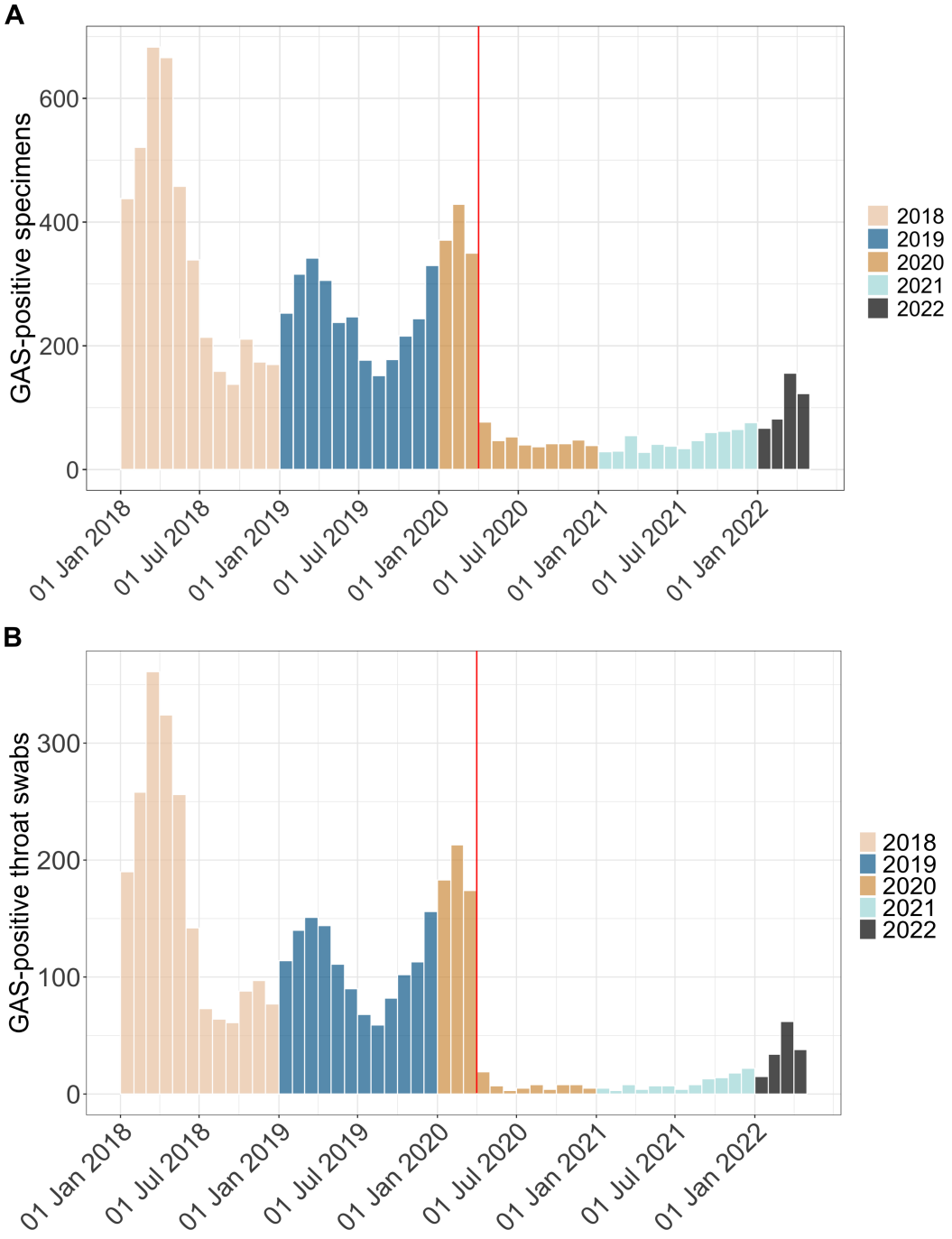


Figure 2.7: A - All Group A *Streptococcus* (GAS)-positive specimens submitted per month to diagnostic laboratories of the GG&C Health Board. B - GAS-positive throat swabs submitted per month to diagnostic laboratories of the GG&C Health Board. In both plots a red line indicates the transition from March to April 2020, which approximates the beginning of the COVID-19 pandemic in Scotland.

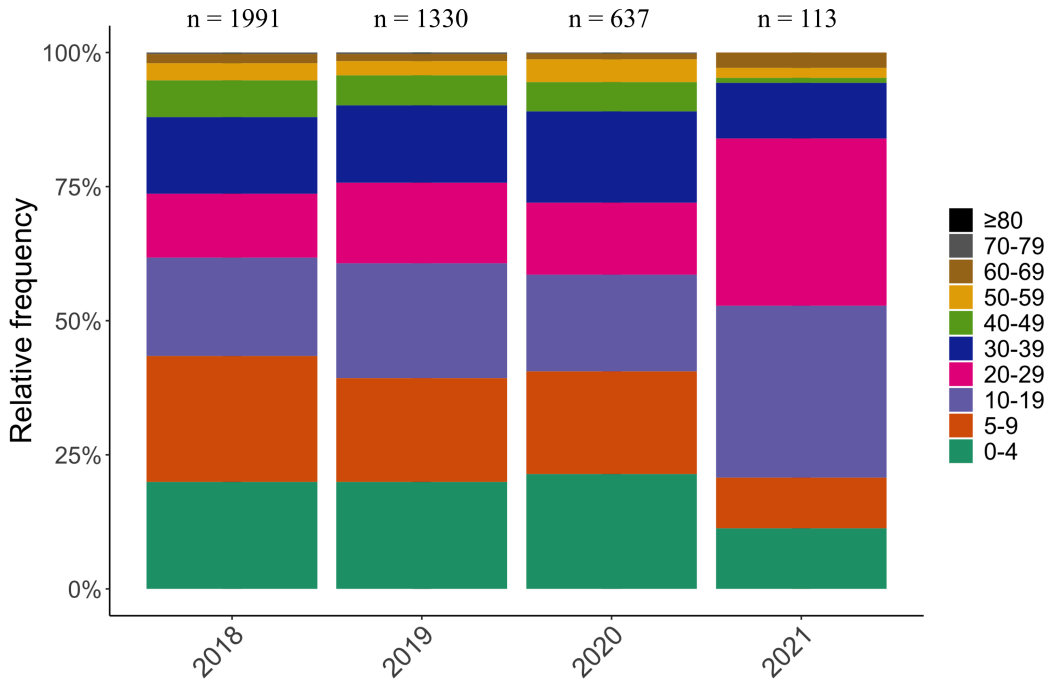


Figure 2.8: Proportion of Group A *Streptococcus* (GAS)-positive throat swabs from specific age groups within the GG&C Health Board from 2018 to 2021. The total number of GAS-positive throat swabs recorded each year is reported above each bar.

2018), with a seasonal pattern of isolation similar to that described for iGAS infections in the whole country. The COVID-19 pandemic was associated with a considerable reduction in the number of GAS-positive specimens submitted to the diagnostic laboratories. The total count of positive samples in March 2020 was 350 while in the following month only 77 specimens were recorded. The lowest count was in April 2021, when just 28 GAS-positive specimens were submitted for diagnostic purposes. From April 2020 to December 2021 the seasonal pattern of diagnoses noticed in the pre-pandemic time was not apparent.

The proportion of GAS-positive throat swabs from specific age groups was calculated and is shown in Figure 2.8. The majority of positive throat swabs (70-80%) was collected from individuals younger than 30 years of age. In 2018, 2019 and 2020 the age-specific proportions of positive throat swabs was found to be comparable. In 2021, however, a larger proportion of GAS-positive throat swabs was collected from individuals of the 10-30 age group.

Resistance rates towards clarithromycin, clindamycin, doxycycline and penicillin each year from 2018 to 2021 are reported in Figure 2.9. In, 2018, 2019 and 2021 the highest rates of resistance were towards doxycycline, whereas in 2020 clarithromycin resistance was marginally the most prevalent form of resistance reported. The rate of resistance towards each antibiotic was consistently lower than 20% in 2018, 2019 and 2020. However, in 2021, every antibiotic tested, with the exception of penicillin, was associated with a resistance rate higher than 25%. No penicillin resistance was detected in the time-frame of the study.

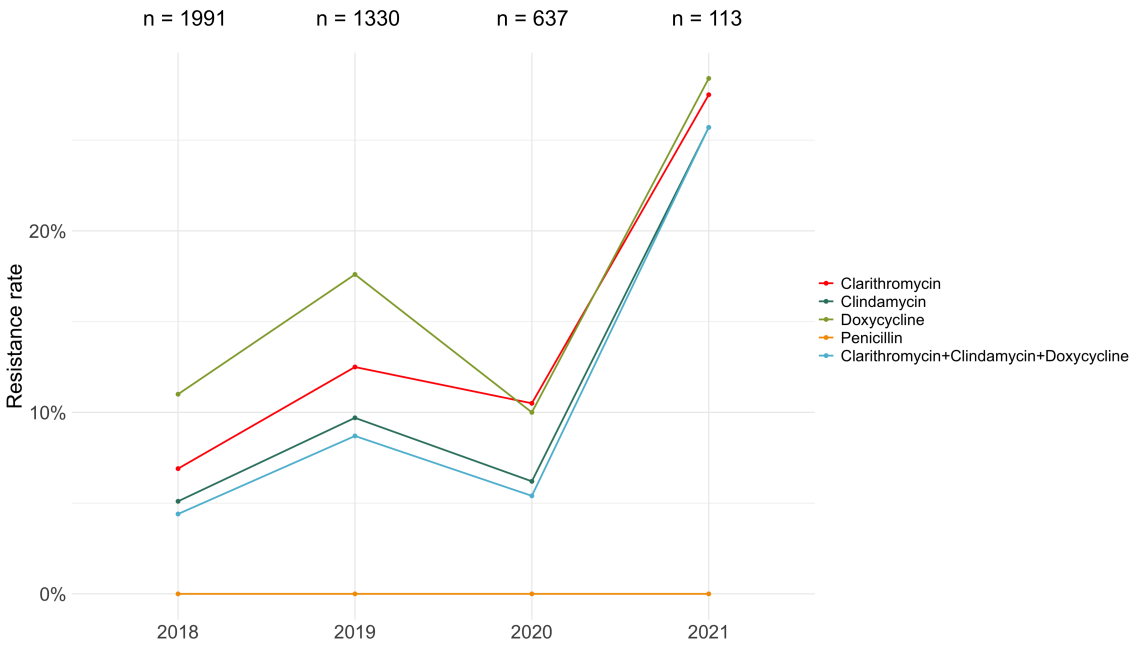


Figure 2.9: Rate of resistance to clarithromycin, clindamycin, doxacycline and penicillin among the Group A *Streptococcus* (GAS) strains isolated from throat swabs in the GG&C Health Board area. The total number of GAS isolates collected each year is reported on top.

The proportion of GAS-positive throat swabs collected from female and male patients was calculated, revealing an over-representation of female specimens. When the gender-specific proportion of positive swabs was calculated for the underage (<18 years) and adult population ($\geq$ 18), a clear difference was noticed (Figure 2.10). While nearly the same proportion (45-55%) of positive throat swabs were collected from males and females younger than 18 years of age, approximately 30% of the adult-derived swabs were collected from males. A z-test for equality of proportions revealed that this difference was statistically significant (p

$< 2.2x10^{-16}$), suggesting that different behaviours in the male and female populations might influence the number of specimens collected from men and women.

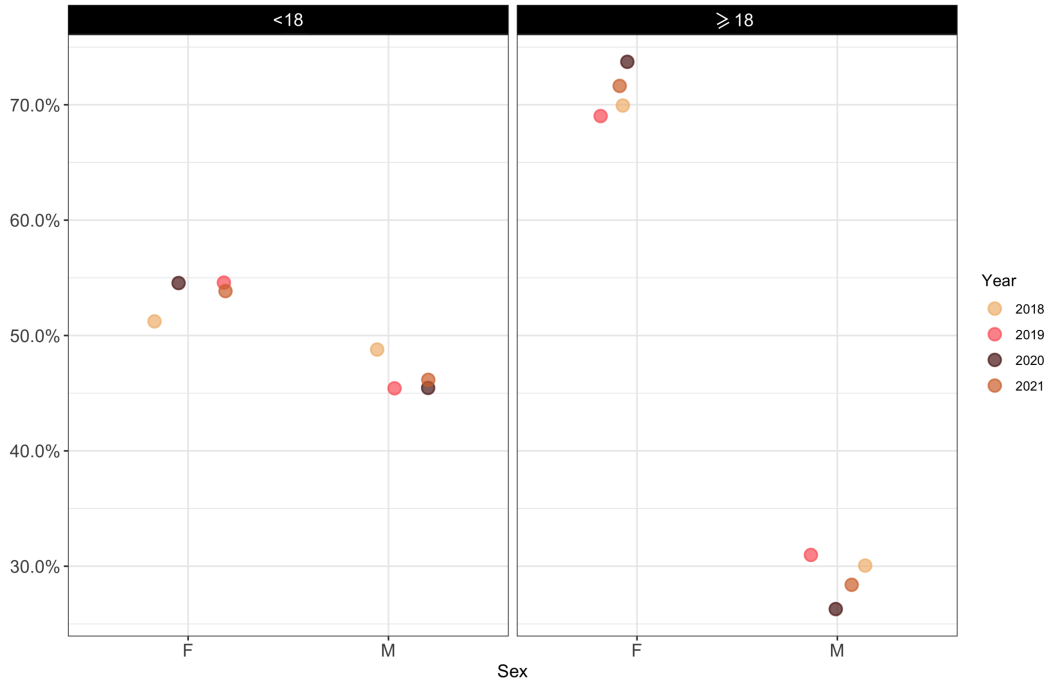


Figure 2.10: Proportion of Group A *Streptococcus*-positive throat specimens submitted to diagnostic laboratories of the GG&C Health Board from 2018 to 2021 in underage (<18) and adult (≥18) males and females.

2.4 Discussion

In this chapter, epidemiological data on invasive GAS infections in Scotland from 2014 to 2021 were presented, visualised and explored. The main focus of this work was investigating the epidemiology of iGAS disease, as it is the only GAS-associated form of infection that is notifiable in Scotland. In this study, all cases of iGAS disease referred to SMiRL from 2014 to 2021 and related metadata, namely date of bacterial isolation, age of patients, strain *emm* type and specimen type were available for analysis. Enhanced surveillance data for invasive infections, which are routinely collected by PHS, could not be accessed due to the pressure that the COVID-19 pandemic exerted on the Scottish healthcare system. Data on all GAS-positive specimens from the GG&C Health Board from 2018 to early 2022 are also presented in this chapter, with particular attention to GAS-positive throat swabs and the characteristics

of individuals sampled.

2.4.1 Incidence of iGAS in Scotland

The yearly incidence of iGAS disease in the Scottish population before the COVID-19 pandemic was as low as 3.9 cases per 100,000 population (2017) and as high as 8.1 cases per 100,000 population (2018). Similar incidence rates have been reported in Canada from 1999 to 2004 (Laupland et al., 2006), in Norway from 2006 to 2009 (Kittang et al., 2011) and in New Zealand from 2002 to 2012 (Williamson et al., 2015). Across the whole of the UK, a lower rate of 3.3 per 100,000 people was documented in 2003-2004 (Lamagni et al., 2008b) although more recent nation-wide data have not been published. Lower rates have also been described in certain regions of Canada between 2001 and 2013 (Athey et al., 2016). Higher iGAS incidence rates appear to be uncommon in high-income countries, with the exception of regions with high a indigenous population density in Canada, the USA, Australia and New Zealand (Carapetis et al., 2005; Athey et al., 2016; Efstratiou and Lamagni, 2016). In the years immediately preceding the COVID-19 pandemic, the iGAS yearly incidence in Scotland was similar to that observed in other high-income countries in recent years. However, over the two years of the COVID-19 pandemic, iGAS incidence rates in Scotland dropped, likely due to the lockdown and social distancing measures in place to contain the spread of SARS-CoV-2. Since the main transmission routes of GAS appear to be person-to-person contact and aerosol droplets (Avire et al., 2021), reduced human interactions likely prevented the spread of GAS strains in the population. Pandemic-related measures, however, might have also reduced GAS incidence indirectly by suppressing the spread of respiratory viruses, which appear to facilitate the occurrence of GAS infections in previously colonised people (Herrera et al., 2016). According to a recently published study, the low rate of *Streptococcus pneumoniae* infections in Israel in 2020 was not associated with a reduction in *S. pneumoniae* carriage and circulation but was instead found to be associated with the disappearance of certain respiratory viruses, such as influenza (Danino et al., 2021). Since *S. pneumoniae* and GAS are phylogenetically related and both share the same ecological niche (Bessen, 2009; Margolis et al., 2010), which is the human upper respiratory tract, it is reasonable to hypothesise that a similar interaction may be responsible for the recent reduction in GAS infections. The annual infection rates of severe influenza

and iGAS disease in Scotland did not show similar trends across the years (Figure 2.1), suggesting that other variables, such as other respiratory viruses and chickenpox (Lamagni et al., 2008a), may have influenced the annual burden of iGAS infection in Scotland. The number of confirmed new cases of iGAS disease per month prior to the COVID-19 pandemic revealed a seasonal pattern characterised by a high incidence over winter and spring and a reduction in the case count during summer and autumn. This seasonal variation has also been described in other epidemiological studies on iGAS infections (Lamagni et al., 2008a, 2009; Olafsdottir et al., 2014). From April 2020 to December 2021, during the COVID-19 pandemic, a reduction in the number of newly confirmed cases per month was noted together with a disruption of the previously observed seasonal pattern, with no clear distinction between winter/spring and summer/autumn, possibly due to the overall low disease incidence. The age-specific incidence of iGAS infections in the Scottish population before the COVID-19 pandemic appeared to mimic a W shaped curve, with two lateral peaks (0-4 years and $\geq$80 years) and a third peak representing the 30-39 age group (Figure 2.3). This finding is in contrast with previously published data on iGAS disease in different European countries, where the age-specific incidence was described as a J shaped curve with only two peaks among infants and the elderly (Lamagni et al., 2008a; Efstratiou and Lamagni, 2016). Other published data, however, have documented in Canada, Denmark and the UK a higher incidence among individuals between 30 and 39 years of age (Davies et al., 1996; Lamagni et al., 2008a; Luca-Harari et al., 2008; Lepoutre et al., 2011). The authors of a study on iGAS epidemiology in Ontario, Canada, revealed that the higher-than-expected incidence observed among people aged 30-39 was due to puerperal sepsis in women and arthritis in men (Davies et al., 1996). In a study of iGAS in Denmark, an incidence peak was also documented among individuals 30-39 years of age, with most of the cases involving female patients (Luca-Harari et al., 2008). Puerperal sepsis, however, occurred only in 35% of female cases aged 30-39 (Luca-Harari et al., 2008), suggesting other factors may have been responsible for the high incidence rate in that age group. Data regarding the age-specific incidence in the UK show an unusually high incidence rate in males aged 25-44 (Lamagni et al., 2008a). In the current study, clinical data were not available and so it is impossible to investigate the association between disease manifestations and infection incidence in different age groups. A recently published report on the prevalence of drug use in Scotland

identified that the highest rate of drug misuse was among people aged 25-34 (https://www.isdscotland.org/health-topics/drugs-and-alcohol-misuse/drugs-misuse/prevalence-of-problem-drug-use/). Since drug use, particularly the use of injecting drugs, is a recognised risk factor to develop iGAS disease (Lamagni et al., 2008c), drug misuse could be one possible explanation for the high incidence of infection of people in their thirties, although people in their twenties do not seem to be equally affected. Further investigations are required to clarify the association between age, specifically in the age group 30-39, and iGAS disease in Scotland. The incidence during the first year of pandemic (2020) was lower across all age ranges compared to previous years, although a similar age-specific pattern was still noticeable. During the second year of pandemic, however, the total number of iGAS cases was so low that any age-specific incidence peak was likely to be due to chance.

2.4.2 Scottish iGAS *emm* types

Since the start of the COVID-19 pandemic, lockdowns and social restrictions have had an unprecedented impact on contemporary society and it was unknown how this would influence the epidemiology of other infectious diseases. With that in mind, the overall iGAS *emm* type diversity during and prior to the pandemic was first assessed by calculating the SID with 95% CI for each study year (Figure 2.4). SID values calculated in the present study are similar to those estimated for iGAS disease in Asia, Latin America, Middle East and other high-income countries but lower than those for Africa and the Pacific region (Steer et al., 2009). SID values in Scotland showed a tendency to increase from 2014 to 2019, with particularly high values in 2018 and 2019, possibly due to the upsurge of iGAS cases in 2018. SID values for 2020 and 2021 are comparable to those of previous years, suggesting that the COVID-19 pandemic did not have a significant effect on the overall invasive *emm* type diversity in Scotland. The effect of the pandemic on the occurrence of invasive *emm* types was also assessed using the BC dissimilarity index to measure year to year compositional change. The BC dissimilarity index is normally used to compare the population richness between two geographical sites. In this case, however, it was used to compare different time periods (years) for the same geographical area, i.e. the whole Scotland. Since the BC dissimilarity index is influenced by the size of the samples to compare, and since differences

in the total number of invasive strains are to be expected from year to year (particularly in the face of the pandemic), the index was calculated with respect to the relative abundance of each *emm* type rather than their absolute count. Thus, the adjusted BC dissimilarity indices represented a measure of *emm* type turnover from year to year. A considerable change in *emm* type turnover was noted from 2017 to 2018. The BC index for 2019 and 2020 was not dissimilar than that measured in previous years. The index for 2020 and 2021, however, was considerably higher than any other index from previous years. These results indicate that the COVID-19 pandemic had a noticeable effect on the year-on-year *emm* type turnover from the first to the second year of pandemic, although the same effect was not apparent from the year preceding the pandemic to the first year of pandemic. It should be appreciated, however, that the total number of invasive disease strains isolated in 2021 was very small (n = 46), meaning *emm* type relative abundance measurements may provide spurious results. The six most prevalent *emm* types over the eight years encompassed by this study were *emm*1 (20%, 414/2020), *emm*89 (8%, 177/2020), *emm*12 (7%, 135/2020), *emm*4 (5%, 97/2020), *emm*76 (5%, 91/2020) and *emm*28 (4%, 86/2020). If we consider just the pre-pandemic years, however, the only *emm* types that were consistently among the ten most prevalent per year were *emm*1, *emm*89, *emm*12, *emm*4 and *emm*28. Together with *emm*3, these five *emm* types are considered the most prevalent among invasive disease cases in high-income countries worldwide (Lepoutre et al., 2011; Steer et al., 2009; Efstratiou and Lamagni, 2016; Villalón et al., 2021). Although the total count of invasive *emm*76 isolates was high, this *emm* type virtually disappeared after 2015. Similar to what was observed for *emm*76, temporary upsurges of other *emm* types were observed from 2014 to 2021 (Figure 2.6). The reasons for these *emm*-specific short-term upsurges are not always clear. Invasive disease outbreaks due to uncommon *emm* types have been documented occasionally in high-risk communities such as patients of the same hospital, inmates of the same correctional facility or injecting drug users of the same area (Levy et al., 2003; Lamagni et al., 2008c; Trell et al., 2020). On other occasions, however, uncommon *emm* types appear to be temporarily successful at spreading and causing disease in the general population without obvious underlying epidemiological connections (Darenberg et al., 2007; Luca-Harari et al., 2008; Chiang-Ni et al., 2011). The introduction of novel *emm* types to which most of the population has never been exposed has been suggested as a mechanism behind the sporadic upsurge of iGAS cases with uncommon

emm types (Luca-Harari et al., 2009; Southon et al., 2020). Most of the uncommon *emm* types isolated in Scotland from 2014 to 2021, however, were responsible for a very low proportion of the annual iGAS disease burden. Based on a visual inspection of Figure 2.6 and in line with the BC dissimilarity indices previously described, the relative frequency of invasive *emm* types recorded in 2020 was comparable to that of 2019. In 2021, however, the *emm*-specific invasive disease burden did not resemble that of 2020 and 2019. The most evident change was in the prevalence of the five dominant *emm* types, those that were consistently frequent across the pre-pandemic years and which are considered globally relevant. While the 2021 disease burden of *emm*28 and *emm*89 was similar to that of previous years, the relative frequency of *emm*1, *emm*4 and *emm*12 in 2021 was lower than that observed in the past. Importantly, no cases of iGAS disease attributable to *emm*4 and *emm*1, which is the globally predominant *emm* type, were documented in 2021. These two *emm* types, together with *emm*3.93 and *emm*6, were the most commonly isolated genotypes from primary school aged children exposed to scarlet fever in six different locations in London, UK (Cordery et al., 2022). *Emm*1 appeared to be more common in children than in adults in France (Lepoutre et al., 2011) while *emm*4 was significantly more common in children than adults in a study of iGAS epidemiology in Iceland (Olafsdottir et al., 2014). The latter study also reported that *emm*28 was significantly more common in adults (Olafsdottir et al., 2014). One of the main consequences of the COVID-19 pandemic in Scotland was the closure of schools and the strict social distancing measures in place when they re-opened, including the compulsory use of face masks and the requirement to self-isolate in the event of a student testing positive for SARS-CoV-2. These circumstances may have caused a progressive decline in the spread and overall prevalence of *emm* types more commonly carried by children. Since *emm*1 and *emm*4 appear to be very common strains circulating among children, it may be hypothesised that the COVID-19 restrictions had a particularly strong disruptive effect on these *emm* types.

2.4.3 iGAS isolation site

The majority of iGAS strains isolated in Scotland from 2014 to 2021 originated from blood samples, followed by cutaneous tissue and respiratory tract specimens (Table 2.1). The proportions of each specimen type before and during the COVID-19 pandemic were comparable, suggesting the relative frequency of iGAS clinical manifestations were unchanged during

2020 and 2021. It should be appreciated, however, that clinical data for the invasive cases considered in this study were not available and that iGAS referral criteria may have biased the type of samples collected (`https://www.legislation.gov.uk/asp/2008/5/schedule/1/2010-01-01`).

2.4.4 GAS isolations from the GG&C Health Board

As occurred for iGAS cases, the submission of GAS-positive specimens in the GG&C Health Board suddenly dropped from March to April 2020, when the COVID-19 pandemic was officially declared and the first national lockdown enforced. For the rest of 2020 and throughout 2021, the number of positive specimens per month remained consistently low. Over the first few months of 2022, which corresponded with a progressive reduction of the COVID-related rules in Scotland, a slow increase in the number of specimens submitted each month was noted. The majority of GAS-positive throat swabs (around 75%) submitted each year were collected from patients younger than 30 years of age. Older patients were swabbed less frequently, possibly because they are more prone to develop severe forms of infection that require secondary care interventions. The proportion of GAS-positive throat swabs per age group per year were comparable from 2018 to 2020. In 2021, however, an increase in the proportion of swabs from individuals aged 10-29 was recorded, which corresponded to 60% of all the swabs that year. However, due to the limited number of swabs submitted to diagnostic laboratories in 2021 (n = 113), even small, randomly occurring differences might be responsible for noticeable fluctuations in the proportions of swabs from different age groups. We cannot exclude, however, that the age-specific changes occurring from 2020 to 2021 were associated with specific variables, such as behavioural patterns in teenagers and young adults. Lower perceived risk of SARS-CoV-2 among teenagers and young adults (Commodari and La Rosa, 2020), for example, might have increased their chances of being exposed to GAS infection when the rest of the population was being more cautious.

Almost all the GAS isolates from throat swabs were tested for antimicrobial sensitivity towards clarithromycin, clindamycin, doxycycline and penicillin. While no isolate was found to be resistant to penicillin, the rate of resistance towards other antibiotics varied across the years. In 2021, a rise in resistance rates towards all antibiotics tested was noted. This is

unlikely due to a real increase in the prevalence of AMR and is probably a consequence of changed diagnostic practices in primary care during the COVID-19 pandemic. Interactions with general practitioners during the pandemic were reduced to a minimal level to limit spread of SARS-CoV-2. It may be assumed that throat swabs were submitted to diagnostic laboratories particularly for challenging cases, such as prolonged, recurring or complicated infections, which all increase the chance of antimicrobial resistance acquisition. Excluding 2021, resistance rates ranged from 10 to 18% for doxycyline, from 7 to 13% for clarithromycin and from 5 to 10% for clindamycin. The rates of resistance detected among GAS isolates from the GG&C Health Board in 2018, 2019 and 2020 were comparable to those previously described in the literature (Arvand et al., 2000; Imöhl et al., 2010).

The proportion of GAS-positive throat swabs from males up to 18 years of age was significantly higher than those from adult males. Conversely, the proportion of GAS-positive throat swabs from females up to 18 years of age was significantly lower than that from adult females. Since gender is not considered a risk factor for GAS colonisation and infection (Avire et al., 2021), this discrepancy may be a consequence of different behavioural patterns in adults. Throat swabs are mainly collected from non-invasive upper respiratory infections, which are frequently mild and self-limiting and rarely require secondary care interventions. Contacting a physician and discussing diagnostic options, thus, is often a matter of personal choice. Underage patients are unlikely to make autonomous decisions about their own health. Unsurprisingly, similar numbers of throat swabs from females and males up to 18 years of age were submitted to diagnostic laboratories of the GG&C Health Board each year from 2018 to 2021. Among adults, who are generally responsible for deciding whether to seek medical intervention or not, males appeared to be significantly less likely than females to have a throat swab submitted for diagnostic purposes. These findings match published data on the different help-seeking behaviours of adult males and females (Galdas et al., 2005).

2.4.5 Summary

In this chapter, the epidemiology of iGAS infections in Scotland from 2014 to 2021 was presented, with a particular focus on the effect of the COVID-19 pandemic and related restrictions. A reduction in the iGAS annual incidence was detected during the two years of the

COVID-19 pandemic, particularly in 2021. The volume of new iGAS strains isolated each month from January 2014 to December 2021 revealed a seasonal pattern of high incidence in winter and spring and low incidence in summer and autumn. Throughout the pandemic, however, no clear differences were apparent and a consistently low number of new invasive strains was recorded each month regardless of the time of the year. The age-specific incidence of iGAS disease in Scotland showed higher incidence rates for infants, the elderly and people aged 30-39 years. It is unclear why individuals in their thirties develop iGAS infections more frequently than others in Scotland, but previous studies have suggested an association with pregnancy, puerperal sepsis and this particular age group. Another possible explanation could be the high rate of drug misuse in people in their late twenties and early thirties. The annual *emm* type diversity in Scotland was comparable throughout the study period. The year-on-year *emm* type turnover, however, showed a degree of variability, being higher in the transition of the first to the second year of pandemic. This was partially confirmed when the relative annual frequency of isolation of invasive *emm* types was assessed. Two of the dominant *emm* types, namely *emm*1 and 4, did not cause any case of invasive disease during 2021, while other *emm* types were associated with a higher-than-usual disease burden. Overall, the effects of the COVID-19 pandemic were most noticeable in 2021, suggesting that the most profound change in the epidemiology of iGAS infections took several months to become evident. A considerable reduction in the number of GAS-positive specimens submitted each month to diagnostic laboratories of the GG&C Health Board during the COVID-19 pandemic compared to pre-pandemic was observed. A demographic change in the proportion of GAS-positive throat swabs in 2021 was also noted, with higher percentages of specimens from teenagers and young adults than in previous years. Antimicrobial susceptibility rates for throat swab isolates were comparable to those described in previous studies. In 2021, however, rates were particularly high, potentially due to different diagnostic practices influenced by the COVID-related restrictions. A significant difference in the proportion of male and female positive throat swabs collected from underage and adult individuals was detected. This may be caused by different behavioural patterns in adult males compared to underage males, who are less likely to take autonomous decisions regarding their health needs. In conclusion, the epidemiology of GAS disease, particularly invasive infections, in Scotland was significantly impacted by the COVID-19 pandemic and it is noteworthy that

emm type 1, which was the most frequently associated with invasive disease, was not reported in 2021.

Chapter 3

Genomic characterisation of *Streptococcus pyogenes emm*5.23, a recently emerged genotype causing invasive disease in Scotland

3.1 Introduction

More than 200 *emm* types and subtypes of *Streptococcus pyogenes* have been identified to date with some *emm* types being associated with disease more frequently than others. In high income countries, which provide the majority of the available data on *S. pyogenes* epidemiology, *emm* types 1, 12, 28, 3, 4 and 89 constitute almost 60% of all cases of GAS infection (Steer et al., 2009). The reasons for a higher infectivity and virulence of some *emm* types compared to the wider GAS population have been investigated from a molecular point of view and in some cases demonstrated. The high level of virulence of GAS *emm* type 1, for instance, is due to the horizontal acquisition of virulence factors Sda2, SpeA2 (derived from a single point mutation of SpeA1), NAD+-glycohydrolase and SLO (Nasser et al., 2014). In recent years, the emergence of a new *emm*1 lineage, named M1UK and distinguished by 27 SNPs that confer an increased production of the speA toxin *in vitro*, was reported (Lynskey et al., 2019). The high number of cases and the disease severity associated with *S. pyogenes*

*emm*89 was shown to be caused by mutations in transcriptional regulation regions of the genome and by the loss of genes encoding the hyaluronic capsule (Beres et al., 2016). The acquisition of multi-drug resistance, which is often associated with horizontal gene transfer events, is also thought to play a role in successful spread of pathogenic GAS clones (Davies et al., 2015). Overall, published data suggest that variation in the *S. pyogenes* genome arises through both recombination and mutation events, and that some genotypes are associated with a greater virulence phenotype (Beres et al., 2002; Nasser et al., 2014; Beres et al., 2016; Feng et al., 2016; Kachroo et al., 2019). The rate of vertically acquired mutations in *S. pyogenes* was estimated to range from 1.3 to 2.1 SNPs per genome per year (Nasser et al., 2014; Turner et al., 2015) and the rate of recombination to mutation was estimated to be 1.03 (Lee and Andam, 2022).

Since 2018, a rarely isolated genotype of *S. pyogenes*, *emm* subtype 5.23, has been involved in several cases of invasive disease in Scotland. While GAS *emm*5 cases represent less than 2% of the *S. pyogenes* disease burden in high income countries (Steer et al., 2009), *emm*5.23 in Scotland caused 4.71% and 9.82% of all iGAS cases in 2018 and 2019, respectively (SMiRL). Although not commonly isolated, this *emm* subtype was involved in one outbreak of invasive disease in England (Gossain et al., 2016) and has recently been associated with increased mortality in the North West of England (Blagden et al., 2020). An informal enquiry to public health colleagues in Denmark, Finland and Norway revealed that this genotype had not been implicated in invasive infections in those countries in 2018, 2019 and 2020 (personal communication, Statens Serum Institut, Norwegian Institute of Public Health, Public Health Agency of Sweden).

Aim and objectives The aim of this chapter was to characterise *emm*5.23 isolates in Scotland, considering genotypic and phenotypic determinants that could account for its onset and unexpectedly high disease incidence in 2018 and 2019. The following objectives were thus formulated:

- To produce at least one complete and closed genome of *emm*5.23 to use as a reference sequence.
- To detect genomic determinants of AMR in all *emm*5.23 WGS.

- To characterise the AMR phenotype of a subset of *emm*5.23 isolates.
- To determine the presence and characteristics of MGEs integrated in the *emm*5.23 genomes.
- To establish the carriage of virulence genes in the *emm*5.23 genomes.
- To consider the acquisition of polymorphisms that may have an impact on the phenotype of the *emm*5.23 isolates.
- To investigate the *emm*5.23 population structure.
- To characterise differences in the overall gene expression between the two main *emm*5.23 genotypes identified through a preliminary population structure analysis.

3.2 Methods

3.2.1 Epidemiology of iGAS *emm*5.23 in Scotland

In Scotland, a total of 58 cases of *S. pyogenes* invasive disease were associated with the genotype *emm*5.23 between 2015 and 2022. The majority of these cases occurred in 2018 (n = 21) and 2019 (n = 27). Eight cases were reported in 2020 and only a single case was reported in 2015 and in 2022. Invasive *emm*5.23 isolates were collected from blood (n = 40/58, 69%), cutaneous (n = 10/58, 17%) and respiratory specimens (n = 8/58, 14%). The mean age of the patients affected was 50, and the median age was 55, with a range spanning from 1 day to 93 years.

3.2.2 Bacterial isolation and Illumina whole genome sequencing

Group A *Streptococcus* strains causing invasive disease in Scotland are routinely isolated and collected by SMiRL. Isolates of GAS are then genotyped based on the first 180 bases of the *emm* gene sequence encoding the cell surface M protein, according to the CDC guidelines (`https://www.cdc.gov/streplab/groupa-strep/emm-background.html`). In the present study, we included all available GAS *emm*5.23 strains isolated from normally sterile body sites of symptomatic patients in Scotland up until August 2022 (n =

58). Only one isolate dated back to 2015, with the others being collected in 2018 (n = 21), 2019 (n = 27), 2020 (n = 8) and 2022 (n = 1).

Bacterial genomic DNA was extracted after cell lysis with mutanolysin, lysozyme and proteinase K, using the DNeasy 96 Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA extractions were performed on a QIAsymphony automated instrument (Qiagen, Hilden, Germany). Genomic DNA was quantified with a Qubit 3 Fluorometer, paired-end WGS libraries were generated using a Nextera XT DNA Sample Preparation Kit and Index Kit V2 (Illumina, San Diego, CA). Next generation WGS was performed on an Illumina MiSeq platform at SMiRL, Glasgow.

3.2.3 Antimicrobial susceptibility testing

Antimicrobial susceptibility to a panel of antibiotics commonly used to treat Gram positive infections, namely ampicillin, amoxicillin, clindamycin, ceftriaxone, cefotaxime, doxycycline, erythromycin, levofloxacin, meropenem, moxifloxacin, oxacillin, penicillin G, tetracycline and vancomycin, was measured for a subset of the *emm*5.23 isolates (n=25). The Micronaut-S kit (Merlin, Berlin, DE), which employs the broth microdilution (BMD) method, was used to determine the isolates' minimum inhibitory concentrations (MIC) for the tested antibiotics as per EUCAST guidelines v 3.0 (`https://www.eucast.org/ast_of_bacteria/mic_determination/`). The *Haemophilus influenzae* strain ATCC 49766 served as a quality control for the kit and reagents employed. Initially, pure bacterial cultures were plated and grown on Columbia blood agar plates (Oxoid, Waltham, MA) for 48 hours at 37°C. Then, 3 mL saline solution (0.85% NaCl, pH 5.5 to 6.5) was inoculated with each isolate until the suspension density reached 0.5 McFarland (0.44-0.56), as measured using a DensiCHECK Plus instrument. From each saline suspension, 200 μL were taken and added to an 11.5 mL solution of Micronaut H-Medium broth. Equal volumes of 100 μL of each aliquot of H-Medium broth were then distributed in a 96 wells of a Micronaut-S PHE Co-GP03 plate. Ten μL of each aliquot of H-Medium broth were also plated on a Columbia blood agar plate (Oxoid, Waltham, MA) to check purity. Both the Micronaut plates and the purity plates were incubated for 22-24 hours at 37°C. After that, bacterial growth in the Micronaut plates was measured using a Multiskan FC Mi-

croplate photometer (ThermoFisher Scientific, Waltham, MA) and MIC values were interpreted by the Micronaut MCN6 software according to the EUCAST breakpoint table v 12.0 (`https://www.eucast.org/clinical_breakpoints/`). In case of inconsistent or ambiguous results, MIC values were visually inspected.

A subset of isolates tested by BMD (n=18) were also tested using Vitek 2 technology (Ligozzi et al., 2002). The Vitek measurements were previously generated in diagnostic laboratories of the Greater Glasgow and Clyde area as part of routine workflows.

3.2.4 Sequence assembly and quality check

A pipeline to trim low-quality reads, remove PCR-generated duplicates and perform *de novo* assembly was employed and is reported in Appendix B.2.1. Paired-end reads in fastq format were trimmed using ConDeTri v3.11.1 (Smeds and Künstner, 2011) with default settings. The script FilterPCRdupl was then used to remove redundant read copies that may have emerged in the PCR step. Paired-end libraries of trimmed and filtered reads were subsequently *de novo* assembled using SPAdes v3.11.1 (Bankevich et al., 2012) in the only assembler mode. Assembly quality was assessed with QUAST v5.0.2 (Gurevich et al., 2013). Reads of assemblies with a cumulative length greater than 2.25 Mbp and/or a GC% content higher or lower than 2 standard deviations from the mean were considered potentially contaminated. These were submitted to the KmerFinder v3.2 software of the Centre for Genomic Epidemiology to confirm or rule out contamination (Clausen et al., 2016).

3.2.5 MinION sequencing and hybrid assembly

We selected three strains that had a high Illumina read coverage and a good assembly quality based on QUAST analysis. One of the strains was the oldest isolate in our collection, which was isolated in 2015. The remaining two strains were selected because they were representative of the two main *emm*5.23 groups identified in a preliminary phylogenetic analysis. The chosen strains were plated on Todd-Hewitt Broth (THB) agar (ThermoFisher Scientific, Waltham, MA) and single colonies were inoculated in THB for overnight culture at 37°C. Genomic DNA was extracted from overnight broth cultures using the Wizard Genomic

DNA Purification Kit (Promega, Madison, WI). Genomic DNA was quantified with Qubit and quality (absorbance ration A260/A280) was assessed on a Nanodrop spectrophotometer, and the concentration adjusted to approximately 57 ng/μL to comply with the MinION library DNA input. Genomic DNA was then sequenced using MinION technology (Oxford Nanopore Technologies, Oxford, UK). Libraries were prepared using the Oxford Nanopore Rapid Barcoding Kit SQK-RBK004 (Oxford Nanopore Technologies, Oxford, UK) following the manufacturer's instructions. MinION sequencing was performed at the OHRBID laboratory of the School of Veterinary Medicine of the University of Glasgow. The sequencing run output, which was in fast5 format, was converted to fastq format using Guppy basecaller v3.6.0 (Wick et al., 2019). Guppy was also used for demultiplexing, i.e. to separate the fastq reads according to the isolate from which they originated. These long MinION reads were assembled together with MiSeq generated short Illumina reads using Unicycler v0.4.8 (Wick et al., 2017) in the hybrid assembly mode. Finally, annotations were produced using the National Centre for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline v5.0 (Tatusova et al., 2016). A list of the bioinformatic commands used to analyse the MinION sequences can be found in Appendix B.2.2.

3.2.6 Mobile genetic elements

MGEs were detected using SRST2 v0.2.0 (Inouye et al., 2014) coupled with a published database of *S. pyogenes* phage and integrative and conjugative element (ICE) integrase and virulence genes, as previously described (Southon et al., 2020). A bash script that takes trimmed fastq reads as input files and runs SRST2 against the MGE integrase and virulence gene database was used (Appendix B.2.3). The predicted presence of known integrase and MGE-associated virulence genes detected by SRST2 was then confirmed in the assembled genomes using the Basic Local Alignment Search Tool BLAST v2.9.0 (Camacho et al., 2009). Since each identified integrase gene has one or a few known integration sites within the *S. pyogenes* genome, once integrase genes were detected it was possible to infer the exact position and sequence of the MGE within the three *emm*5.23 closed genomes.

3.2.7 Multi locus sequence typing and virulence gene identification

SRST2 v0.2.0 (Inouye et al., 2014) was used to identify the MLST of the *S. pyogenes emm*5.23 strains. An MLST database for *S. pyogenes* (updated as of the 30/04/2020) was downloaded from PubMLST (Jolley et al., 2018) and SRST2 was run on the *emm*5.23 WGS against this database (Appendix B.2.4).

The presence of virulence genes in the GAS *emm*5.23 genomes was established using the command-line tool ARIBA v3.1.0 (Hunt et al., 2017) coupled with VFDB (Virulence Factors Database) (Chen et al., 2005). Similarly to SRST2, ARIBA allows the identification of coding sequences (e.g. antimicrobial resistance, virulence or plasmid associated genes) by aligning local assemblies against a database. ARIBA was run using default settings. An updated version of VFDB (as of the 01/04/2020) was downloaded and used for virulence gene detection as described in Appendix B.2.4.

Both SRST2 and ARIBA were also used to confirm each isolate's *emm* type, previously assigned by PCR amplification and Sanger sequencing. The database used for this purpose was the *S. pyogenes emm* gene database curated by the Centers for Disease Control and Prevention (`https://www2.cdc.gov/vaccines/biotech/strepblast.asp`).

3.2.8 Antimicrobial susceptibility genotype

The presence of known genes and mutations conferring AMR was assessed using the command-line tool ARIBA with default settings coupled with the comprehensive antibiotic resistance database (CARD) (last updated 20/01/2021) (Alcock et al., 2020), as described in Appendix B.2.4.

3.2.9 Polymorphism detection and phylogenetic analysis

The WGS data from *emm*5.23 isolates originating from 2018, 2019, 2020 and 2022 were screened for the presence of non-synonymous polymorphisms against two reference strains, namely the single 2015 isolate, iGAS426 (accession number: CP067008), and the *emm*5

Manfredo strain (accession: AM295007). Only non-synonymous polymorphisms were considered because, unlike synonymous ones, they determine amino acid sequence changes that can cause phenotypic alterations. Polymorphisms were detected using Snippy v4.4.5 (Seemann, 2015). Snippy uses Freebayes (`https://arxiv.org/abs/1207.3907`) to identify both SNPs and insertions/deletions (indels) from reads and assembled sequences mapped to a reference genome. Snippy also predicts the function of detected variations using snpEff (Cingolani et al., 2012). Particular emphasis was given to the detection of polymorphisms in the *pbp* genes of the *emm*5.23 strains compared to the historical Manfredo strain, which does not contain any known *pbp* mutations associated with reduced susceptibility to β-lactams (Beres et al., 2022). Instructions to use Snippy are presented in Appendix B.2.5.

In order to get meaningful insights into the evolution of the *emm*5.23 population, polymorphism detection and phylogenetic tree construction were carried out on the Scottish *emm*5.23 isolates and on an available sample (n=29) of English *emm*5.23 isolates. The English isolates were included in order to increase the overall sample size and improve the power of the phylogenetic inference. The English isolates were provided by UKHSA and originated from patients with invasive (n=22) and non-invasive (n=7) disease from 2012 to 2020. Ten English isolates had previously been subjected to WGS by UKHSA using Illumina HiSeq technology, so their sequencing raw reads were already available for bioinformatic analysis. For the remaining English isolates (n=19), bacterial samples were delivered to SMiRL and WGS was undertaken as previously described. Genomic sequences were trimmed and assembled as previously described and assembly quality was checked before proceeding with further analyses. The core SNP alignment produced using Snippy was used for phylogenetic tree construction (Appendix B.2.5). In order to limit the phylogenetic inference to vertically acquired genomic variations, SNPs located in the MGEs were excluded from the core SNP alignment used for phylogenetic tree construction. The command-line tool IQ-TREE (Nguyen et al., 2015) was used to build a maximum likelihood phylogenetic tree with standard non-parametric bootstrap analysis using the best-fit substitution model function (Kalyaanamoorthy et al., 2017). The web-based program iTOL (Letunic and Bork, 2021) allowed the visualisation of the phylogenetic tree, which was midpoint rooted.

3.2.10 Transcriptomic analysis

Differences in overall gene expression between the two main genotypes identified through phylogenetic analysis were determined via total RNA sequencing and bioinformatic analysis.

RNA extraction and sequencing

Total RNA from two isolates in late exponential growth phase, each represented by four biological replicates, was extracted. For each of the eight samples, 500 μL of bacterial overnight culture in THB were inoculated into 9.5 mL of fresh THB and grown for six hours at 37°C. Total RNA was then extracted using protocol five of the QIAGEN RNeasy Mini Kit according to the manufacturer's instructions. RNA samples were treated with the QIAGEN RNase-free DNase to remove any contaminating residual DNA. RNA concentration and quality was later assessed using, respectively, Nanodrop and Bioanalyzer technologies.

RNA sequencing was performed at Glasgow Polyomics (University of Glasgow, Glasgow). The total RNA was depleted of ribosomal RNA using the QIAGEN FastSelect (bacteria) system, followed by library creation using the Illumina TruSeq Stranded Total RNA kit. Libraries were then sequenced on an Illumina NextSeq2000 instrument to an average of 10 million paired-end reads, each 100 bp long.

Bioinformatic analysis

Transcriptomic data were analysed on a University of Glasgow Galaxy server (Giardine et al., 2005). Sequencing reads were trimmed using Trimmomatic (Bolger et al., 2014) with default settings, and the quality of trimmed reads determined with FastQC (Andrews, 2010). HISAT2 (Sirén et al., 2014) was then used to align the trimmed reads to the reference genome iGAS426 and HTSeq-count (Anders et al., 2015) was employed to count the number of reads that mapped to each feature of the reference genome. Differential gene expression between the replicates representative of each *emm*5.23 genotype was determined using DESeq2 with default settings (Love et al., 2014). A series of visualisations of the output from DESeq2 were produced with Searchlight2 (Cole et al., 2021). Finally, a gene enrichment analysis of the genes differentially expressed between the two genotypes was carried out using the online tool GSEA-Pro v3 (`http://gseapro.molgenrug.nl/`).

3.3 Results

All supplementary material for this chapter can be found in Appendix B. Isolate and WGS metadata are provided in Table B.1.

3.3.1 AMR phenotype

Only tetracycline resistance was observed in the *emm*5.23 isolates tested. MIC values for the panel of antibiotics tested using BMD are reported in Table 3.1. Interpretations of MIC values, based on the EUCAST breakpoints for both the BMD and VITEK data, are summarised in Figure 3.1.

Table 3.1: Minimum inhibitory concentrations (MIC) expressed as mg/L of 25 Scottish *emm*5.23 isolates tested for antimicrobial susceptibility. Based on the EUCAST v 12.0 breakpoint table, MIC values that correspond to resistance are coloured in red.

Antibiotic class	β-lactams							Fluoroquinolones		Glycopeptide	Lincosamide	Macrolide	Tetracyclines	
Antibiotic	Amoxicillin	Ampicillin	Cefotaxime	Ceftriaxone	Meropenem	Oxacillin	Penicillin G	Levofloxacin	Moxifloxacin	Vancomycin	Clindamycin	Erythromycin	Doxycycline	Tetracycline
GAS breakpoints	0.25	0.25	0.25	0.25	0.25	0.25	0.25	2	0.5	2	0.5	0.5	2	2
Isolate														
S.426	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.453	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤2	=0.5	≤1	≤0.125	≤0.125	≤2	>4
S.458	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	=4
S.451	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.477	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.368	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.359	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.378	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.358	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.352	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	=4
S.367	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	=4
S.487	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.394	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	=0.0625	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.384	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.391	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤1	=0.25	≤1	≤0.125	≤0.125	=4	>4
S.179	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.158	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.134	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.169	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.138	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤1	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.254	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	=0.5	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.270	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.293	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤1	=0.25	≤1	≤0.125	≤0.125	≤2	>4
S.315	≤0.125	≤0.0625	≤0.0625	=0.125	≤0.125	≤0.125	≤0.03125	≤0.5	=0.25	≤1	≤0.125	≤0.125	>4	>4
S.314	≤0.125	≤0.0625	≤0.0625	≤0.0625	≤0.125	≤0.125	≤0.03125	≤0.5	≤0.125	≤1	≤0.125	≤0.125	≤2	>4

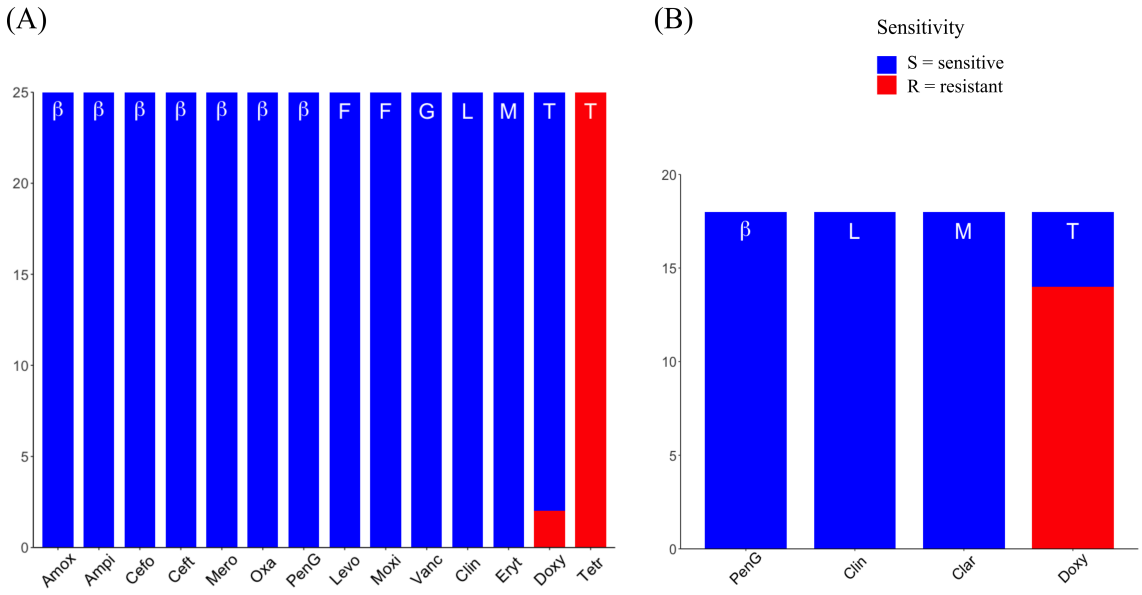


Figure 3.1: Count of sensitive and resistant isolates tested with the broth microdilution technique (A) and VITEK technology (B). In both cases, EUCAST breakpoints were used to interpret MIC data. In the legend, S = sensitive, I = sensitive, increased exposure, R = resistant. Antibiotic names are abbreviated and appear on the X-axis (Amox = Amoxicillin, Ampi = Ampicillin, Cefo = Cefotaxime, Ceft = Ceftriaxone, Mero = Meropenem, Oxa = Oxacillin, PenG = Penicillin G, Levo = Levofloxacin, Moxi = Moxifloxacin, Vanc = Vancomycin, Clin = Clindamycin, Eryt = Erythromycin, Doxy = Doxycycline, Tetr = Tetracycline, Clar = Clarythromycin). White letters on top of each bar represent antibiotic classes (β = β-lactams, F = fluoroquinolones, G = glycopeptides, L = lincosamides, M = macrolides, T = tetracyclines). All isolates tested had the same AMR genotype.

3.3.2 Illumina and MinION sequencing

Four Illumina-sequenced WGS of Scottish *emm*5.23 strains showed signs of contamination and were then removed from further analyses. Long reads WGS data from three *S. pyogenes emm*5.23 isolates produced with MinION technology were merged with short reads derived from Illumina sequencing to generate three complete closed genomes. Sequence characteristics of the newly generated closed genomes, which were deposited to the NCBI database under the GenBank accession numbers CP067008 (iGAS426), CP067009 (iGAS391) and CP067010 (iGAS376), are presented in Table 3.2.

Table 3.2: Sequence characteristics of the three closed genomes of *Streptococcus pyogenes emm*5.23 generated by hybrid assembly of long MinION reads and short Illumina reads.

Isolate	**iGAS376**	**iGAS391**	**iGAS426**
Year of isolation	2018	2019	2015
Source	Blood	Blood	Blood
Genome size (bp)	1,897,124	1,897,129	1,897,111
GC content (%)	38.6	38.6	38.6
No. of reads	28,201	95,719	90,232
No. of sequenced bp	192,673,926	451,467,050	379,063,977
Genome coverage	101x	237x	199x
No. of genes	1,913	1,913	1,913
No. of CDSs	1,824	1,824	1,824
No. of coding genes	1,775	1,775	1,775
No. of rRNAs (5S, 16S, 23S)	6, 6, 6	6, 6, 6	6, 6, 6
No. of tRNAs	67	67	67
Accession number	CP067010	CP067009	CP067008

3.3.3 Mobile genetic elements

The presence of MGEs and their associated virulence genes was determined using SRST2 coupled with a published database of *S. pyogenes* integrase and virulence genes. The integrase genes, which are associated with specific integration sites in the GAS genome, allowed the detection of MGEs.

The same five MGEs were found in each *emm*5.23 genome (Figure 3.2). Four of these elements were prophages while the last one was a composite element made of a prophage and an ICE. The following virulence genes were detected in MGE regions: *sdn*, *spd1*, *spd3*, *hylP*, *spd4* and *speC*. A BLAST search against the NCBI database allowed an additional characterisation of the MGEs detected. While the prophage regions show high similarity with prophages already described in GAS strains, the composite element was not found in other publicly available GAS genomes. The sequence displaying the greatest similarity to the *emm*5.23 ICE in the NCBI database (99% sequence identity) is integrated in the genome

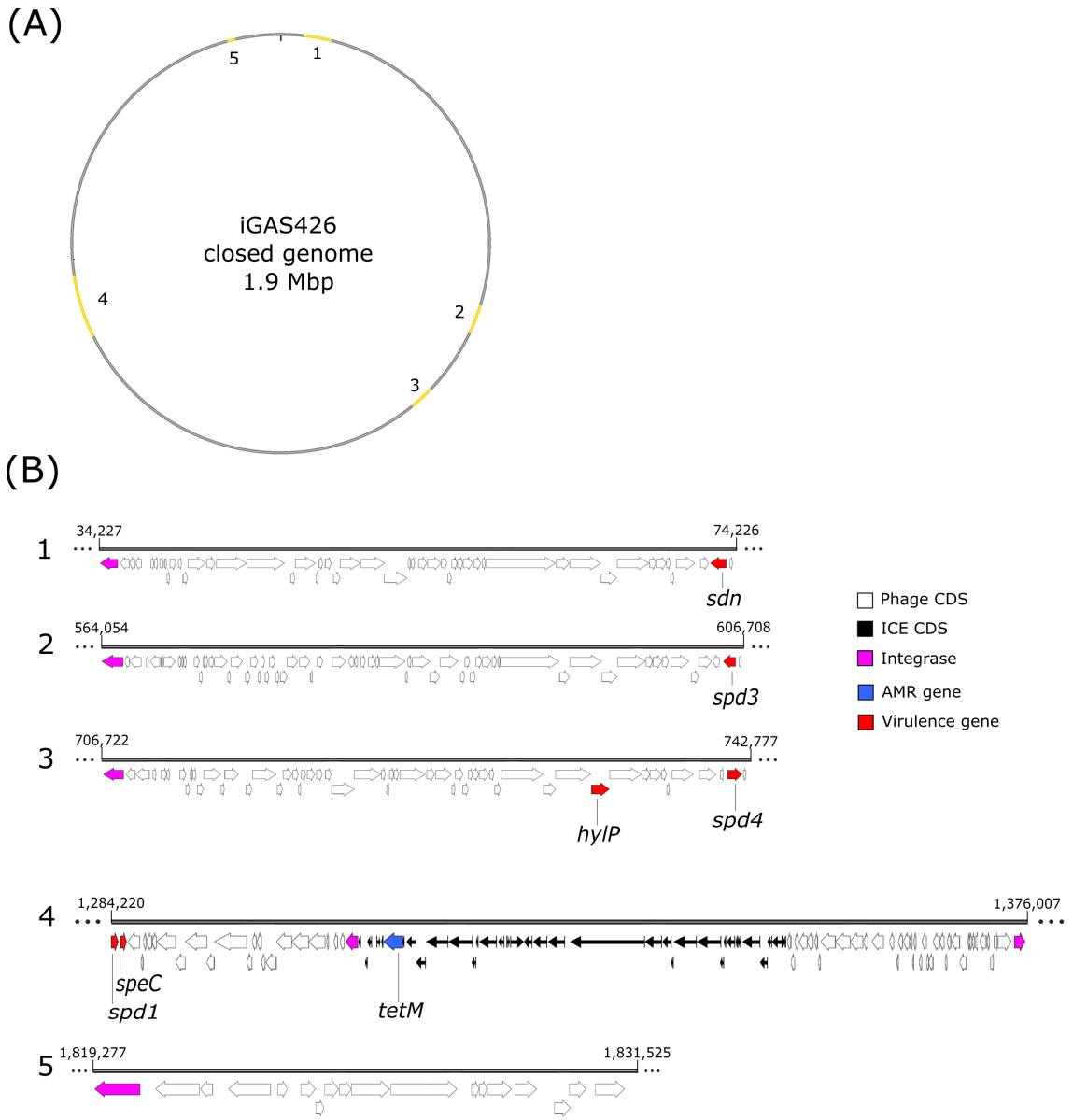


Figure 3.2: (A) Location, in the reference strain iGAS426, of the 5 mobile genetic elements (MGEs) detected in each GAS *emm*5.23 genome. (B) Characteristics of the coding sequences (CDS) belonging to each MGE. MGEs 1, 2, 3 and 5 consist entirely of phage-origin sequences, while MGE 4 has both phage and integrative and conjugative element-derived sequences. MGEs orientation in the figure is 5'-3'. Integrase, antimicrobial resistance and virulence genes are highlighted. Nucleotide positions indicating the beginning and end of each MGE are also reported.

of a strain of *Filifactor alocis* (accession CP002390), a human opportunistic pathogen that colonises the oral cavity. Based on the NCBI non-redundand nucleotide database matches identified by BLAST, this ICE appears to derive from multiple recombination events that involve at least two conjugative transposons and a bacterial intron, carries the AMR gene

tetM, which confers resistance to tetracyclines.

3.3.4 MLST and virulence gene identification

Emm type prediction using SRST2 and ARIBA confirmed that all the strains analysed were *emm* subtype 5.23. SRST2 also revealed that all the *emm*5.23 isolates belonged to MLST 99.

The virulence genes identified by ARIBA from the full VFDB database are reported in Table 3.3. All 54 isolates of *emm*5.23 shared the same virulence gene profile. Most of the virulence genes detected (26/44) are thought to be part of the GAS accessory genome (Davies et al., 2019), reflecting the variability of GAS pathogenic mechanisms.

Table 3.3: List of virulence genes and associated virulence factors identified in this study along with their functions. Virulence genes that are present in > 99% of the GAS genomes included in a work by Davies et al. are considered part of the GAS core genome (Davies et al., 2019).

Virulence genes	Virulence factors	Main virulence function(s)
	Core genome	
eno	Enolase	Adhesion
hylA	Hyaluronidase	Invasion
plr/*gapA*	Plasmin-binding protein	Adhesion
ropA	Trigger factor	Virulence factors maturation
sagA,*B*,*C*,*D*,*E*,*F*,*G*	Streptolysin S	Invasion
scpA	Streptococcal C5a peptidase	Immune evasion
scpC	Serine protease SpyCEP	Immune evasion
sdaB	Streptodornase	Invasion
shr	Streptococcal hemoprotein receptor	Invasion
slo	Streptolysin O	Immune evasion/invasion
speB	Cystein protease SpeB	Immune evasion/invasion
spyA	ADP-ribosyltransferase SpyA	Immune evasion
	Accessory genome	
cfa	CAMP factor	Adhesion/invasion

Table 3.3 continued from previous page

Virulence genes	Virulence factors	Main virulence function(s)
cpa	Pilus ancillary protein I	Adhesion
cppA	C3-degrading protease	Immune evasion
emm5	M5 protein	Adhesion/immune evasion
endoS	Endoglycosidase EndoS	Immune evasion
fbp54	Fibronectin-binding protein Fbp54	Adhesion
graB	GRAB	Adhesion/protection
hasA,*B*,*C*	Hyaluronic capsule	Adhesion/immune evasion
htrA	Serine protease HtrA	Stress response
hylP	Phage hyaluronidase	Invasion
isdE	Heme transporter	Adhesion/immune evasion
lmb	Laminin-binding protein	Adhesion
*spd*1,3,4	Deoxyribonuclease	Invasion
mtsA	Metal ABC transporter	Stress response
prtF2	Surface-anchored protein	Adhesion
sclA	Collagen-like surface protein	Adhesion/invasion
sda3	Streptodornase	Invasion
sdn	DNA endonuclase	Invasion
ska	Streptokinase	Invasion
smeZ	Streptococcal mitogenic exotoxin Z	Inflammation/invasion
speC	Streptococcal pyrogenic exotoxin C	Inflammation/invasion
speG	Streptococcal pyrogenic exotoxin G	Inflammation/invasion

3.3.5 AMR genotype

The same AMR gene profile was shared by all genomes analysed. Only two genes associated with antimicrobial resistance were detected: *lmrP* and *tetM*. Based on the AMR gene content, the *emm*5.23 population was expected to express phenotypic resistance to four antibiotic classes: lincosamides (*lmrP*), macrolides (*lmrP*), streptogramin (*lmrP*) (Putman et al., 2001) and tetracyclines (*lmrP* and *tetM*) (Morse et al., 1986; Putman et al., 2001). Snippy output was used to detect mutations in the *pbp* genes of the *emm*5.23 isolates compared to those of strain Manfredo, which does not contain any known mutations conferring reduced susceptibility to β-lactams (Beres et al., 2022). No differences between the *pbp* genes of reference strain Manfredo and those of the *emm*5.23 strains were found, allowing us to conclude the *emm*5.23 population analysed does not carry any known *pbp* gene mutation associated with reduced sensitivity to β-lactam antibiotics.

3.3.6 Polymorphism detection and phylogenetic analysis

No non-synonymous polymorphisms shared by all the Scottish *emm*5.23 isolates were detected. The majority of the genomes analysed (46/54), however, carried the same three identical non-synonymous polymorphisms. One is a SNP involving the lactate oxidase gene *lctO* (C54A - Phe18Leu). Another is a nucleotide insertion (T -> AT) upstream of the gene *greA*, in position -37. The *greA* gene encodes the transcription elongation factor GreA. The third mutation is a SNP that caused an early stop codon in the gene encoding the enzyme Acetyl-CoA C-Acyltransferase (C112T - Gln38*). When BLASTed against the NCBI database, the three polymorphisms were not identified in any other published sequences.

Phylogenetic analysis of the 54 Scottish and 29 English *emm*5.23 WGS partitioned the sequence dataset into two main groups that for ease of interpretation we refer to as English and Scottish groups (Figure 3.3). The English group comprised 28 of the 29 English isolates and eight of 54 Scottish strains. Within the English group, the average number of core pairwise SNPs was 18 and four pairs of clonal clusters (0 core SNPs) were detected. The Scottish group comprises 46 Scottish isolates and only one English strain. The average number of core pairwise SNPs in the Scottish group was four and six clonal clusters, involving a total

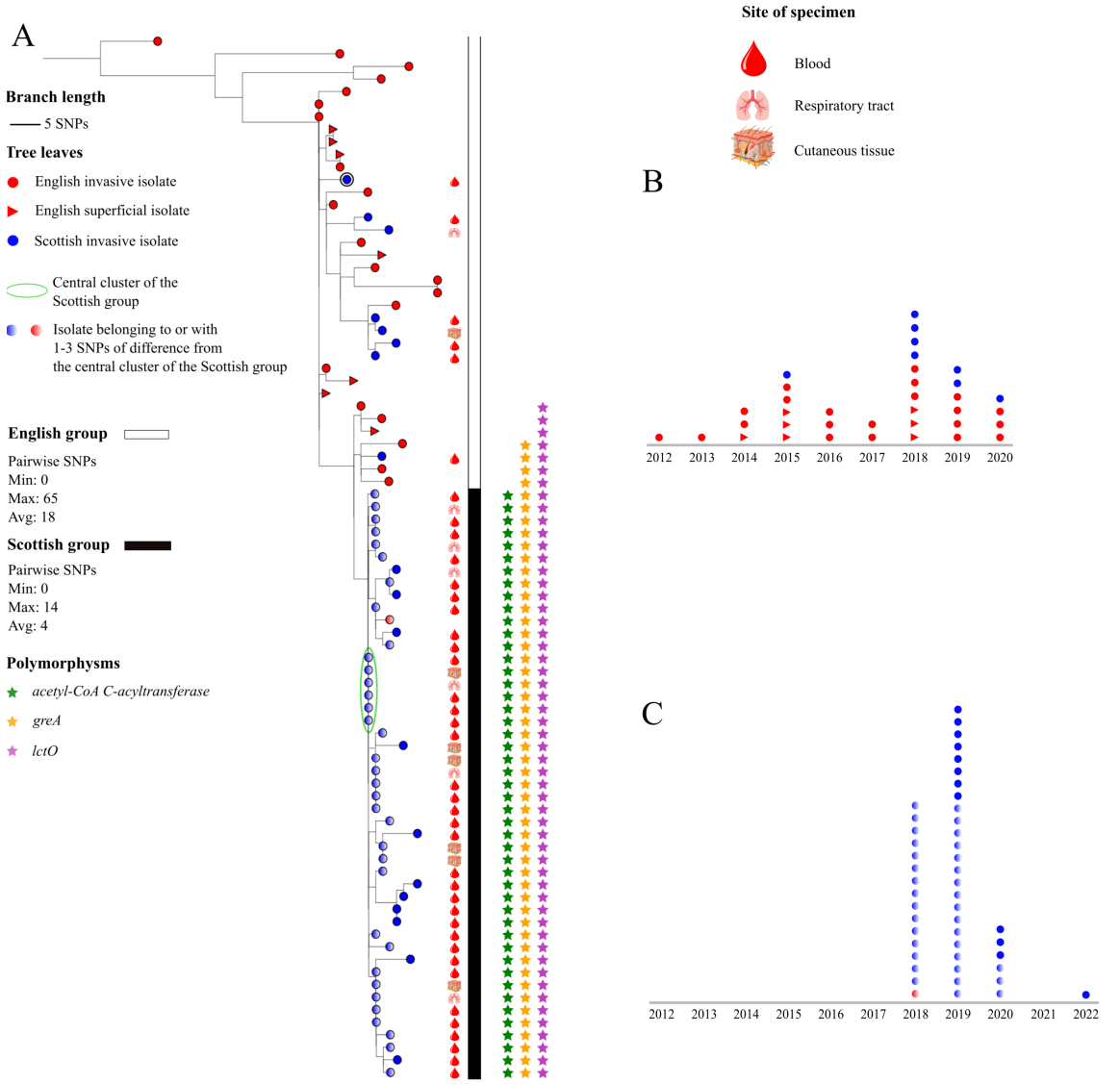


Figure 3.3: A - Core genome maximum likelihood single nucleotide polymorphism-phylogeny showing the evolutionary relatedness of the Scottish *emm*5.23 isolates and a randomly selected sample of English isolates. The tree was midpoint rooted. For ease of interpretation, the part of the tree dominated by English isolates is called the "English group" and the one dominated by Scottish strains is called the "Scottish group". Specimen site is reported next to each Scottish isolate. The presence of three non-synonymous polymorphisms found in all the isolates of the Scottish group is also visually represented. The three isolates whose genomes were closed by hybrid assembly are highlighted on the tree by a black outer circle. These isolates are, from top to bottom, iGAS426 (or S.426), iGAS391 (or S.391) and iGAS376 (or S.376). B - Temporal distribution of isolation of the English group strains. C - Temporal distribution of isolation of the Scottish group strains.

of 22 isolates, were identified. The tree topology suggested that the isolates of the Scottish group originated from a relatively recent common ancestor, represented by a clonal cluster of six isolates, here referred to as the "central cluster" of the Scottish group. Thirty-six of the 47 isolates of the Scottish group (77%) had between 0-3 SNPs differentiating them from the central cluster genotype (Figure B.1). Among the remaining isolates of the Scottish group, no more than seven SNPs differentiating them from the central cluster genotype were detected. All the isolates clustering in the Scottish group shared the three previously described mutations, further confirming of the clonal origin of this phylogenetic group. Only one Scottish isolate clustering in the English group carried two of the aforementioned non-synonymous polymorphisms.

3.3.7 Transcriptomic analysis

A preliminary transcriptomic analysis was performed on two Scottish isolates representative of, respectively, the "English group" and the "Scottish group", with the aim of detecting differences in overall gene expression between the two genotypes. The isolate selected to represent the "English group" genotype was S.134 and the one chosen to represent the "Scottish group" was S.188. Each isolate was represented by four replicates and, thus, a total of eight samples were analysed. As shown in Figure 3.4 A, the overall transcriptomic variation between the English isolate and Scottish isolate is limited, and transcriptomic heterogeneity was apparent for replicates of the English isolate. Only 14 genes were found to be statistically significantly over or under-expressed in the Scottish isolate compared with the English isolate, as indicated in Figure 3.4 B. In particular, nine genes showed higher expression (Figure 3.4 C) and five genes showed lower expression (Figure 3.4 D) in the Scottish isolate compared with the English isolate. Details on the level of expression, p-value and expected products of the 14 statistically significantly differentially expressed genes are reported in Table 3.4.

Gene enrichment analysis did not reveal any functional association of the genes differentially expressed in the English and Scottish isolates, which may be tentatively interpreted to suggest that no phenotypic differences exist between the isolates.

Table 3.4: Gene product and function, as well as differential expression and p adjusted values, of the statistically significantly differentially expressed genes in the Scottish isolate compared with the English isolate. A positive log2fold corresponds to higher expression in the Scottish isolate.

Gene product	Protein function	log2fold	Fold change	p.adj
Phage lysin	Involvement in phage lytic cycle	1.17	2.25	9.03E-05
Pbp2A	Cell wall synthesis	0.78	1.72	9.03E-05
YhgF	Transcriptional accessory protein	0.95	1.93	9.03E-05
Ldh	Fatty acid metabolism	0.89	1.85	9.03E-05
PtsIIA	Sugar uptake	1.18	2.27	4.60E-04
CutC	Copper homeostasis	0.72	1.65	5.40E-03
NrdE	DNA synthesis	0.8	1.74	0.01
AgaS	Lactose catabolism	0.79	1.73	0.03
DnaJ	Chaperone function	0.79	1.73	0.04
Phage protein	Hypothetical protein	-1.04	0.49	4.53E-06
RecO	DNA repair	-0.75	0.59	9.03E-05
Prs	Nucleotide synthesis	-0.75	0.59	1.10E-04
SibA	Putative secreted protein	-0.75	0.59	4.60E-04
Hk	Signal transduction	-0.7	0.62	0.03

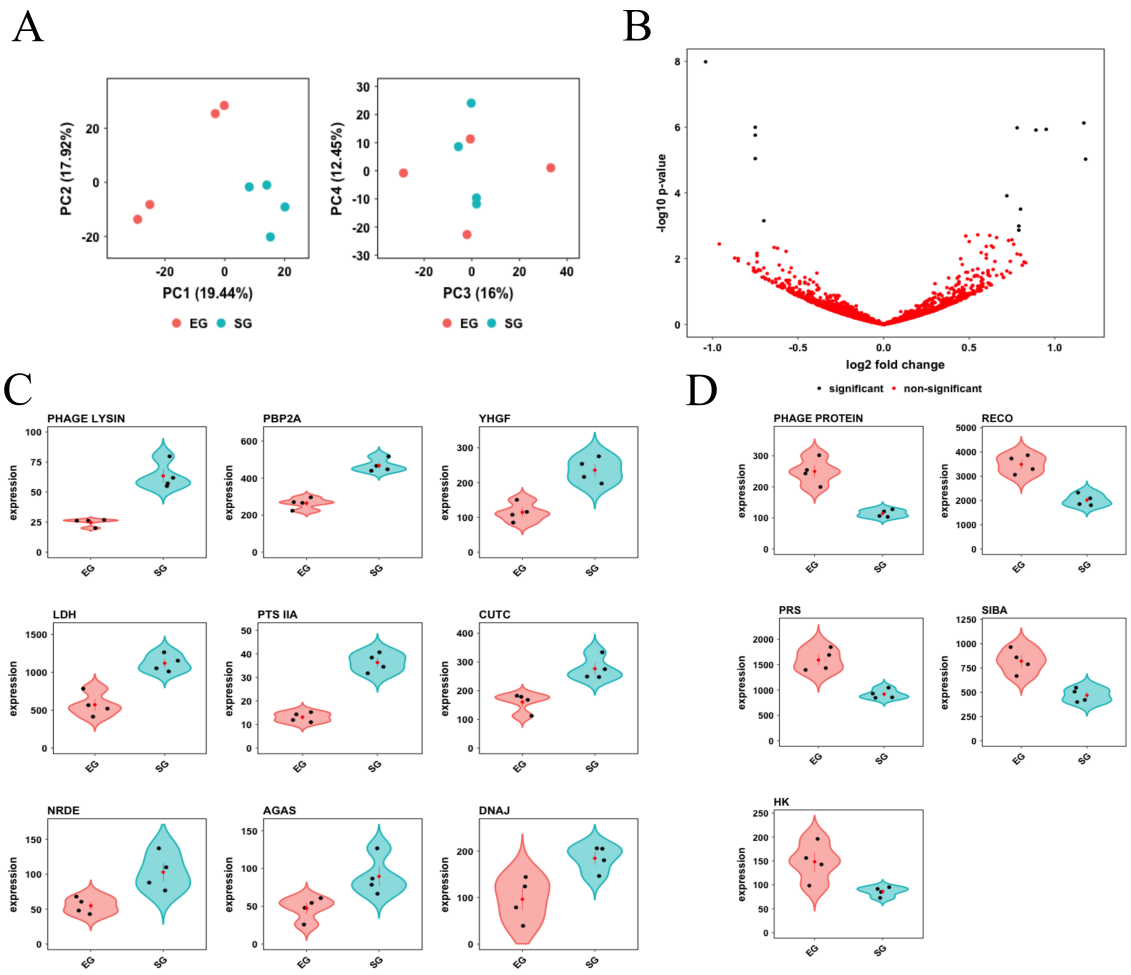

A
PC2 (17.92%)
PC1 (19.44%)
EG SG
PC4 (12.45%)
PC3 (16%)
EG SG
B
-log10 p-value
log2 fold change
significant non-significant
C
PHAGE LYSIN
PBP2A
YHGF
LDH
PTS IIA
CUTC
NRDE
AGAS
DNAJ
expression
EG
SG
D
PHAGE PROTEIN
RECO
PRS
SIBA
HK

Figure 3.4 ***(previous page)*:** Summary plots of the transcriptomic analysis for isolates representative of the English group (EG) and Scottish group (SG). A - Gene expression data Principal Component Analysis (PCA) scatterplots. Showing PC1 vs PC2 (left) and PC3 vs PC4 (right). Individual samples are represented by dots, pink dots are the EG replicates and blue dots are the SG samples. The percentage of total variation explained by each component is given in the x and y-axis as appropriate. To control for over-representation of very highly expressed genes, all gene expression values were scaled on a gene-by-gene basis using the z-score transformation, prior to the PCA. B - Volcano plot for the comparison between SG and EG. Significantly differential genes (p.adj <0.05, absolute log2fold >0.5) are shown in black and non-significant genes in red. A positive fold change indicates higher expression in SG than in EG. C - Violin plots showing the nine statistically significantly (p.adj <0.05) genes with higher expression by log2fold change in the SG samples. Each dot is a single sample, with sample groups given on the x-axis and gene expression on the y-axis. The mean and standard error for each sample group is given as a red dot and line. In the plots, each gene is named after its product. D - Violin plots showing the five statistically significantly (p.adj <0.05) lower genes by log2fold change in the SG samples. Each dot is a single sample, with sample groups given on the x-axis and gene expression on the y-axis. The mean and standard error for each sample group is given as a red dot and line. In the plots, each gene is named after its product.

3.4 Discussion

The present study aimed to investigate the role of the *emm*5.23 genotype in the spread and pathogenic potential of iGAS in Scotland. This genotype was involved in a relative increase of invasive disease cases which peaked in 2018-2019. Although all GAS strains can cause clinical disease in humans (Terao, 2012), particular *emm* types are known to cause disease more frequently than others (Steer et al., 2009). The higher infectivity and virulence of some GAS strains has been ascribed to the acquisition of certain virulence genes via horizontal gene transfer, the occurrence of mutations in genes involved in transcriptional regulation and virulence (Nasser et al., 2014), and to multi-drug resistance (Davies et al., 2015), defined as resistance to at least one antibiotic in three or more drug classes (Magiorakos et al., 2012). GAS outbreaks, however, do not simply depend on factors which are intrinsic to the infecting pathogen strain and are influenced also by the environment and host susceptibility. GAS strains that are not considered hypervirulent can still cause outbreaks of invasive disease under certain circumstances, such as exposure of multiple at-risk individuals such as the elderly, hospitalised patients or intravenous drug users to a common source of infection (Ruben et al., 1984; Harkness et al., 1992; Lamagni et al., 2008c).

Most of the genomic analyses conducted in this study required the use of a good quality reference sequence to compare to the short-read Illumina WGS. The only publicly available closed genome of a *S. pyogenes emm*5 strain was the Manfredo sequence. This genome originated from a GAS strain isolated in the US in 1952 (Holden et al., 2007), and thus it was expected to be phylogenetically distant from our population and unsuitable to use as a reference. Due to the lack of other publicly available genomes phylogenetically related to the Scottish *emm*5.23 population, we decided to produce at least one closed and annotated *emm*5.23 genome to use as a reference for further analyses. The hybrid assembly approach we used to produce three closed genomes maximises the benefits of both Illumina and Oxford Nanopore technologies. If short-read Illumina sequences tend to be less prone to sequencing errors than long-read Oxford Nanopore sequences, the latter allow the complete coverage of large genomic portions and the production of a closed genome (Wick et al., 2017). Considering the importance of horizontal gene transfer in the acquisition of virulence characteristics and AMR in GAS (Nasser et al., 2014; Davies et al., 2015; Beres et al., 2016), we first estab-

lished the presence of MGEs in the *emm*5.23 WGS. We found that all the *emm*5.23 genomes sequenced shared the same MGEs, which consisted of four prophages and one composite element. The part of the composite element that was not of prophage origin was almost identical (99% identity) to an MGE identified in a strain of *F. alocis*, a pathogen of the oral cavity (accession CP002390) (Aruni et al., 2015). This MGE appeared to have originated from the recombination of two different ICEs. The first component was described in 2016 in an isolate of *S. dysgalactiae* (Wang et al., 2016) where it was not associated with increased bacterial virulence (Wang et al., 2016). The second component had a high similarity (89% identity) with an ICE described in a GAS *emm*12 clone involved in a scarlet fever outbreak in mainland China and Hong Kong in 2011 (Hsieh and Huang, 2011; Chen et al., 2012). Although a genomic study on the epidemic *emm*12 population attributed the isolates' virulent phenotype to phage-encoded proteins, the authors suggested that the AMR properties conferred by that ICE have had a major role in the selection and expansion of the population (Davies et al., 2015). Importantly, the portion of the *emm*12 ICE shared by the *F. alocis* strain and by the *emm*5.23 isolates carried the tetracycline resistance gene, *tetM*. No other AMR genes, however, were detected in the *emm*5.23 MGE. Four of the five prophage regions identified carried at least one virulence gene. The presence of these virulence genes, namely *sdn*, *spd1*, *spd3*, *spd4*, *hylP* and *speC*, has not, to the best of our knowledge, been linked to increased pathogenicity of GAS strains. The *speC* gene, carried by the prophage region inserted in the composite element, has been associated with scarlet fever (Silva-Costa et al., 2014; Davies et al., 2015). None of the virulence and AMR genes detected in MGE regions has been linked to increased pathogenicity of GAS strains in previous studies, suggesting that MGEs are unlikely to have played a role in the upsurge of *emm*5.23 cases of invasive disease in Scotland.

Next, we investigated the presence of virulence genes across the whole genome of the *emm*5.23 isolates. In previous studies, certain virulence genes have been associated with the successful spread of virulent strains (Nasser et al., 2014; Beres et al., 2016; Feng et al., 2016). In the current work we identified 44 virulence genes, responsible for the synthesis of 31 virulence factors. Eighteen of the virulence genes detected in the Scottish *emm*5.23 population are considered part of the GAS core genome, which was defined as the collection

of GAS genes that are shared by >99% of the 2083 genomes included in a recently published study (Davies et al., 2019). The *slo* gene, encoding the secreted toxin SLO, plays an important role in hyper-virulent *emm* types such as *emm*1, 12 and 89 (Nasser et al., 2014; Beres et al., 2016; Feng et al., 2016), particularly when coupled with the *nga* gene encoding the NAD-glycohydrolase exotoxin, whose presence in the *emm*5.23 genome was confirmed by BLAST searching. In the *emm*1, 12 and 89 epidemic strains, however, SLO and NAD-glycohydrolase production alone is insufficient to account for their pathogenic potential, which is also influenced by mutations in transcriptional regulation genes (Nasser et al., 2014; Beres et al., 2016; Feng et al., 2016). When compared to the virulence gene set of the historic *emm*5 strain Manfredo, the Scottish *emm*5.23 population differed only in the acquisition of the streptodornase encoding gene *sdn*. Overall, previously described virulence gene profiles associated with hypervirulence in epidemic GAS strains were not detected in the *emm*5.23 population (Beres et al., 2002; Nasser et al., 2014; Beres et al., 2016; Kachroo et al., 2019). Based on our findings, there is no evidence to suggest that the virulence gene profile of the Scottish *emm*5.23 population is responsible for an enhanced pathogenicity.

The acquisition of multi-drug resistance is thought to play a role in the successful spread of GAS strains (Davies et al., 2015). Moreover, it is important to know the AMR profile of recently emerged strains for surveillance purposes and to inform about clinical management of GAS infections. Therefore, we decided to characterise the AMR genotype and phenotype of the Scottish *emm*5.23 isolates. The potential for genotypic-based resistance to different antibiotic classes was assessed by analysing the sequences for the presence of AMR-conferring genes and mutations. Only two genes associated with AMR were found in every isolate, namely *lmrP* and *tetM*. This AMR gene profile has been associated with possible resistance to MLSB and to tetracyclines resistance (Morse et al., 1986; Putman et al., 2001). Given the recently reported emergence of *S. pyogenes* strains with reduced sensitivity to penicillins due to single point mutations in genes encoding penicillin-binding proteins (Vannice et al., 2019; Musser et al., 2020; Beres et al., 2022), we also assessed the presence of mutations involving *pbp* genes. When we looked for mutations known to be associated with reduced sensitivity to β-lactams in the *pbp* genes (Vannice et al., 2019; Musser et al., 2020), we did not find any. We tested 25 of the Scottish

*emm*5.23 isolates for sensitivity to a panel of antibiotics commonly used to treat Gram positive infections. We established MIC values with the BMD method using the *H. influenzae* strain ATCC 49766 as a control. Although the EUCAST guidelines recommend the use of *S. pneumoniae* strain ATCC 49619 for principal quality control of GAS AST (`https://www.eucast.org/ast_of_bacteria/quality_control/`), we were unable to obtain the expected ranges of antimicrobial susceptibility for this strain. Since the protocol of the GAS AST assay is very similar to the one for the *H. influenzae* strain ATCC 49766, for which we were able to obtain the expected ranges of antimicrobial susceptibility, we used the ATCC 49766 strain to quality control the kit and reagents, which were employed as per the manufacturer's instructions. We found that all isolates tested were resistant to tetracycline, as expected from their genotype. Only two isolates were resistant to doxycycline. Based on our findings, we suggest that the AMR genotype detected in the *emm*5.23 population confers full resistance to tetracycline and limited resistance to doxycycline. This finding is in concordance with previously generated data on *S. pneumoniae* harboring the *tetM* gene (Dallas et al., 2013). It should be noted, however, that in the present study doxycycline MICs were very close to breakpoint values for all isolates classified as susceptible (Table 3.1). Interestingly, 14 of the 18 isolates tested with Vitek 2 technology in the diagnostic laboratories of the Greater Glasgow and Clyde area were classified as resistant to doxycycline. Since the isolates tested by diagnostic laboratories were the same tested with BMD in the present study, this discrepancy is unlikely to be ascribed to biological reasons and we hypothesise that the different phenotypic results may be linked to the methodology applied. In fact, Vitek 2 has shown a tendency to overestimate MIC values for tetracyclines in *S. pneumoniae* strains (Goessens et al., 2000). Unfortunately, we do not have MIC values generated via Vitek 2 testing, so a direct comparison with the data produced using BMD is not possible. It should be noted that none of the 25 isolates tested with BMD, nor any of the 18 isolates tested with Vitek 2 technology, expressed phenotypic resistance to lincosamides and macrolides, as expected from genotypic data. We concluded that the presence of the *lmrP* gene on its own, which was associated with lincosamide, macrolide and streptogramin resistance in previous studies (Putman et al., 2001), is not responsible for lincosamide and macrolide resistance *in vitro* in the isolates tested. This finding is in agreement with the observation that genotypic data cannot always correctly infer AMR phenotype (WHO, 2020), highlighting the impor-

tance of *in vitro* testing. Based on the AMR characteristics of the *emm*5.23 isolates analysed, we concluded there is no sign of multi-drug resistance in this bacterial population.

The potential for acquisition of non-synonymous mutations that may influence bacterial phenotype was taken into account as a possible mechanism to favour virulence and strain transmission. Relative to the reference genomes used, we did not detect any non-synonymous polymorphisms shared by all *emm*5.23 isolates. We did observe, however, that the majority of the *emm*5.23 isolates (46/54) carried three identical non-synonymous polymorphisms. The first one was an amino acid change in the *lctO* gene, which is involved in the L-lactate metabolism (Taniai et al., 2008). The second polymorphism was a SNP that determined an early stop codon in the gene expressing Acetyl-CoA C-Acyltransferase, which is implicated in the fatty acids metabolism (Röttig and Steinbüchel, 2013). Finally, we detected a nucleotide insertion upstream (position -36) of the *greA* gene, which is a transcription elongation factor (Yuzenkova et al., 2014). These mutations do not involve genes known to be directly associated with bacterial virulence, hence their role in pathogenesis remains uncertain. Since the *lctO* gene and the gene expressing Acetyl-CoA C-Acyltransferase are both involved in metabolic pathways, we cannot exclude that the polymorphisms identified confer some biological fitness advantage. However, based on published data on the emergence of hypervirulent genotypes (Nasser et al., 2014; Beres et al., 2016), it appears unlikely that the polymorphisms detected in this study have a significant influence on GAS virulence.

A maximum likelihood core SNP phylogenetic tree was constructed using all Scottish isolates and a randomly selected sample of English isolates, the latter included to increase the scope of the analysis (Figure 3.3). Our results suggested a phylogenetic distinction between the isolates of the two main groups. Most of the Scottish isolates in the "English group" were collected in 2018, when a generalised increase in the incidence of iGAS cases may have been associated with the high incidence of influenza in the 2017-2018 winter (`https://webarchive.nationalarchives.gov.uk/ukgwa/20220401215804/https://www.gov.uk/government/statistics/annual-flu-reports`). In the Scottish group, 30% of the Scottish isolates were from 2018 and the remaining 70% were from 2019, 2020 and 2022. The distribution of the previously described

polymorphisms, which were shared by all the isolates of the Scottish group, reflected the monophyletic nature of the majority of the Scottish isolates. This suggests that 46 isolates responsible for invasive disease in Scotland were phylogenetically closely related and shared a relatively recent common ancestor. The tree topology indicated that the most likely common ancestor of the Scottish group isolates is a cluster of six clones, which have no more than 7 SNPs of difference from the remaining isolates of the Scottish group.

The phenotypic impact of the three polymorphisms that characterised most of the Scottish isolates was investigated through total RNA sequencing and transcriptomic analysis. We compared the overall gene expression of four replicates of an isolate from the Scottish group with four replicates of an isolate from the English group. Overall, we did not find major differences in the transcriptomes of the two isolates. We noticed, however, among-replicate heterogeneity, which could partially explain the limited diversity detected. When we focused on those genes that were flagged as statistically significantly differentially expressed in the Scottish group isolate compared with the English group isolate, we found nine genes with higher expression and five with lower expression. These genes did not appear to share any functional relationship, which was confirmed by gene enrichment analysis. These results suggest that there is no significant difference in the transcriptome, and consequently the phenotype, of the two isolates tested. It should be noted, however, that gene expression in bacteria can be influenced by environmental conditions, as demonstrated by the activation of stress response pathways under challenging circumstances (Dalton and Scott, 2004). In this work, the chosen isolates were grown in the nutrient-rich broth THB and no other growth conditions were tested for due to financial constraints. It was thus impossible to investigate the occurrence of transcriptomic shifts driven by environmental conditions in the two *S. pyogenes emm*5.23 isolates. Since the growth medium utilised is not a good representation of the environmental conditions that *S. pyogenes* strains face when interacting with human hosts, this should be regarded as a limitation of the transcriptomic work described. If I had the chance to carry out more experimental work to better characterise the gene expression in *S. pyogenes emm*5.23 strains, I would analyse the transcriptome of more isolates (e.g. six isolates representative of the English group and six representative of the Scottish group genotypes) grown in multiple media (e.g. C medium, which is rich in peptide but poor in glu-

cose, or chemically defined media that are deficient in metals). Although the transcriptomic results obtained in this work may be influenced by the small sample size and by the specific environmental conditions, all the data collected thus far indicated that a relevant phenotypic difference between isolates from the English group and isolates from the Scottish group is unlikely. Genomic similarity and phylogenetic tree characteristics strongly suggest that 46 of the total 58 cases of *emm*5.23 invasive disease registered in Scotland in recent years can be considered linked by an unknown means of transmission or common factor, and so represent a potential outbreak of infection (Engelthaler et al., 2016; Pightling et al., 2018).

The genomic investigation carried out in this work revealed that the GAS *emm*5.23 genotype responsible for a relative increase in cases of invasive disease in Scotland peaking in 2018-2019 does not appear to contain a genotypic background analogous to currently recognised hypervirulent strains. The expansion of a clade of isolates sharing three identical non-synonymous polymorphisms, (*lctO*, *greA* and an early stop codon in the gene expressing Acetyl-CoA C-Acyltransferase) is interesting and further studies investigating their pathogenic phenotype are required. Although phylogenetic analysis indicates linkage between isolates, due to the limited metadata available we were not in a position to establish epidemiological connections among patients. A combination of environmental and host-specific factors may be responsible for a chain of events that led to the relative increase in this *emm* subtype in Scotland in the studied period. Our results highlight the need to implement coordinated detection systems to monitor and alert health protection teams to take effective action in investigation and control of linked iGAS increases in Scotland.

Chapter 4

Genomic epidemiology, virulence and antimicrobial resistance of the multi-host pathogen *Streptococcus canis*

4.1 Introduction

Streptococcus canis is a Gram positive, ß-haemolytic, Lancefield Group G *Streptococcus* (Devriese et al., 1986). While normally found on the skin and mucosal membranes of dogs and cats not showing clinical signs of infection (Lysková et al., 2007b), *S. canis* can occasionally be involved in canine and feline skin and soft tissue infections (Lysková et al., 2007b) or, in rare occasions, invasive and severe forms of disease (Prescott et al., 1995; Matsuu et al., 2007; Pesavento et al., 2007). Although dogs and cats appear to be the main reservoirs of this bacterial species, *S. canis* has been isolated from and implicated in disease in several mammalian host species, including cattle (Król et al., 2015; Eibl et al., 2021) in which it is associated with sub-clinical mastitis, and humans (Galpérine et al., 2007). In humans, *S. canis* infections are uncommon but occasionally result in severe clinical manifestations, such as septicaemia (Taniyama et al., 2017; Zaidi and Eranki, 2019; Lederman et al., 2020) and endocarditis (Lacave et al., 2016; Mališová et al., 2019).

Over the past two decades, reports of life-threatening cases of *S. canis*-associated disease in humans (Lacave et al., 2016; Taniyama et al., 2017; Mališová et al., 2019; Zaidi and Eranki, 2019), together with the ever increasing popularity of dogs and cats as family pets, have drawn attention towards this bacterial species. A number of studies have investigated *S. canis*, but compared to other streptococcal species, work on this pathogen is limited and much remains unclear. For instance, while epidemiological studies have provided information regarding the carriage and incidence of infection in pets and humans, to date they have been limited in scope and number (Galpérine et al., 2007; Lysková et al., 2007b,a; Lamm et al., 2010; Guerrero et al., 2018). The emergence of AMR in *S. canis* isolates has also been documented (Galpérine et al., 2007; Lysková et al., 2007a; Pinho et al., 2013; Fukushima et al., 2020b), but its burden is unclear and the underlying biological determinants are not fully known. Several virulence mechanisms have been investigated (DeWinter and Prescott, 1999; Fulde et al., 2013; Hitzmann et al., 2013; Yoshida et al., 2021) but, similarly, little is known about the carriage of virulence genes in the global *S. canis* population. Two main systems for genotypic strain classification, an MLST scheme (Pinho et al., 2013) and one based on the allelic variations of the *scm* gene (Pinho et al., 2019; Fukushima et al., 2020a), have been developed. However, neither of these schemes have been tested against high-discrimination typing techniques such as core genome SNP typing. Finally, a lack of host-specificity of *S. canis* has been previously proposed, based on the identification of the same MLST isolates in different host species (Pinho et al., 2019). However, traditional MLST is based on the allelic variations of only seven housekeeping genes across the entire genome, making it a relatively low-discrimination approach; caution should be exercised in making inferences about bacterial population structure and evolution on the basis of such sparse genetic markers (Tsang et al., 2017).

The use of high-throughput sequencing data, such as WGS, to study bacterial populations has been increasingly employed in recent years and is now considered a key element in our understanding of pathogen epidemiology and evolution (Schürch et al., 2018). Many of the knowledge gaps surrounding *S. canis*, some of which were highlighted in the previous paragraph, may be addressed by WGS population analysis, which has never hitherto been applied to the study of this bacterial species.

Aim and objectives The aim of this chapter was to perform for the first time a WGS-based analysis of a collection of *S. canis* strains (n=59) from different hosts and geographic locations in order to gain valuable insights into the epidemiology of this multi-host pathogen. The following objectives were defined to address the aim of the chapter:

- To assess the prevalence and molecular mechanisms of AMR in *S. canis*.
- To determine the presence and distribution of virulence genes in this bacterial species.
- To evaluate the accuracy of MLST and SCM classification systems against core genome SNP typing.
- To look for signs of host adaptation among *S. canis* isolates.

4.2 Methods

4.2.1 Isolate collection, whole genome sequencing and sequence assembly

Thirty-nine isolates of *S. canis* from dogs, cats, humans and a seal collected in the UK between 2002 to 2021 were available for this study. Twenty-eight animal derived isolates were provided by the University of Glasgow Veterinary Diagnostic Services (VDS, Glasgow, UK), one human isolate by the Glasgow Royal Infirmary (GRI, Glasgow, UK) and ten human isolates by Public Health England (PHE, Colindale, UK). All frozen glycerol-stored isolates were grown overnight on THB agar (Thermo Scientific, Loughborough, UK) at 37°C and submitted to SMiRL for WGS. The DNeasy 96 Blood and Tissue Kit (Qiagen, Hilden, Germany) was used to extract genomic DNA according to the manufacturer's instructions. DNA was purified with the QIAsymphony extraction instrument (Qiagen, Hilden, Germany) and quantified using Qubit dsDNA BR Assay Kit on a Qubit 3 Fluorometer. Paired-end sequencing libraries were prepared with a Nextera XT DNA Library Preparation Kit and Index Kit V2 (Illumina, Cambridge, UK) and paired-end sequencing carried out using Illumina MiSeq technology. Raw sequencing reads were trimmed using ConDeTri (Smeds and Künstner, 2011) and contigs assembled using SPAdes v 3.14.0 (Bankevich et al., 2012). The quality of the assemblies was assessed using QUAST v 5.0.2 (Gurevich et al., 2013).

4.2.2 Genome dataset

All the publicly available genomes of *S. canis* (n=20) from the NCBI genomes database as of May 2022 and the newly generated genomic sequences were analysed in the current study. The publicly available sequences derived from *S. canis* isolates originated from different geographic locations, namely South Korea, Japan, Europe and the US. Overall, the dataset studied was composed of WGS from isolates derived from two cattle, 9 cats, 32 dogs, 14 humans, one unspecified animal and one common seal. Additional information, including accession numbers, for each WGS included in this work is provided in Table C.1.

4.2.3 Antimicrobial resistance

The presence of genes associated with AMR was determined using ARIBA v 3.1.0 (Hunt et al., 2017) and the CARD database (Jia et al., 2016), as of August 2021. Commands used to run ARIBA are presented in Appendix B.2.4. Only short-sequence reads can be analysed by ARIBA, hence, four published genomes were excluded from this step. Three of the excluded genomes were generated using long-read sequencing technologies and one did not have publicly accessible sequencing reads. ARIBA was also used to extract sequences of *gyrA*, *gyrB*, *parC* and *parE* genes from each genome, using as a reference the corresponding genes from the publicly available strain HL_98_2 (accession number CP053789.1). Sequence alignments of *gyrA*, *gyrB*, *parC* and *parE* were then produced using MAFFT v 7 (Katoh et al., 2009) and visually inspected with MEGAX v 10.1.7 (Kumar et al., 2018) to assess the presence of mutations associated with quinolone resistance in *S. canis* (Fukushima et al., 2020b).

For the 39 available isolates of *S. canis*, MIC to a panel of antibiotics commonly used to treat Gram-positive infections, namely ampicillin, amoxicillin, clindamycin, ceftriaxone, cefotaxime, doxycycline, erythromycin, levofloxacin, meropenem, moxifloxacin, oxacillin, penicillin G, tetracycline and vancomycin, were determined using BMD as per EUCAST guidelines `https://www.eucast.org/ast_of_bacteria/mic_determination/`. Bacterial cultures were plated and grown on Columbia blood agar plates (Oxoid, Basinstoke, UK) for 48 hours at 37°C. For each isolate, saline solutions

(0.85% NaCl, pH 5.5 to 6.5) were inoculated with bacterial cells to reach a density of 0.5 McFarland (0.44-0.56). Bacterial suspensions were then added to a solution of Micronaut H-Medium broth (BioConnections, Knypersley, UK). Equal volumes of H-Medium broth were then distributed in a 96 well Micronaut-S PHE Co-GP03 plate. Plates were incubated for 22-24 h at 37°C. After that, bacterial growth was measured by a Multiskan FC Microplate photometer (Thermo Scientific, Loughborough, UK) and MIC values were interpreted by the Micronaut MCN6 software according to the EUCAST breakpoint values v 12.0 (`https://www.eucast.org/clinical_breakpoints/`).

4.2.4 Virulence genes

The presence of genes homologous to known bacterial virulence factors was assessed in this collection of genomes using the command line version of BLASTn v 2.9.0 (Camacho et al., 2009), coupled with VFDB (Chen et al., 2005), as of October 2021. A positive match was considered to be one with at least 20% sequence identity, at least 90% gene coverage, a bit score > 50 and an e-value < 10^{-10} (Pearson, 2013). This parameter choice allowed for a conservative approach that maximised the specificity of the search. The script used to run BLASTn with VFDB is reported in Appendix C.2.1.

4.2.5 Strain typing

Multi-locus sequence types were determined for all the 59 genomes using the PubMLST database (Jolley and Maiden, 2010) and the software mlst (`https://github.com/tseemann/mlst`), as shown in Appendix C.2.2. SCM sequences were extracted from all genomes using ARIBA and assigned an SCM type according to the typing scheme developed by Fukushima *et al.* (Fukushima et al., 2020a).

4.2.6 Phylogenetic analysis

A core genome SNP alignment of all the 59 WGS in the current database was generated with snippy v 4.4.5 (Seemann, 2015). A maximum likelihood phylogenetic analysis with 100 standard nonparametric bootstrap replicates was then carried out using IQ-TREE v 2.1.4 (Nguyen et al., 2015) to produce a core SNP phylogenetic tree. The tree was visualised

and annotated using RStudio v 2022.7.2.576 (RStudio Team, 2020) and the packages ggtree (Yu et al., 2017) and phytools (Revell, 2012). An outgroup of four *Streptococcus dysgalactiae equisimilis* genomes (accession numbers: ASM1419289v1, 44503_D02, 42197_A02, 46166_D01) was used to root the tree. Core genome SNP (CGS) types were identified in the core SNP phylogeny using TreeCluster (Balaban et al., 2019) with a threshold of 0.017 and the Max method (Appendix C.2.2). These settings allowed for the detection of clusters with a maximum of 2,000 pairwise core SNP difference between isolates. IQ-TREE was then used to generate two additional core SNP phylogenies, one constrained to be monophyletic for the MLST types and one constrained to be monophyletic for the SCM types.

4.2.7 Comparative phylogenetic analysis

Both constrained phylogenies were compared to the unconstrained core SNP tree using an Approximately Unbiased (AU) test (Shimodaira, 2002) with 10,000 nonparametric bootstrap replicates on IQ-TREE (Appendix C.2.2). The AU test evaluates different tree topologies under the null hypothesis that the trees tested provide an equally good explanation of the dataset in use. The accuracy of strain clustering according to the MLST and SCM schemes was then determined in comparison to the newly identified CGS types using the adjusted Wallace (AW) coefficient (Severiano et al., 2011) with 95% CI on the Comparing Partitions website (`http://www.comparingpartitions.info/?link=Home`).

4.2.8 Pangenome-wide association analysis

Genomes were annotated using Prokka v 1.14.6 (Seemann, 2014). An *S. canis* pangenome was then generated with Panaroo (Tonkin-Hill et al., 2020). Scoary v 1.6.16 (Brynildsrud et al., 2016) was later used to carry out a pangenome-wide association (pan-GWAS) analysis to investigate the potential overrepresentation of specific genetic markers in strains from different host species. For the pan-GWAS analysis, p-values were adjusted using the Benjamini-Hochberg method (Ferreira and Zwinderman, 2006), to correct for false positives while undertaking multiple statistical testing. All commands used for pangenome analysis are provided in Appendix C.2.3.

4.2.9 Accessory genome network

The accessory gene diversity of the population studied was determined with GraPPLE (`https://github.com/JDHarlingLee/GraPPLE`), which calculates the pairwise similarity between genomes, expressed as a proportion of shared accessory genes (Appendix C.2.4). Accessory genes were considered those shared by $\leq$99% of the analysed genomes, based on the pangenome analysis results. The pairwise distance matrix produced by GraPPLE was visualised on Graphia v 2.2 (Freeman et al., 2020) in the form of a network. For a proper network visualisation, edges were reduced using the k-nearest neighbour (k-NN) algorithm, calculated with edge weight, k=5 and descending order.

4.3 Results

All supplementary material for this chapter can be found in Appendix C (these are indicated with the letter C in front of the sequential number).

4.3.1 Antimicrobial resistance

I initially focussed on the detection and identification of AMR-conferring genes in the genomic sequences. A total of six different genes associated with AMR was found among the genomes analysed (Table 4.1). These genes are *ermA* and *ermB*, associated with MLSB resistance (Yu et al., 1997), *lsaC*, associated with lincosamide resistance (Malbruny et al., 2011), and *tetM*, *tetO* and *tetS*, all associated with tetracycline resistance (Roberts, 2005). Seventeen of the 55 genomes tested (31 %) were positive for the presence of at least one AMR gene. Only three genomes were positive for more than one AMR-conferring gene, all having both the *ermB* and *tetO* genes. The most common AMR gene in the dataset was *tetO* (8/55) and tetracycline resistance appeared to be the most prevalent, based on the genotype (14/55). If the results from publicly available genomes are excluded, which may have specifically been sequenced because of their AMR characteristics, the prevalence of at least one AMR-conferring gene in our cohort of *S. canis* isolates is 23 % (9/39). None of the previously reported single point mutations in the QRDR regions of *gyrA*, *gyrB*, *parC* and *parE* were detected in this cohort of genomes.

Table 4.1: Genomic determinants of antimicrobial resistance detected among the 55 *Streptococcus canis* short-read sequenced genomes.

AMR determinants	Newly generated sequences	Publicly available sequences	Total
ermA	1/39 (3 %)	0/16 (0 %)	1/55 (2 %)
ermB	2/39 (5 %)	2/16 (13 %)	4/55 (7 %)
lsaC	1/39 (3 %)	0/16 (0 %)	1/55 (2 %)
tetM	1/39 (3 %)	3/16 (19 %)	4/55 (7 %)
tetO	6/39 (15 %)	2/16 (13 %)	8/55 (15 %)
tetS	0/39 (0 %)	2/16 (13 %)	2/55 (4 %)
ermB + *tetO*	2/39 (5 %)	1/16 (6 %)	3/55 (5 %)
Quinolone-conferring mutations	0/39 (0 %)	0/16 (0 %)	0/55 (0 %)
Any AMR determinant	9/39 (23 %)	8/16 (50 %)	17/55 (31 %)

Thirty-nine *S. canis* isolates were tested for antimicrobial sensitivity towards 14 commonly used antibiotics, representative of six antibiotic classes, namely ß-lactams, fluoroquinolones, glycopeptides, lincosamides, macrolides and tetracyclines. MIC values for the antibiotics tested are reported in Table C.2. Based on the values reported in the EUCAST breakpoint table v 12.0 for Group G streptococci, resistance was detected against lincosamides (clindamycin), macrolides (erythromycin) and tetracyclines (tetracycline and doxycycline). All isolates tested were fully sensitive to ß-lactams. The concordance between AST and genomic inference was very high (95 % agreement) and the carriage of AMR-associated genes among the 39 isolates tested is illustrated in Figure 4.1. Nine of 39 isolates tested (23 %) were resistant to at least one antibiotic and for eight of these a genomic determinant of AMR could be identified. One isolate was found to be resistant to tetracycline and no resistance-conferring gene was detected within its genome. Conversely, one isolate

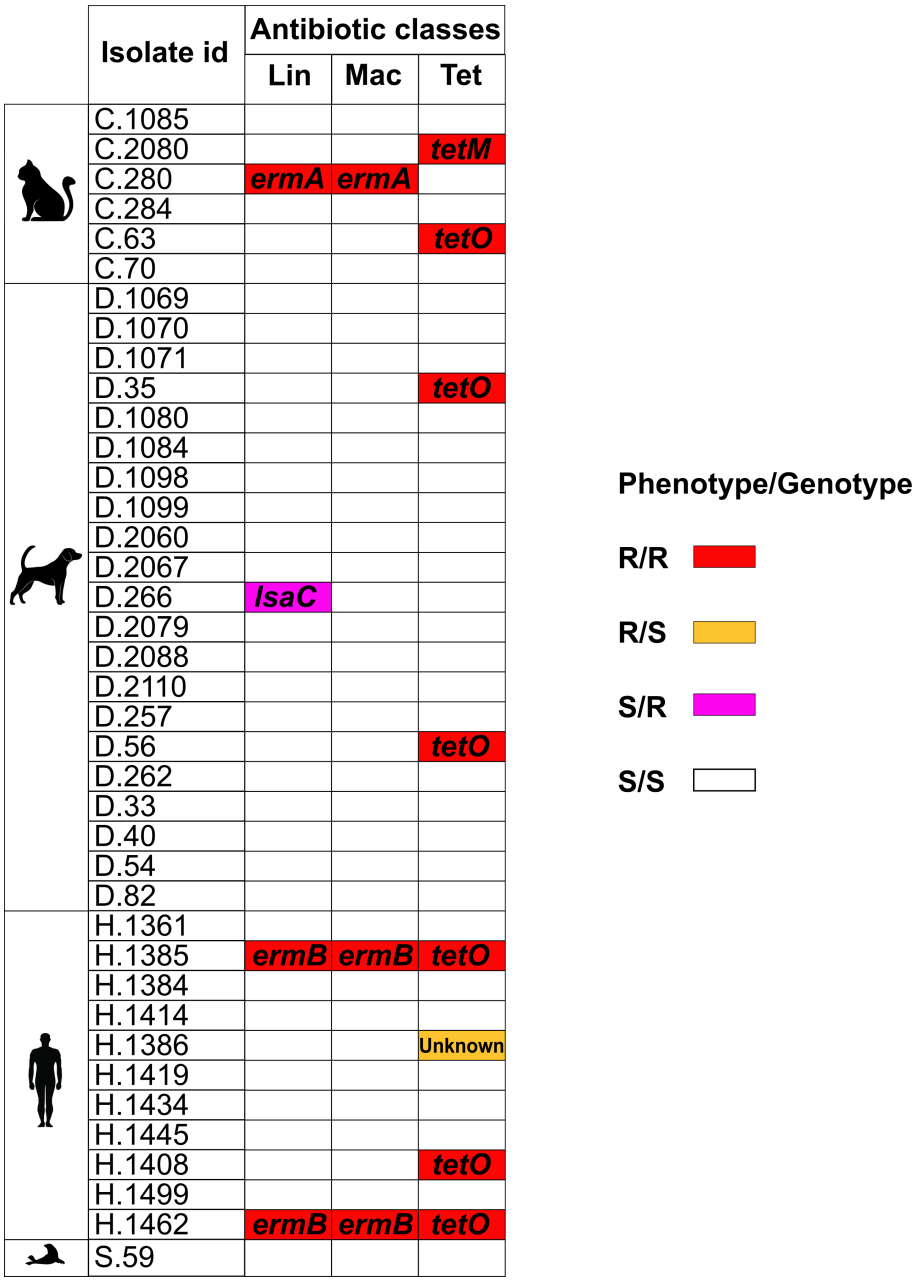


Figure 4.1: Concordance between antimicrobial susceptibility testing results (phenotype) and presence of antimicrobial resistance (AMR)-associated genes (genotypes) in the 39 *Streptococcus canis* isolates tested. Each row represents one isolate and isolates are grouped based on the host (cat, dog, human and seal) from which they were collected. Since phenotypic resistance was found only towards lincosamides (clindamycin), macrolides (erythromycin) and tetracyclines (doxycycline and tetracycline), only results referring to these antibiotic classes are reported. Genomic determinants of AMR are reported for isolates showing phenotypic and/or genotypic resistance profiles. Lin = lincosamides; mac = macrolides; tet = tetracyclines; R = resistance; S = sensitivity.

that was predicted to be resistant to lincosamides was fully sensitive to all antibiotics tested.

4.3.2 Virulence genes

A total of 19 genes homologous to known virulence genes within the VFDB was found among the 59 genomes analysed (Figure 4.2). Fourteen of these genes (74%) were detected in every genome, suggesting they could be part of the species core genome. One gene, homologous to the hyaluronidase-encoding gene *hyl*, was detected in all but one genome. Forty-two WGS were positive for the carriage of a homolog to the *ssp-5* gene, which encodes an agglutinin receptor. Seven genomes carried homologs to one or more of the following: *aspA*, *mf3*, *sda* and *smeZ*. The *smeZ* gene, whose product is the streptococcal mitogenic exotoxin Z, is found in some *S. pyogenes* strains and when present it is usually integrated into the chromosome. All the virulence genes found in the current WGS dataset have been described in other pathogenic streptococcal species, such as *S. pyogenes*, *S. agalactiae* and *S. pneumoniae* (Liu et al., 2019). The *scm* gene, encoding the universally present virulence factor *S. canis* M-like protein (Pinho et al., 2019), was not detected using the previously described methodology. The *scm* gene is, to date, the only confirmed *S. canis*-specific virulence gene and it is not currently found in the VFDB. The presence of such gene, however, was confirmed in all isolates using as a reference publicly available *scm* sequences and command line BLAST.

4.3.3 Population analysis

A core SNP maximum-likelihood phylogeny was built with the aim of providing a high-resolution representation of the evolutionary relationships among isolates (Figure 4.3). Since the two existing typing schemes proposed for *S. canis*, the MLST and SCM systems, have never been validated against a highly accurate typing technique such as core genome SNP typing, both MLST- and SCM-constrained core SNP maximum-likelihood phylogenies were also constructed and compared to the unconstrained core SNP tree.

The results of the AU statistical test used to compare the phylogenies indicate that both constrained trees are significantly different from the unconstrained one (Table 4.2). It can

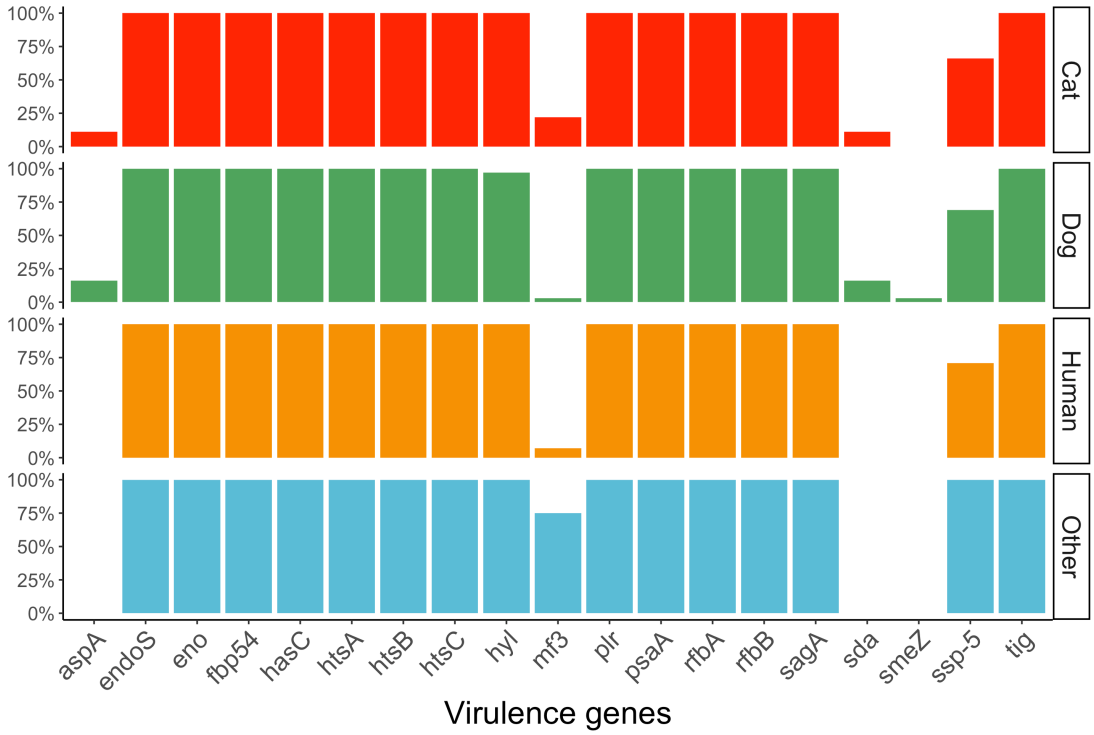


Figure 4.2: Proportion of genomes, grouped by host species, carrying genes homologous to known virulence genes within the VFDB. The group "Other" includes two genomes from bovine isolates, one genome from an unspecified animal isolate and one genome from a seal isolate.

therefore be concluded that the clusters predicted by the MLST and SCM schemes are not a perfect fit to the core SNP data.

Table 4.2: Approximately unbiased statistical comparison of multi-locus sequence typing (MLST) and *Streptococcus canis* M-like protein (SCM)-constrained phylogenies using the core single nucleotide polymorphism unconstrained tree as a reference. Log likelihood, difference in log likelihood and p-values are reported for each tree comparison.

Tree tested	ln L	Diff ln L	p-value
Core SNP MLST-constrained	-501089.9	2631	6.68×10^{-7}
Core SNP SCM-constrained	-508213.29	9754.4	1.48×10^{-40}

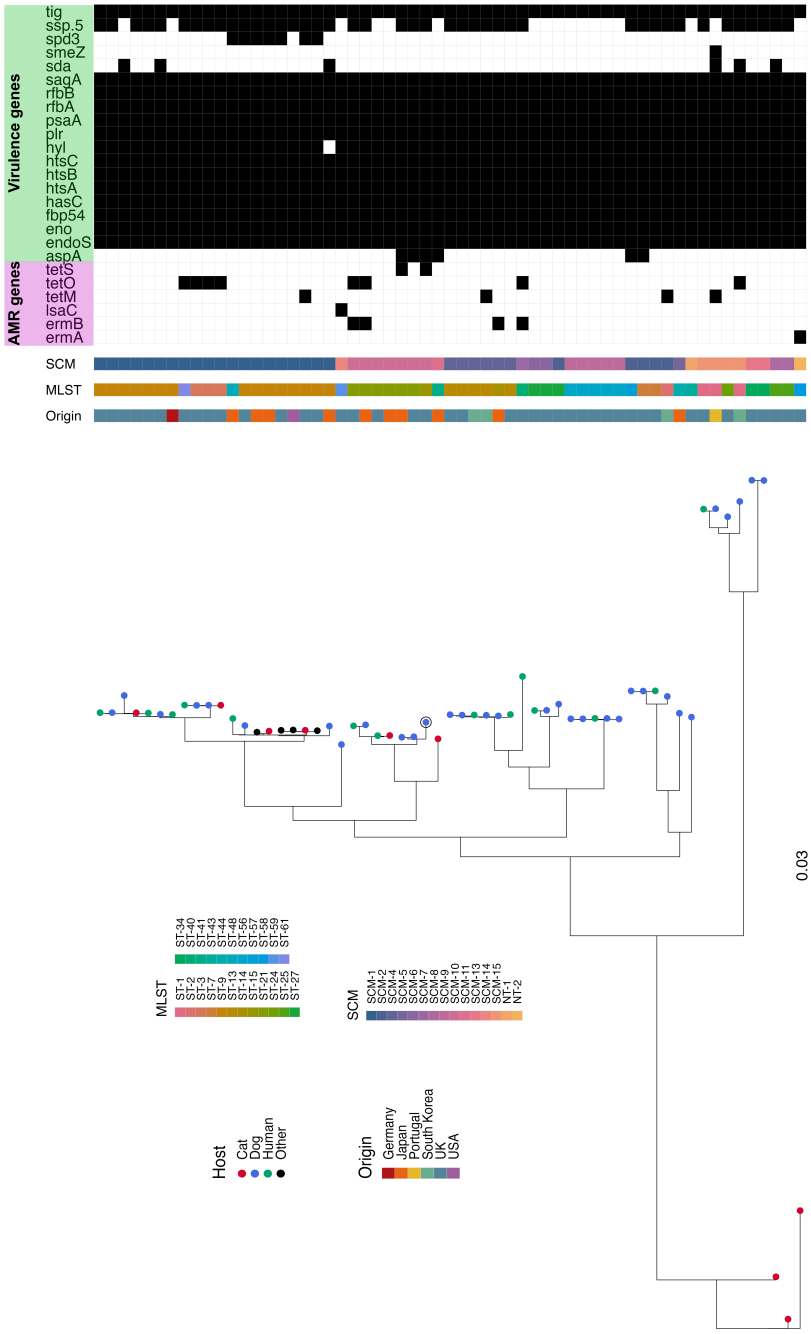


Figure 4.3: Maximum likelihood core single nucleotide polymorphism phylogenetic tree of the 59 *Streptococcus canis* whole genome sequences analysed. The tree was rooted using an outgroup of four *Streptococcus dysgalactiae equisimilis* genomes. Isolates are coloured according to the host species from which they were isolated. The reference genome used for phylogenetic analysis, namely D.HL_98_2 (accession number: ASM1099386v2), is highlighted on the tree by a black outer circle. Tree annotations regarding the isolates' geographical origin and classification according to multi-locus sequence types (MLST) and *S.canis* M-like protein (SCM) Fukushima types are reported. NT stands for new types and indicates those *scm* alleles that were not classified in the work of Fukushima et al. A heatmap shows which genomes possess (black squares) or which do not possess (white squares) antimicrobial resistance and virulence genes detected in previous analyses.

Eighteen CGS types were found in the core SNP phylogeny by TreeCluster (Appendix C.1). The AW coefficient of the MLST and SCM schemes compared to the CGS clustering was, respectively, 0.575 (CI 0.495-0.654) and 0.540 (CI 0.467-0.613). The AW coefficient can be interpreted as the probability of two isolates belonging to the same group according to one scheme (MLST or SCM) to also belong to the same group according to another scheme (CGS). Since the CIs of the two AW coefficients obtained are overlapping, we can conclude that there is no evidence that either of the competing schemes to type *S. canis* strains is more accurate than the other in representing core SNP diversity.

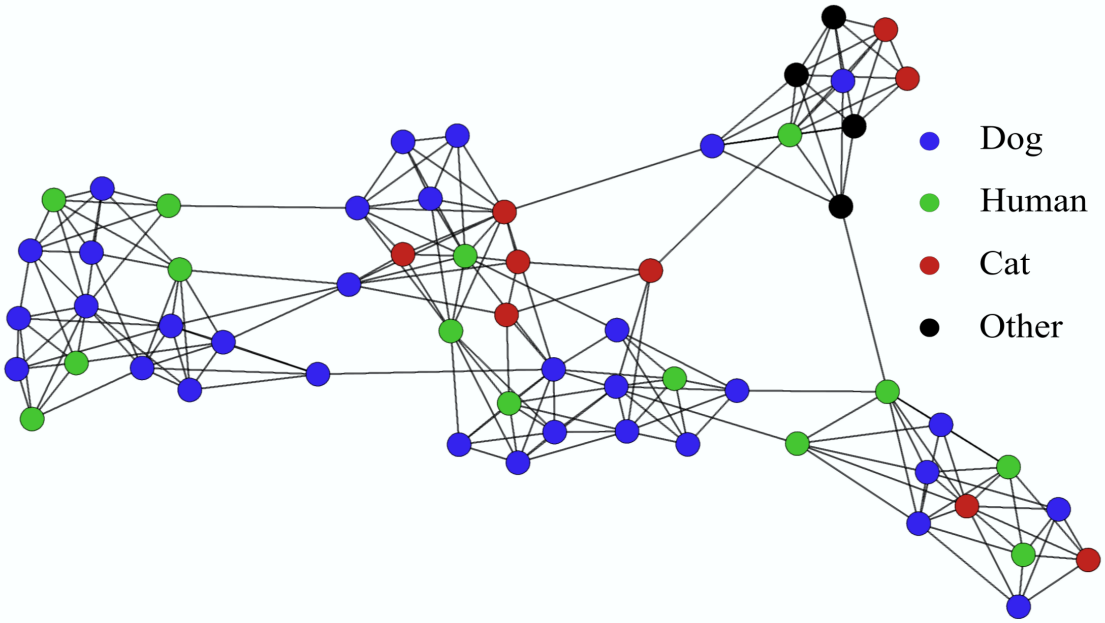


Figure 4.4: Accessory genome network for the *Streptococcus canis* strains investigated. Each dot represents the accessory genome of a bacterial isolate. Dots connected and/or clustering have a similar accessory genome. Dots are coloured according to the isolate host species.

Streptococcus canis core SNP phylogeny was also used to investigate the relatedness of isolates from different host species. The core SNP tree indicated that isolates from different hosts frequently cluster together, suggesting a lack of host adaptation in the *S. canis* cohort analysed (Figure 4.3). In one case, the core SNP distance between a human and a dog isolate was found to be as low as 25 SNPs (Appendix C.2). Diversity within the *S. canis* accessory genome was also evaluated. A network based on the presence/absence of accessory genes

was created and is depicted in Figure 4.4. In agreement with the core SNP phylogenetic tree, the accessory gene network did not show any signs of host-specificity in the current *S. canis* dataset.

The newly generated *S. canis* pangenome comprised 1432 core genes, defined as those shared by at least 99% of the genomes. The majority of the remaining 2,994 genes that constitute the *S. canis* accessory genome are shared by ten or fewer strains (Appendix C.3). The pan-GWAS analysis revealed no specific genomic traits overrepresented in different host groups across the 59 genomes analysed.

4.4 Discussion

Streptococcus canis has been recognised for decades as a pathogen of multiple mammalian species, but many aspects of its biology and disease epidemiology remain unclear. In this work, for the first time, a genomic-based approach was used to study *S. canis*, with the aim of clarifying some important epidemiological characteristics such as AMR prevalence, virulence gene distribution and overall population structure. This study represents the largest *S. canis* genome collection (n=59) currently available.

A set of animal (n = 28) and human (n = 11) derived newly generated *S. canis* WGS was analysed together with all the publicly available *S. canis* genomes (n = 20). Given the limited knowledge about AMR prevalence and its genetic determinants in *S. canis*, the available collection of genomes was scanned for the presence of AMR-encoding genes and mutations previously associated with quinolone resistance (Fukushima et al., 2020b). Six resistance encoding genes, *ermA*, *ermB*, *lsaC*, *tetM*, *tetO* and *tetS*, were detected across the genomic dataset. All these genes, except for *lsaC*, have previously been detected in *S. canis* isolates resistant to macrolides, lincosamides and tetracyclines (Lysková et al., 2007a). When expressed, gene *lsaC* confers high levels of lincosamide resistance through an efflux mechanism as has previously been described in *S. agalactiae* strains (Malbruny et al., 2011). In our study, 17 genomes carried at least one AMR gene and only three genomes carried two. None of the previously described mutations associated with quinolone resistance in *S. canis* isolates was detected in this collection of genomes. In order to assess the accuracy of the

genomic-based AMR predictions, the MIC values for a panel of antibiotics commonly used to treat streptococcal infections were determined for the 39 available isolates. Resistance was detected towards lincosamides, macrolides and tetracyclines. Resistance to these antibiotic classes appears to be commonly encountered in *S. canis* strains isolated from dogs, cats and humans, according to the literature (Galpérine et al., 2007; Lysková et al., 2007a; Pinho et al., 2013). Of the 39 isolates tested, 23 % (9/39) were resistant to at least one antibiotic class. The prevalence of resistance in the current study was lower than that described in the literature, where, for example, tetracycline resistance is estimated to be expressed by approximately 30-40 % of isolates (Galpérine et al., 2007; Lysková et al., 2007a; Pinho et al., 2013). The genomic predictions of AMR generated in the current work matched the phenotypic results for 8 out of 9 isolates. One canine isolate was fully sensitive to all antibiotics tested, despite lincosamide resistance being anticipated based on the carriage of the *lsaC* gene (Malbruny et al., 2011). This may suggest that the *lsaC* gene was not expressed or another compensatory mechanism prevented the phenotypic lincosamide resistance being expressed *in vitro*, although lincosamide resistance *in vivo* cannot be excluded. One human isolate, in contrast, showed tetracycline resistance without carrying any known tetracycline resistance gene. Since the acquisition of rRNA mutations that can be associated with tetracycline resistance was ruled out, we suggest that the acquisition of novel resistance mechanisms could be responsible for the observed resistance phenotype (Thaker et al., 2010). A difference in the prevalence of AMR in human, 36% (4/11), vs companion animal isolates, 18% (5/28), was noticed. Although this result may be biased by the underlying reasons for which samples were collected, by the year, host, country of isolation and small sample size, it may be hypothesised that antimicrobial use in human medicine has increased the selective pressure on *S. canis* strains carried by humans. Importantly, no ß-lactam resistance was encountered in this study, suggesting that first-line antimicrobials such as amoxicillin clavulanate and penicillin G are still a suitable option to treat *S. canis* infections.

The presence and distribution of homologs to known virulence genes was determined in the current dataset. Our approach, which relied on the use of a virulence gene database as a reference, limited the search to well established genes of important human and animal pathogens. The *scm* gene, for example, was at first not detected in the dataset because

it was absent from the VFDB. The carriage of *scm*, which is a *S. canis*-specific virulence gene, was later confirmed in all isolates. The chosen thresholds to define gene homology in our study were much stricter than those used in a previous work on a single *S. canis* genome (Richards et al., 2012) and, as a result, the number of positive matches obtained in the current study was considerably lower than those observed in the previous work. Nineteen virulence gene homologs were found in the present *S. canis* genome dataset. In other pathogenic streptococci, the corresponding virulence genes are involved in tissue adhesion, tissue invasion and immune response evasion (Kadioglu et al., 2008; Walker et al., 2014). Seventeen of the gene homologs detected (i.e. all except *aspA* and *ssp-5*) correspond to virulence genes also found in *S. pyogenes* (Walker et al., 2014), and 9 of these genes (*eno*, *fbp54*, *hasC*, *hyl*, *plr*, *rfbA*, *rfbB*, *sagA* and *ropA*) are considered part of the *S. pyogenes* core genome (Davies et al., 2019), providing further evidence of the close evolutionary relatedness between *S. canis* and *S. pyogenes* (Lefébure et al., 2012). A gene homologous to a *S. pyogenes* superantigen, *smeZ*, was found in one *S. canis* genome (accession number: SAMEA4968065). Streptococcal superantigens are potent exotoxins that play an important role in severe forms of infection (Sriskandan et al., 2007). Some superantigens are phage-encoded, allowing for intra and inter-species recombination events via lateral transduction, but *smeZ* is chromosome-encoded (Unnikrishnan et al., 2002) and the mechanism of acquisition of this gene in *S. canis* is unknown. To our knowledge, this is only the second time that a superantigen homolog has been found integrated in a *S. canis* genome (Igwe et al., 2003). In the current study, only the carriage of homologs to known virulence genes was considered, but the presence of unknown *S. canis*-specific virulence genes or genes that would be identified with less stringent settings cannot be ruled out.

A genome-based population analysis of *S. canis* was then carried out. First, a core-SNP phylogenetic tree was constructed. The core-SNP phylogeny, which is regarded as a highly accurate system to assess bacterial strain relatedness (Tsang et al., 2017), was initially used as a reference to validate the MLST and SCM typing schemes. The AU test was employed to compare the core SNP phylogenetic tree to MLST-constrained and SCM-constrained core SNP trees, revealing that both the MLST and SCM classification systems fail to represent with high accuracy *S. canis* population diversity. Since both the MLST and SCM schemes are

based on very limited fractions of the bacterial genome, this result is unsurprising and confirms published data (Tsang et al., 2017). In order to determine whether one typing scheme was more accurate than the other in predicting core SNP-based clustering, AW coefficients with 95 % CIs for the MLST and SCM systems in comparison with the CGS typing were calculated. The CI for both AW coefficients overlapped and therefore there was no evidence that one scheme performs better than the other. The SCM scheme, however, requires the sequencing of only a single gene instead of the seven loci that form the MLST system (Pinho et al., 2019; Fukushima et al., 2020a), making it easier to implement in a diagnostic laboratory. Conversely, the *S. canis* MLST database undergoes constant curation, while no formal SCM database has been developed to date. In light of our findings, we discourage the use of the MLST and SCM systems for fine-typing applications such as outbreak investigation, and we suggest the creation and curation of an SCM database to facilitate the identification of *S. canis* lineages.

The core SNP phylogenetic tree was also used to investigate evolutionary relationships among the isolates. No host-specific or country-specific clustering was observed in the phylogeny, suggesting that the same strains circulate across host and geographical boundaries. The presence of a significant host-specific or country-specific clustering in the core SNP tree was not tested for due to the small genome dataset available. The accessory genome plays an important role in shaping the evolution of bacterial pathogens (Croll and McDonald, 2012). An accessory genome network for the *S. canis* strains analysed was constructed and visually explored. We focussed on host-specific clustering, hypothesising that the accessory genome may influence host adaptation in *S. canis*. Similar to the core genome phylogeny, however, the accessory genome network showed no evidence of host-specific clustering.

Finally, an *S. canis* pangenome was constructed using the 59 sequences included in this study. A total of 4,426 genes, 1,432 of which classified as core genome, were identified in the *S. canis* pangenome. Compared to the core genome of *S. pyogenes*, which was defined by 1,306 coding sequences from 2,083 isolates in a work by Davies et al. (Davies et al., 2019), the *S. canis* core genome is somewhat larger. This may be related to the limited number of sequences included in the current study. The *S. canis* pangenome was used to perform a

pan-GWAS analysis that searched for the over-representation of genetic markers among the different host species. In *S. agalactiae*, for example, host adaptation seems to be driven by the presence of a limited number of genes in specific lineages (Crestani et al., 2021). In the current work, pan-GWAS analysis revealed that there were no genomic traits significantly associated with specific hosts. It should be noted, however, that the dataset utilised was smaller than similar studies in other streptococci, limiting the reliability of any statistical findings, including the pan-GWAS results. Nevertheless, our findings regarding core and accessory genome diversity and pan-GWAS analysis indicate a paucity of host adaptation for the pathogen *S. canis* which further strengthens the case for considering *S. canis* a multi-host pathogen with zoonotic potential.

Chapter 5

Data visualisation in the public health sector

5.1 Introduction

It is important to visually display data in order to easily and quickly communicate otherwise complex information. Data visualisation, as a field, encompasses a multidisciplinary approach to rendering data in graphical form and this has developed to become an essential part of every branch of science (Aparicio and Costa, 2015). While a good diagram can be an excellent means of communication, a poor graphical choice can confuse and mislead the target audience, being less clear and informative than a well-written piece of text (Tufte, 1985; Bateman et al., 2010). Research in the field of visual perception has highlighted how particular elements of a graph may be perceived more accurately than others by the human eye (Cleveland and McGill, 1987). Areas, volumes and colour hues within a figure, for instance, tend to be more prone to interpretation errors than angles, lengths and positions along a common scale (Cleveland and McGill, 1987). Basic guidelines for the successful use of different types of diagram or graph have been developed based on an appreciation of data visualisation concepts. Pie charts, for example, rely on area judgements and are prone to misinterpretation while box plots appear accurate but less intuitive than other graphs (Pierce and Chick, 2013; Siirtola, 2019). Due to the varied use of data visualisation in countless different contexts, it is difficult to define a set of universal rules to systematically apply to graph design and production. This is particularly evident when it comes to elaborate visual-

isations, such as infographics and maps, for which it may be difficult to find a good balance between graphical simplicity and memorability (Tufte, 1985; Bateman et al., 2010). In order to determine the most accurate and appropriate graphical choice, it is important to consider the dataset available, the message to be conveyed and the target audience.

The public health sector makes constant use of data visualisations to communicate to both scientific audiences and the general public (McCrorie et al., 2016). In the context of infectious diseases of public health importance, for instance, a wide array of visualisation methods may be used to inform different audiences about the burden, distribution, transmission and biological characteristics of infectious agents (Polonsky et al., 2019). These visualisations may be as simple as bar charts and histograms or as complex as phylogenetic trees and interactive maps (Polonsky et al., 2019). Since public health visualisations have the ultimate purpose of educating in order to protect the health and welfare of the population, their accuracy is essential, and so this consideration should be prioritised over all other elements that might influence the design process.

In the current work we address a gap in knowledge surrounding the accuracy of data visualisations used to communicate messages of public health relevance. Using a semi-qualitative approach, we investigated the visual preferences of a sample of participants with different public health-related backgrounds towards a set of visualisations that represent the epidemiology of iGAS disease in Scotland. The only currently available visualisations on the epidemiology of iGAS infections in Scotland are outdated and are simple in design (`https://www.hps.scot.nhs.uk/a-to-z-of-topics/streptococcal-infections/`). Although this study did not aim to draw universally applicable conclusions about the best practice for producing visualisations, we nonetheless made an initial step towards a more mindful approach to the graphical presentation of data of public health importance. This was achieved through the administration of an online survey to volunteers working in the public health sector. The purpose of the survey was to guide participants through the appraisal of a set of visualisations representing different aspects of the epidemiology of iGAS disease in Scotland in order to address the following research questions:

- Does the use of comparative data help with understanding the extent and range of the primary data?
- Are temporal trends best depicted by presenting data serially or by overlaying them?
- Is the use of visually simple graphical elements preferred over more informative but "busier" figures?
- Is the comparison of frequency measures best achieved in vertical or horizontal order?
- Is it easier to interpret outbreak data using straight or curved lines in a phylogenetic tree?

Aim and objective The overall aim of this work was to identify some basic guidelines that can be helpful for public health workers to communicate with each other accurate messages concerning the epidemiology of infectious diseases, particularly iGAS disease. The objective of the work was to synthesise the feedback obtained through the online survey to produce optimised figures.

5.2 Methods

All supplementary material for this chapter can be found in Appendix D (these are indicated with the letter D in front of the sequential number).

5.2.1 Data collection

Descriptive epidemiological data on iGAS disease in Scotland from 2014 to 2019 were collected as part of the routine diagnostic service by the Bacterial Respiratory Infection Service of SMiRL and were presented in chapter 2. Data from 2020 and 2021 were not available at the time this work was carried out. For each confirmed case of iGAS disease, information regarding date of isolation, age of the patient and *emm* type of bacterial isolates were gathered. Data were cleaned and re-formatted for later analyses using Microsoft Excel (Microsoft Corporation, 2021). Descriptive epidemiology visualisations were all produced on RStudio

v1.4.1103 (RStudio Team, 2020). This software, although not intuitive to use, allows the creation of good quality graphics with a high degree of flexibility.

Scottish demographic data were collected from the National Records of Scotland website (`https://www.nrscotland.gov.uk`) accessed on the 07/10/2021.

Genomic epidemiological data of GAS genotypes involved in cases of invasive disease in Scotland, namely *emm*5.23, were derived from WGS analyses described in chapter 3. Briefly, all invasive *emm*5.23 isolates collected in Scotland from 2015 to 2020 were whole-genome sequenced and core SNP maximum likelihood phylogenetic analysis was performed. The resulting phylogenetic tree was visualised on iTOL (Letunic and Bork, 2021) and annotated with Inkscape (Bah, 2007).

This work was structured as a collection of five distinct studies, each one focusing on a different dataset and investigating a different aspect of data visualisation. An outline of the five studies is reported below:

- Study 1, Y-iGAS. The yearly cumulative incidence of iGAS disease in Scotland between 2014 and 2019 was used to investigate the following question: does the use of comparative data help with understanding the extent and range of the primary data?
- Study 2, M-iGAS. The monthly burden of iGAS disease in Scotland from 2014 to 2019 was used to investigate the following question: are temporal trends best depicted by presenting data serially or by overlaying them?
- Study 3, A-iGAS. The yearly cumulative incidence of iGAS disease in different age groups in Scotland from 2014 to 2019 was used to investigate the following question: is the use of visually simple graphical elements preferred over more informative but "busier" figures?
- Study 4, E-iGAS. The *emm* type-specific invasive disease burden of iGAS in Scotland from 2014 to 2019 was used to investigate the following question: is the comparison of frequency measures best achieved in vertical or horizontal order?

- Study 5, P-iGAS. Phylogenetic trees that communicate iGAS outbreak data were used to investigate the following question: is it easier to interpret outbreak data using straight or curved lines in a phylogenetic tree?

5.2.2 Visualisations

The R code used to generate the visualisations discussed in this section is presented in Appendix D.1.

Yearly cumulative incidence of iGAS disease (Y-iGAS) The principle behind this component of the study was to investigate whether presenting comparative data helps the user understanding the primary data. Two alternate visualisations were made to represent the yearly incidence of iGAS disease in Scotland over the six years considered (Figure 5.1). Both Figures 5.1A and B were designed as a graph, having 'year' on the X-axis and incidence of disease on the y-axis. Figure 5.1A is a simple bar chart with all bars having the same colour and no background grid. Figure 5.1B is a line chart that allows a comparison of iGAS disease incidence in Scotland and incidence of confirmed hospitalised cases of influenza in a network of acute trusts in England. Influenza data are used here to help the viewer understand the magnitude of iGAS incidence in Scotland. Although influenza data refer to a subset of the English population and not to the whole Scottish population, they are still useful as a proxy of the incidence of severe cases of influenza across the UK. Influenza data are used exclusively in Figure 5.1B and not in its counterpart. Figure 5.1B was designed as a line chart because the slope of the line may be used to communicate the magnitude of changes from year to year.

Monthly burden of iGAS (M-iGAS) The main principle behind this component of the study was to investigate if temporal trends are best depicted by presenting data serially or by overlaying them. The monthly incidence of iGAS disease in Scotland was visualised in the form of two histograms and a line chart, as shown in Figure 5.2. Both Figures 5.2A and B are histograms showing the incidence of iGAS disease per month from the beginning of 2014 to the end of 2019. In both cases the x-axis represented the entire study period. Figure 5.2A is monochromatic and has a background grid to help identify the number of

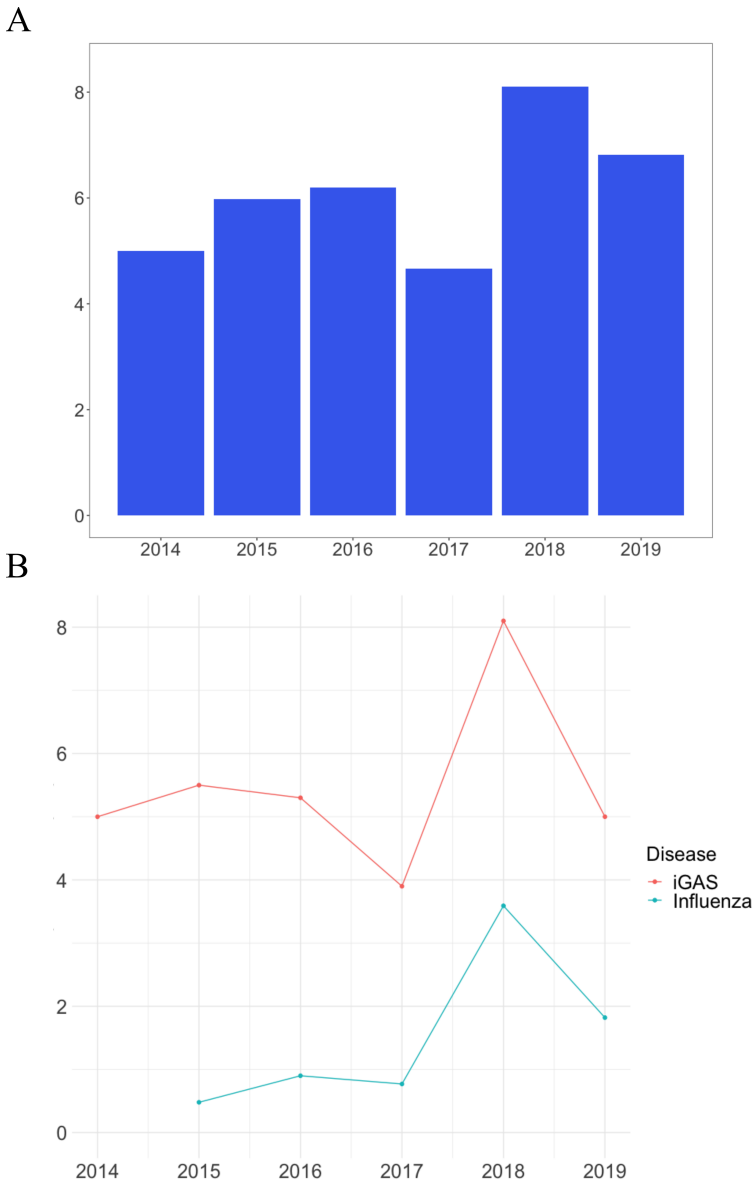


Figure 5.1: A - Yearly incidence of invasive Group A *Streptococcus* (iGAS) disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people. B - Pink line - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people. Blue line – Mean weekly incidence of confirmed hospitalised cases of influenza expressed as cases per 100,000 people in the flu season from 2015 to 2019. Influenza data was collected by the USISS (UK Severe Influenza Surveillance Systems) sentinel scheme, a sentinel network of acute trusts in England who report weekly aggregate numbers on laboratory confirmed influenza hospital admissions at all levels of care.

confirmed new cases per month. In Figure 5.2B bins were coloured according to the year and no background grid was added. Figure 5.2C is a line chart where each line corresponds

to one of the six years considered. The X-axis of this figure represents the months of the year. Figure 5.2C was designed with the aim of highlighting seasonal patterns of disease incidence rather than year and month-specific incidence values.

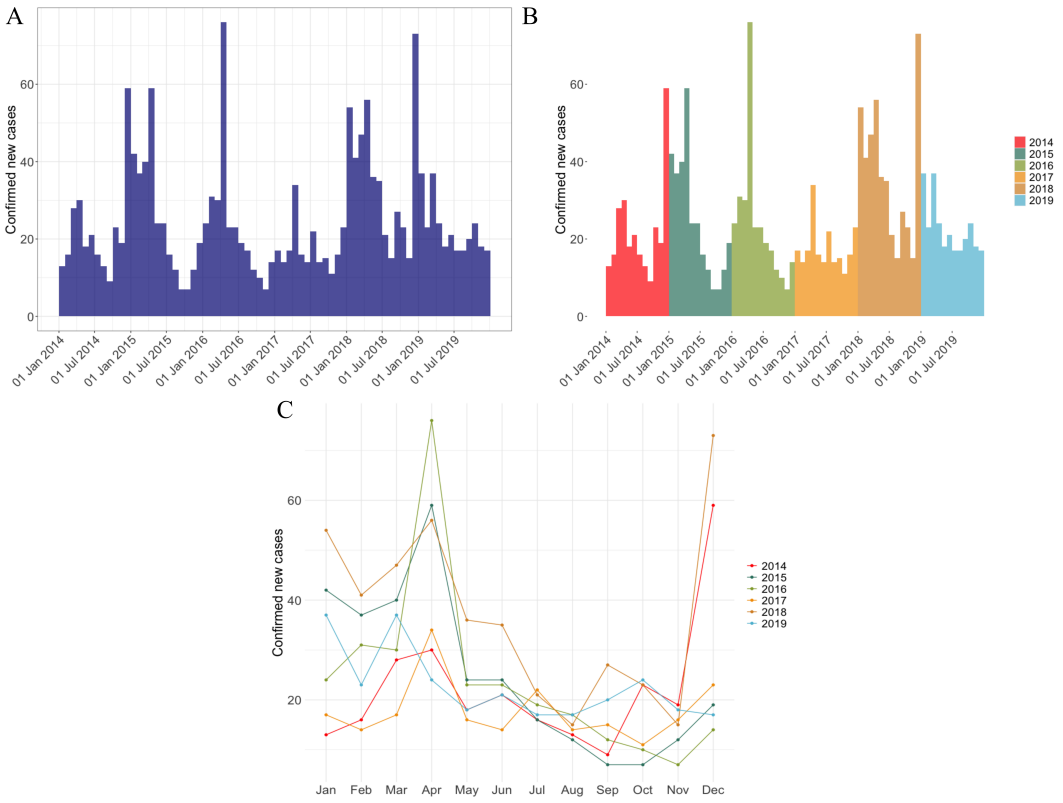


Figure 5.2: A, B, C - Monthly burden of invasive Group A *Streptococcus* disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

Age-specific incidence of iGAS disease (A-iGAS) The main principle behind this component of the study was to investigate whether the use of visually simple graphical elements is preferred over more informative but 'busier' figures. The yearly incidence of iGAS disease in different age groups in Scotland was visualised through three different design options (Figure 5.3). Figure 5.3A is a simple box plot resulting from the age-specific incidence of iGAS disease from 2014 to 2019 (six data points); colours here are used exclusively for aesthetic purposes. Figure 5.3B is a scatter plot superimposed on a box plot. In this figure, boxes are empty and dots are coloured according to year. This allows appreciation of not only the variability of age-specific incidence throughout the study period, but also the age-specific incidence values observed for each year. Figure 5.3C is a black and white scatter plot with background grid. In this visualization, each dot represents the age-specific incidence of

iGAS infection for each of the six years considered, although it is impossible to deduce the year associated with each dot.

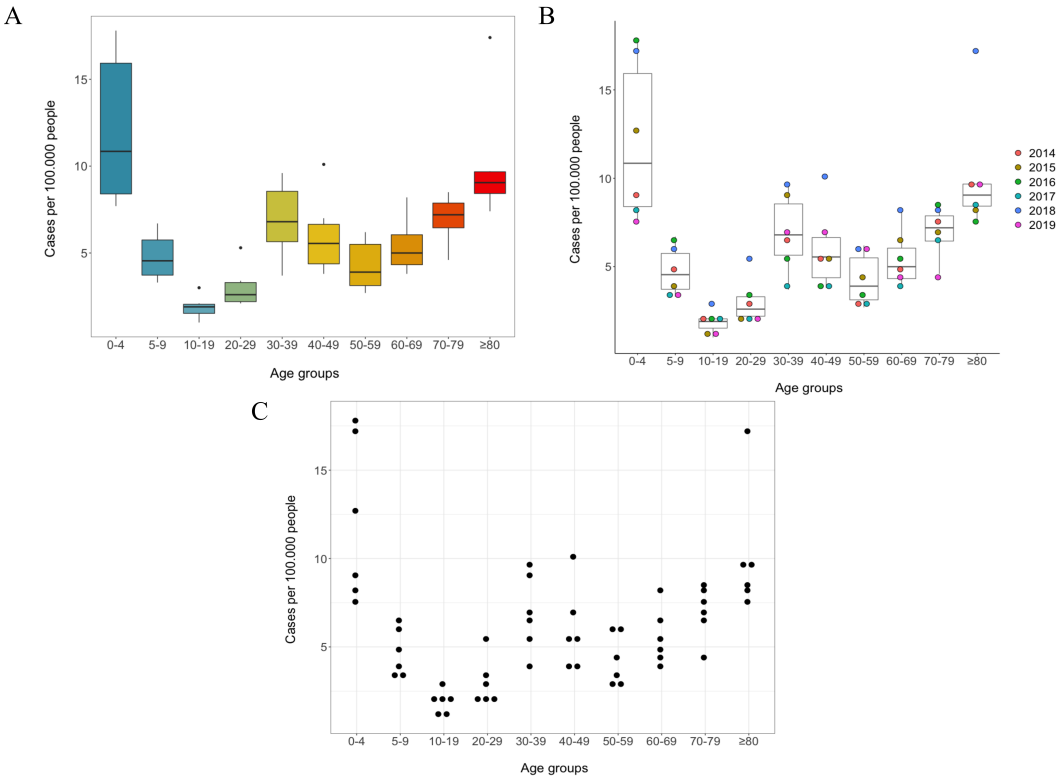


Figure 5.3: A, B, C - Age-specific incidence of invasive Group A *Streptococcus* disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

***emm*-specific burden of invasive disease (E-iGAS)** The main principle behind this component of the study was to investigate whether the comparison of frequency measures is best achieved in vertical or horizontal order. GAS strains are classified as *emm* types. The invasive disease burden of the 15 most common *emm* types circulating in Scotland during the study period was visually rendered using three design options, as shown in Figure 5.4. Figure 5.4A comprises six bar charts, each one reporting the number of isolates of a specific *emm* type per year. Bars are coloured according to the year they refer to and a background grid is added to aid data interpretation. Figure 5.4B is similar to Figure 5.4A, except that neither colour nor a background grid is used and *emm* type relative frequency is expressed rather than actual isolation number. Figure 5.4C is a stacked bar chart of the relative frequency of isolation of *emm* types per year. Each bar corresponds to a year and colour is used to indicate *emm* type.

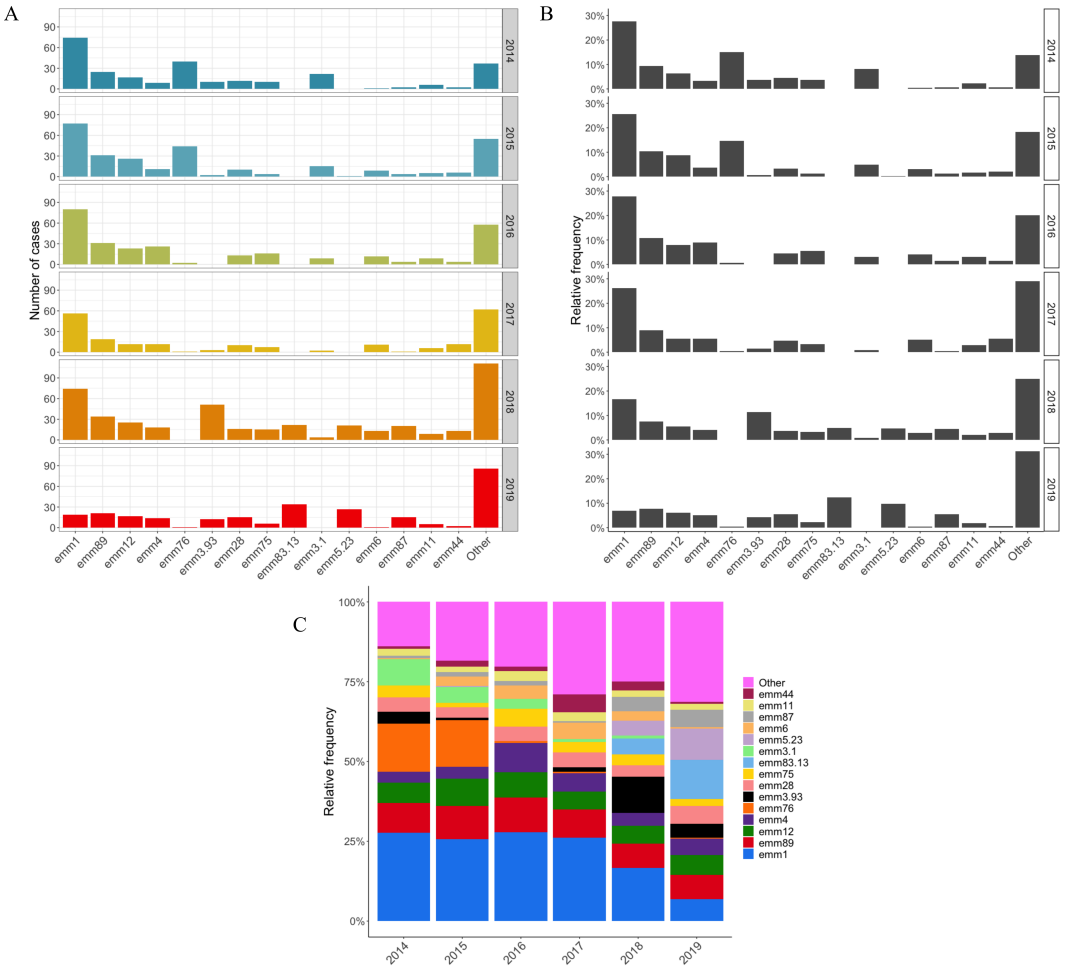


Figure 5.4: A - Absolute frequency of isolation from normally sterile body sites of the 15 most common *emm* types circulating in Scotland from 2014 to 2019. B, C - Relative frequency of isolation from normally sterile body sites of the 15 most common *emm* types circulating in Scotland from 2014 to 2019.

iGAS outbreak data using phylogenetic trees (P-iGAS) The main principle behind this component of the study was to investigate whether it is easier to interpret outbreak data using straight or curved lines in a phylogenetic tree. The same phylogeny, constructed as described in Chapter 3, was rendered as a circular and a rectangular tree using iTOL (Letunic and Bork, 2021) (Figure 5.5). Both tree topographies are commonly used in research publications but no data is available on the relative ease of interpretation.

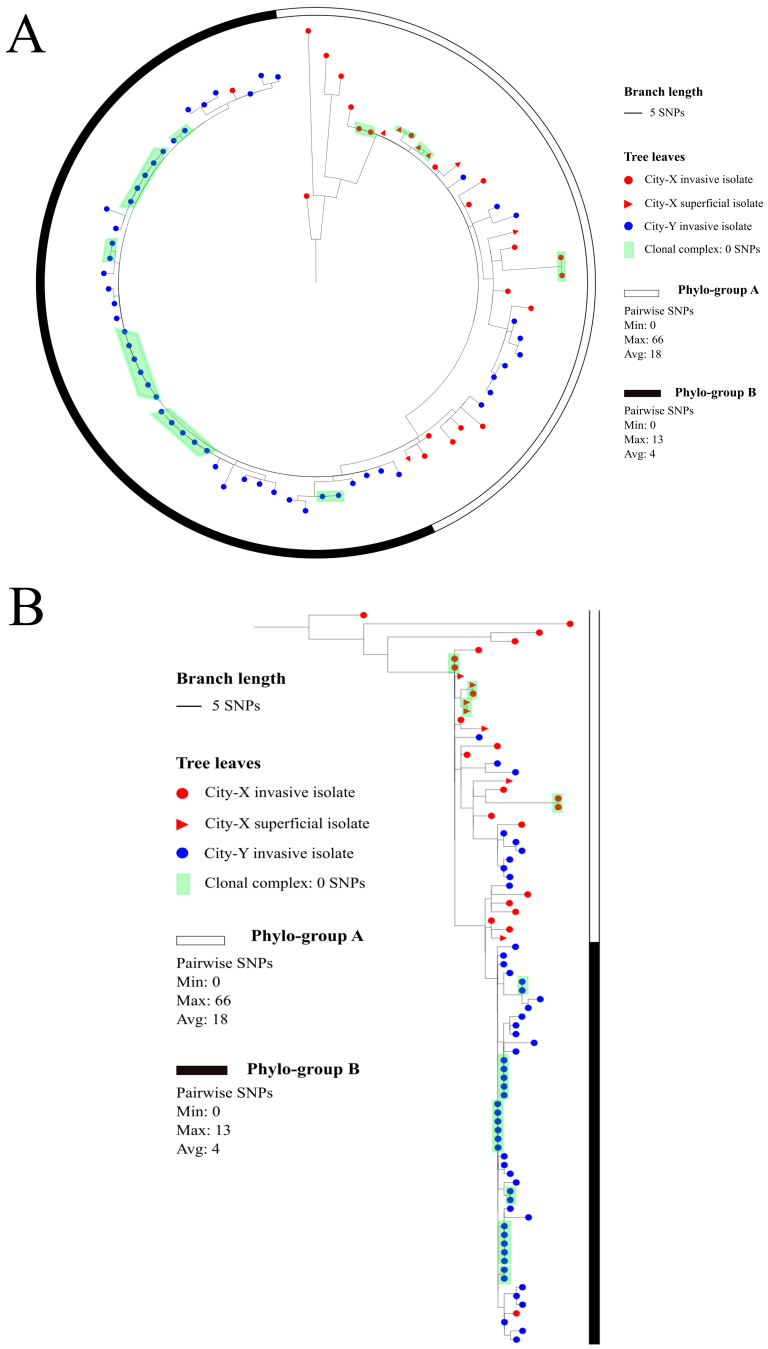


Figure 5.5: A, B - Core single nucleotide polymorphism phylogenetic trees of all the *emm*5.23 isolates involved in invasive disease in Scotland from 2015 to 2020. Both trees were midpoint rooted.

5.2.3 Online survey

An online survey was designed to gather user-experience data on the different visualisations generated. A suitable webform was created on the Google Forms platform (`https://www.google.co.uk/forms/about/`). The survey comprised multiple choice and

short open questions. Ethical approval to administer the survey to people working in the public health sector was sought and received from the University of Glasgow Ethics committee (project number: 200210075). Participants working at SMiRL and UKHSA were recruited via email and responses collected anonymously. A copy of the online survey is presented in Appendix D.1.5.

5.3 Results

Thirty-three responses were obtained from public health workers. The majority of the responses (n = 19, 58%) derived from individuals involved in microbiology laboratory or clinical microbiology activities on a daily basis. Six respondents (18%) identified their occupation as being in public health and four (12%) reported having a career in infectious disease. One respondent identified as a sexual health worker, one as a pharmacist and one as an infection control worker.

5.3.1 Yearly incidence of iGAS disease in Scotland (Y-iGAS)

Figure 5.1A was described as simple and effective by 21 (64%) of the answers and as basic and unremarkable by 11 (33%) of the answers, suggesting that simplicity was positively perceived by most participants. One person described it as a combination of both descriptions. Influenza data were included in Figure 5.1B to help the viewer contextualise the frequency of iGAS infections. The majority of the respondents, n = 27 (82%), found the presence of influenza data in Figure 5.1B helpful, while four (12%) participants thought it was a distracting element. Two respondents commented on the use of different comparison data (e.g. scarlet fever) and on the possible distracting effect of additional epidemiological data depending on the target audience. When asked to suggest how to improve Figure 5.1A, ten (30%) participants answered that the figure did not need any improvements. Among the suggestions received, the most frequent was to make the bars thinner, to add a background grid, to make the figure more visually engaging, to add a trend line and metadata, such as the total number of cases per year. Overall, the feedback received for Figure 5.1A indicated that, although the simplicity of the figure was appreciated, the use of further graphical elements and data would make the visualisation more engaging and informative. As for Figure 5.1B, 45% of

the respondents did not think any adjustments were necessary. Suggestions to improve Figure 5.1B included making the lines thicker, changing the comparison data (e.g. use "flu" data that refer to Scotland or use data regarding a bacterial respiratory infection), using different colours, such as a colour blind friendly palette, expressing iGAS data in the form of bars and influenza data as a line, removing the background grid and increasing the size of the data points.

Based on the feedback received on Figures 5.1A & B, an optimised visual representation of the yearly incidence of iGAS disease in Scotland was produced and is shown in Figure 5.6. This figure combines elements of both options, reporting iGAS incidence in the form of a trend line and a bar chart. Influenza data are rendered in the form of a line chart. In the representation of the iGAS yearly incidence, data points are also bigger, trend lines are thicker and the colours are easily distinguished by colour blind readers.

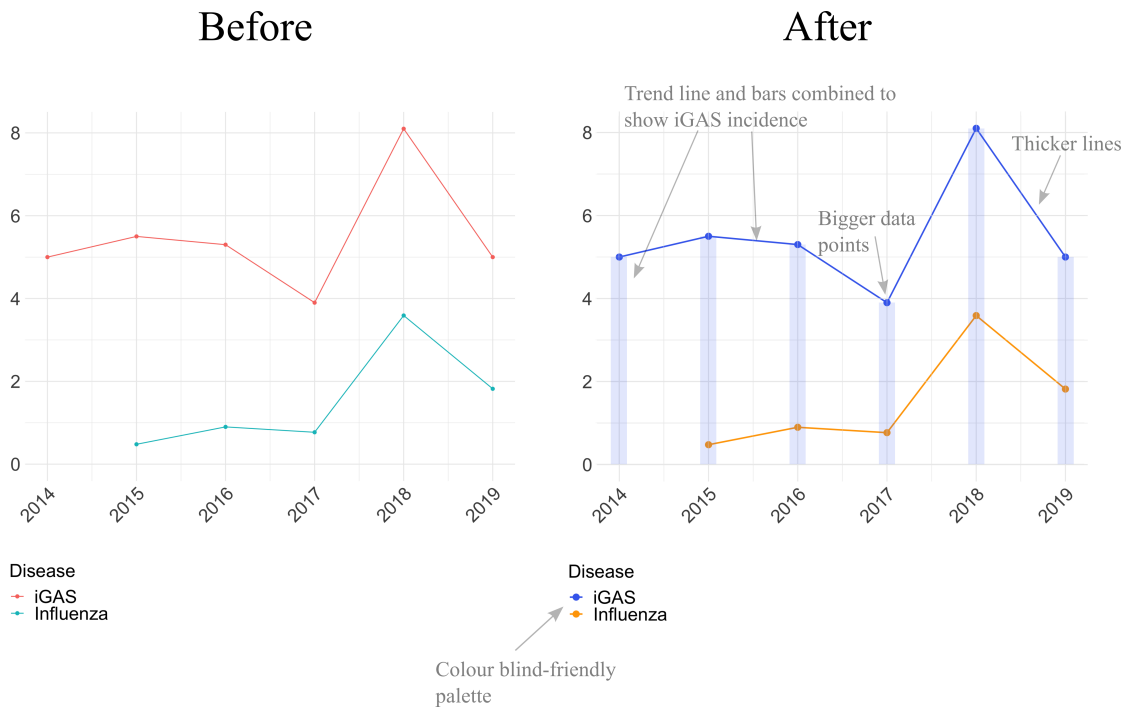


Figure 5.6: Optimised visual representation of the invasive Group A *Streptococcus* yearly incidence.

5.3.2 Monthly burden of iGAS disease (M-iGAS)

In Figure 5.2A, the presence of a background grid was considered useful by 79% of the respondents and distracting by 18% of them. The use of colour in Figure 5.2B was seen

as an improvement from Figure 5.2A by 79% of the participants. When asked to select the statement that better reflected their opinion about Figure 5.2C, 58% of the participants stated that "It is hard to detect the disease burden in a specific point in time" while 42% of them claimed that "It is a good way to compare the seasonal trends of iGAS disease across the six years considered". Most of the participants (49%) found option C to be the best representation of iGAS disease seasonal trend. A considerable proportion of respondents (42%), however, expressed a preference for option B, suggesting that both line charts and histograms are valid choices to represent seasonal trends of disease over time, provided that the former are kept simple and the latter are made visually engaging.

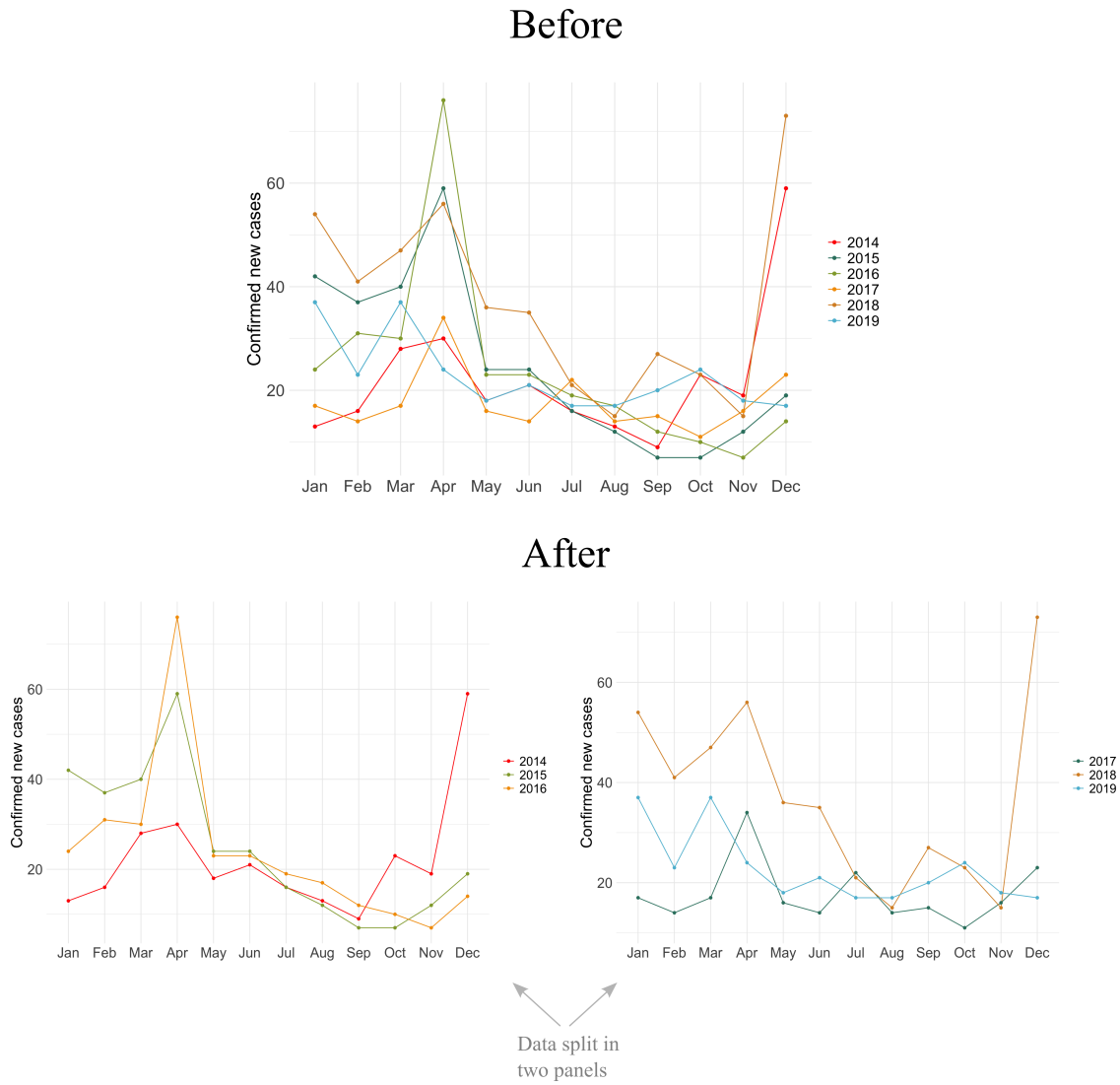


Figure 5.7: Optimised visual representation of the monthly burden of invasive Group A *Streptococcus* disease.

Based on the feedback received on Figures 5.2A, B & C, an optimised visual representation of the monthly burden of iGAS disease in Scotland was produced and is shown in Figure 5.7. Since most of the respondents found Figure 5.2C the best representation of iGAS disease seasonal trend, the new visualisation represents an improved version of that figure. Figure 5.2C was criticised for being too busy, so the optimised visualisation shows the same data split in two panels, facilitating data interpretation.

5.3.3 Incidence of iGAS disease in different age groups (A-iGAS)

The majority of the participants (82%) found that the use of colours in Figure 5.3A attracted their attention and helped them to focus on the data displayed, despite the colours in this visualisation having a purely aesthetic function. In Figure 5.3B, being able to visualise the age-specific incidence for each of the six years was considered as a relevant element by 67% of the respondents, although 46% found the data difficult to interpret due to the many colours used. One third of participants, on the other hand, thought that an average of the age-specific incidence across the years would have been sufficient for this figure. Option C was described as incomplete or too plain by, respectively, 58% and 39% of the respondents. However, it was appreciated for being simple and communicative by 27% of the participants. For the question "Which of the three options better shows the difference in iGAS disease incidence across age groups?" most of the respondents (70%) selected option A, followed by option B (27%) and C (3%).

Based on the feedback received on Figures 5.3A, B & C, an optimised visual representation of the age-specific incidence of iGAS disease in Scotland was produced and is shown in Figure 5.8. Most of the respondents expressed their preference towards Figure 5.3A, appreciating the engaging colour palette employed and favouring the use of a box plot over a scatter plot to represent this dataset. The optimised version of Figure 5.8A differs only from the original by the presence of a background grid, which has been highlighted as a helpful element in the assessment of Figures 5.1 and 5.2. This result, together with previous responses, suggests that the use of a box plot that summarises several data points is preferable to an otherwise busy-looking scatter plot, even though it can be less informative.

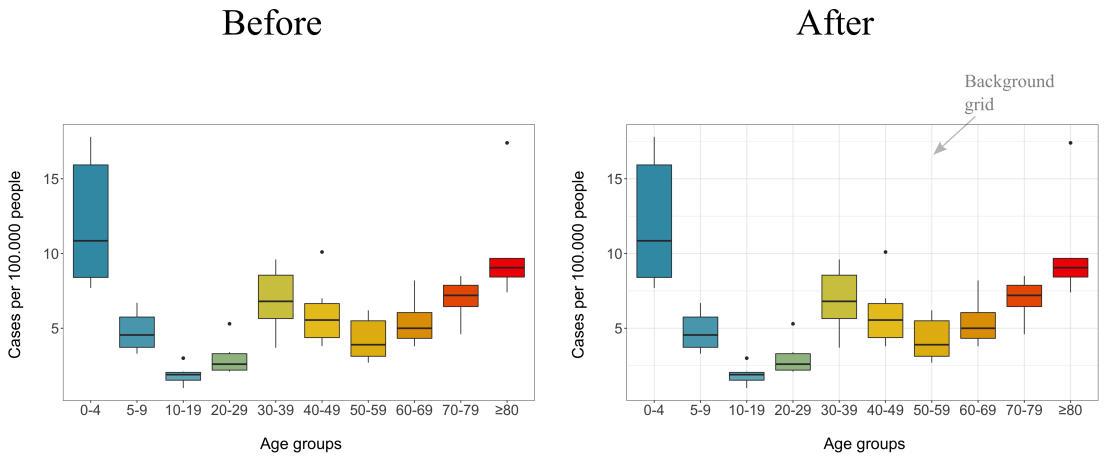


Figure 5.8: Optimised visual representation of the incidence of invasive Group A *Streptococcus* disease in different age groups.

5.3.4 *Emm* type-specific invasive disease burden (E-iGAS)

One of the main differences between Figure 5.4A and B is that the former shows the absolute count while the latter the proportion of isolates of specific *emm* types per year. Most of the participants (61%) expressed a preference for the use of absolute frequencies (option A) to represent this particular dataset. The fact that absolute frequencies were considered the best representation of the *emm* type-specific disease burden suggests that most of the public health workers are particularly interested in the actual number of cases involving each *emm* type on the Scottish population. Twenty-six participants (79%) preferred option A to option B for the use of colours, which were described as an attractive element that helped to retain focused on the displayed data. The presence of a background grid in Figure 5.4A was described as helpful in 36% of the answers, while the absence of a background grid in Figure 5.4B was highlighted as a positive element only by one person. These responses indicate that, despite in most of the cases the background choice not being crucial, for some people the presence of a background grid might have facilitated data interpretation. Referring to positive attributes of Figure 5.4C, 73% of the participants indicated that the figure is easy to interpret and a clear way to represent changing trends in *emm* type prevalence across the years. Among the negative attributes of Figure 5.4C reported by the respondents, the most common was the fact that the figure looked 'too busy' (55% of answers), the difficulty in interpreting the proportion of less common *emm* types (21% of answers) and the colour choice (9%), particularly with regard to colour-blind readers. For the question "Overall, which one of the

three options better represent the fluctuations in *emm*-specific disease burden?", 55% of the participants selected option C, 39% option A and 6% option B.

Based on the feedback received on Figures 5.4A, B & C, an optimised visual representation of the *emm* type-specific invasive disease burden in Scotland was produced and is shown in Figure 5.9. Figure 5.4C was the preferred option of the respondents, so the new visualisation is an improved version of this figure. Fewer *emm* types are shown in the new figure compared to the previous version in order to make the visualisation less "busy" and easier to interpret. Different colours have also been used, in consideration of colour blind readers.

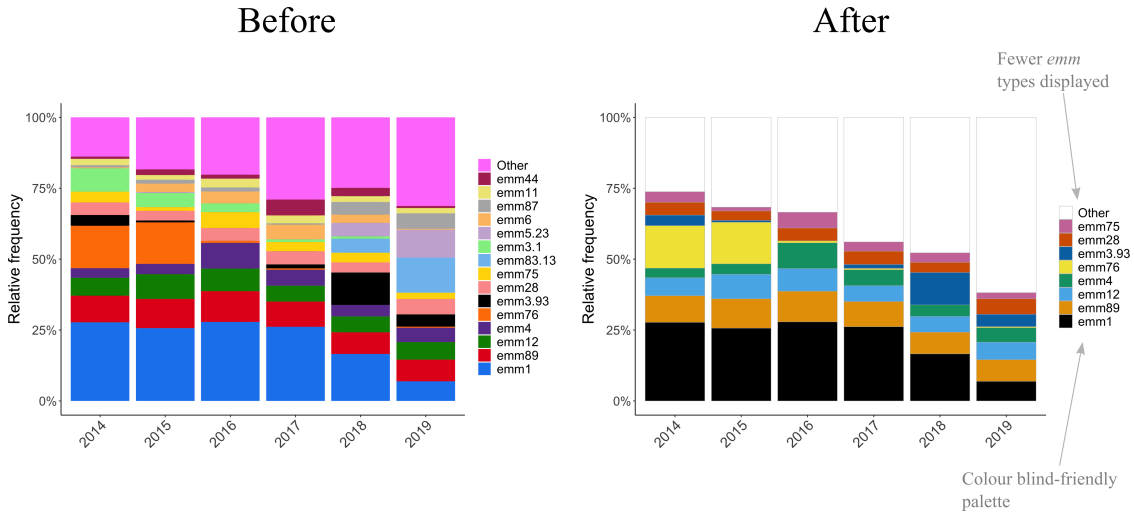


Figure 5.9: Optimised visual representation of the *emm* type-specific invasive disease burden.

5.3.5 Phylogeny (P-iGAS)

The information content of the trees shown in Figures 5.5A & B was identical, i.e. they displayed exactly the same phylogenetic data. However, one phylogeny was represented as a circular tree (option A) and the other one as a rectangular tree (option B). The majority of participants (70%) found Figure 5.5A more "captivating" than Figure 5.5B. The latter, however, was considered easier to read by 64% of the respondents. This suggests that, although circular phylogenies are more engaging and potentially memorable, they are also less intuitive than rectangular ones. The familiarity of public health workers with phylogenetic tree interpretation was probed by asking whether they thought option A and option B displayed the same underlying data. Twenty-one of them (64%) responded that both trees displayed the

same information, as far as they could tell. Nine (27%) said they were unsure, mainly due to the lack of familiarity with this kind of visualisations. Three participants (9%) answered that that option A and B represented different datasets. Altogether, more than a third of the respondents failed to correctly interpret Figure 5.5, indicating how even commonly used visualisations can be misinterpreted by public health workers. It should be remembered, however, that phylogenetic tree construction and interpretation require specialised skills that are not part of the educational background of most people, even among public health workers.

5.4 Discussion

In this work, a semi-structured online survey was used to explore the preferences of public health workers regarding the visual representation of epidemiological data. The study was undertaken in order to gain insight and improve understanding of this area. This is not the first time that data visualisation in the public health sector has been investigated (Park et al., 2022). To our knowledge, however, this is the first work focusing on basic graphical principles rather than complex visualisation tools and techniques. It should be remembered that only a relatively small number of respondents were recruited in the current work, that only a target audience was involved and that many uncontrolled variables might have influenced the answers received (e.g. the screen used, time availability, etc.). The findings of this chapter, thus, should not be considered as generalised conclusions concerning data visualisation in the public health sector. These can, however, can still be helpful to promote a more thoughtful use of data visualisation to communicate public health data. Below, we report and discuss some principles we were able to distil from the analysis of the responses:

1. **The presentation of comparative data can help with understanding the extent and range of a primary dataset.** This principle, which is supported by the results obtained for study 1 (Y-iGAS), is implicitly applied to visualisations describing characteristics of multiple infectious diseases. Having in the same figure data pertaining to both commonly known and uncommon diseases helps to understand the impact of the latter. For example, the interactive visualisation tool MicrobeScope (`https://informationisbeautiful.net/visualisations/`

the-microbescope-infectious-diseases-in-context/) graphically displays different parameters (e.g. fatality rate, average basic reproduction number and incidence) of many fatal infectious diseases worldwide, allowing an intuitive comparison between familiar and uncommon diseases. Another example of this principle being utilised is in a recent work by Bessel and colleagues (Bessell et al., 2020). The first figure in this publication shows the risk of introduction in Scotland of important animal infectious diseases. Also in this case, comparing different datasets allows an easier understanding of the impact of lesser known infections.

2. **Temporal trends are generally best depicted by superimposing data visualisations.** The majority of survey participants indicated that a superimposition of trend lines within a single 12-month time-frame was a better representation of the iGAS seasonal pattern than a histogram showing a progressive transition from 2014 to 2019. Having excessive lines in a single graph, however, proved to be a limitation of Figure 5.2C. The choice of line charts to depict temporal changes in the epidemiology of infectious diseases is quite common. UKHSA, for example, uses line charts when reporting seasonal data on GAS infections in England, as shown in Figure 5.10. It should be noted, however, that a large proportion of participants in the current work preferred a histogram to the line chart option in study 2 (M-iGAS). It is unsurprising that histograms are also commonly employed to represent temporal changes in infectious disease epidemiology, as witnessed by the use of histograms to report COVID-19 daily new cases and daily deaths in the popular web- site Worldometer (https://www.worldometers.info/coronavirus/).

3. **A simple visualisation may be better received than a more informative but cluttered figure.** This principle is supported by the results obtained for study 3 (A-iGAS). Choosing not to include all available information in a figure might sound inappropriate. Efficiently communicating a portion of the data, however, may be better than presenting all available information, which may result in a complex and intimidating diagram. Visual simplicity, or graphical minimalism, has been a source of debate in the field of data visualisation for years (Bateman et al., 2010). An important distinction to be made when discussing visual minimalism is between content and ap-

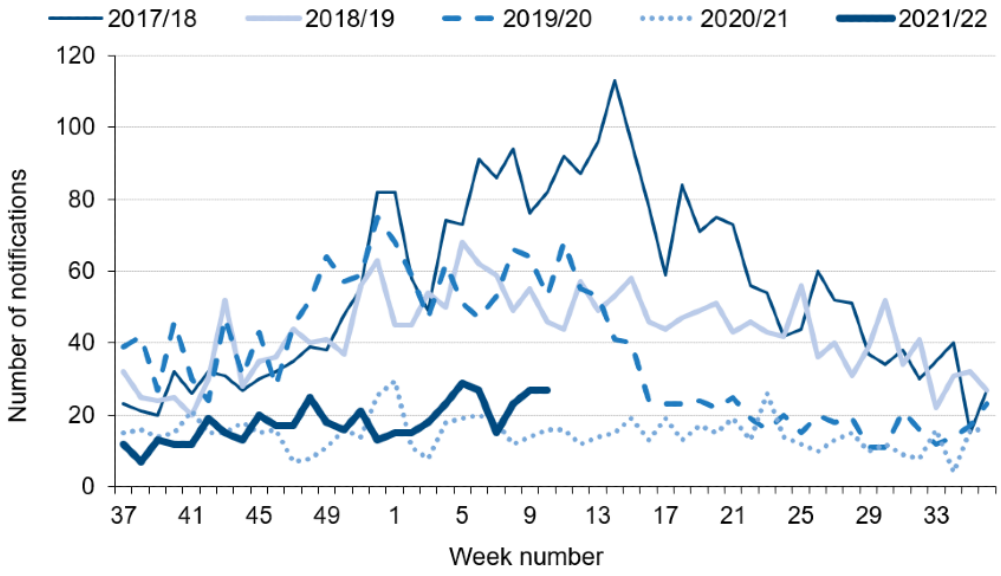


Figure 5.10: Weekly laboratory notifications of invasive Group A *Streptococcus* in England from 2017 to 2018 season onwards. From `https://www.gov.uk/government/publications`

pearance. The former refers to the range of different data displayed, while the latter concerns the overall appearance of a figure. Our results for study 3 (A-iGAS) indicate that public health workers appreciate figures simple in content but not "plain". This concept is at the base of illustrator Nigel Holmes's work (Holmes and Heller, 2006), which uses visual embellishments to turn graphs simple in content into visually engaging figures. The use of visualisations rich in content, however, is common in public health. These visualisations, which can be captivating and extremely informative, can also be overwhelming and intimidating to some readers. Some examples are provided by the interactive visualisations produced by the ECDC, like the COVID-19 vaccine tracker (`https://vaccinetracker.ecdc.europa.eu/public/extensions/COVID-19/vaccine-tracker.html#uptake-tab`).

4. **Comparing frequencies across different data series is easier when graphical elements are arranged horizontally rather than vertically.** This principle is supported by the findings of study 4 (E-iGAS). The overall preference for the use of a stacked bar chart over a series of vertically aligned simple bar charts is likely due to

the demonstrated difficulty in comparing non-adjacent bars in simple bar charts (Talbot et al., 2014). Our findings are in line with the results described by Xiong and colleagues, who concluded that having vertically aligned bar charts favours comparisons of elements within the same group but does not allow intuitive comparisons between elements from different groups (Xiong et al., 2021). Despite these reservations, vertically aligned bar charts have been used by the Scottish government to communicate epidemiological data on COVID-19 transmission, as demonstrated by Figure 9 of the following report: `https://www.gov.scot/publications/coronavirus-covid-19-modelling-epidemic-issue-no-24/`.

5. **Phylogenetic trees constructed using straight branches are easier to interpret than circular trees.** This principle is supported not only by the results obtained in study 5 (P-iGAS), but also by the experimental work of Xu and colleagues, which showed that the use of straight lines in graphs is associated with reduced interpretation time and increased interpretation accuracy compared to curved lines (Xu et al., 2012). Whenever possible, the use of rectangular phylogenetic trees should thus be preferred to circular formats, although circular phylogenies are regularly featured in the literature (Figure 5.11).

6. **Visualisations should appear simple but be visually attractive.** This principle is supported by the responses received in studies 1-4. The majority of the participants expressed a preference for visually engaging graphical options, claiming that the use of a good colour palette could help them to keep focused on a figure. This principle is in contrast with the views of statistician Edward Tufte, who believes that any visual embellishment distracts the reader and hampers data interpretation (Tufte, 1985). The reports published by UKHSA on GAS epidemiology appear to embrace Tufte's approach, showcasing simple and plain graphs as Figure 5.12. On the other hand, limited experimental evidence suggests that visually engaging figures are interpreted as accurately as plain ones, while being more enjoyable and memorable (Bateman et al., 2010).

7. **Background grids may be used instead of blank backgrounds.** This principle is supported by the results obtained in studies 1, 2 and 4. According to the answers

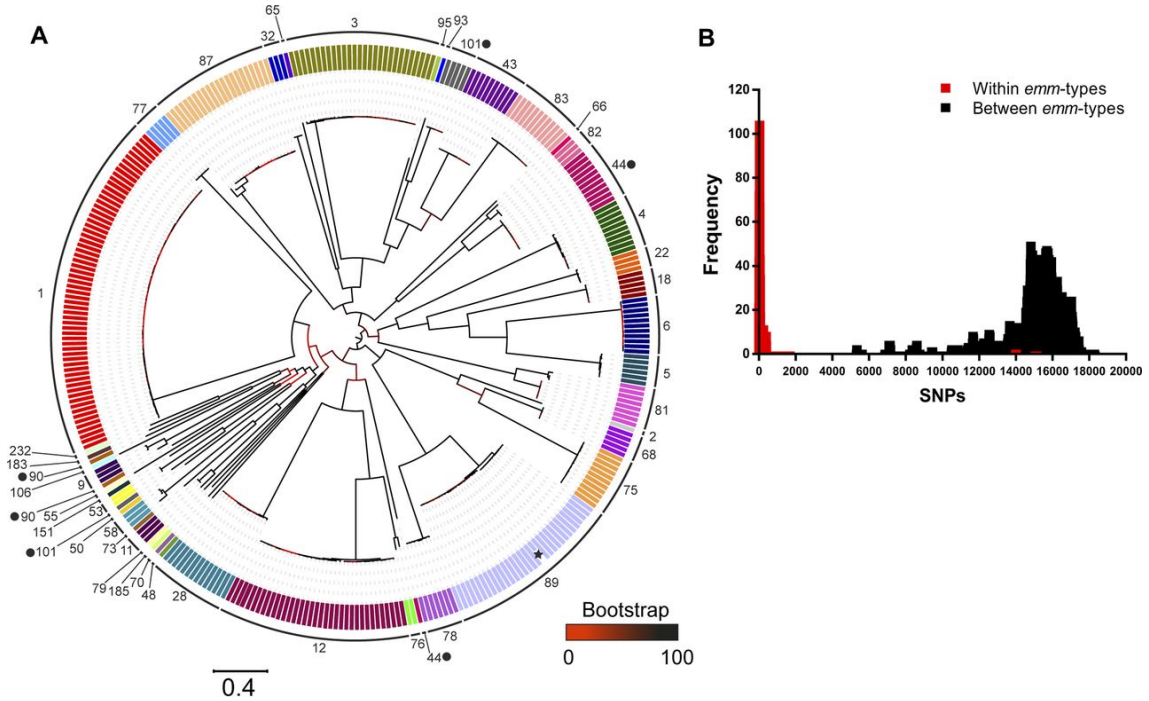


Figure 5.11: Example of a Group A *Streptococcus* core single nucleotide polymorphism maximum likelihood phylogeny from Turner et al. (2019).

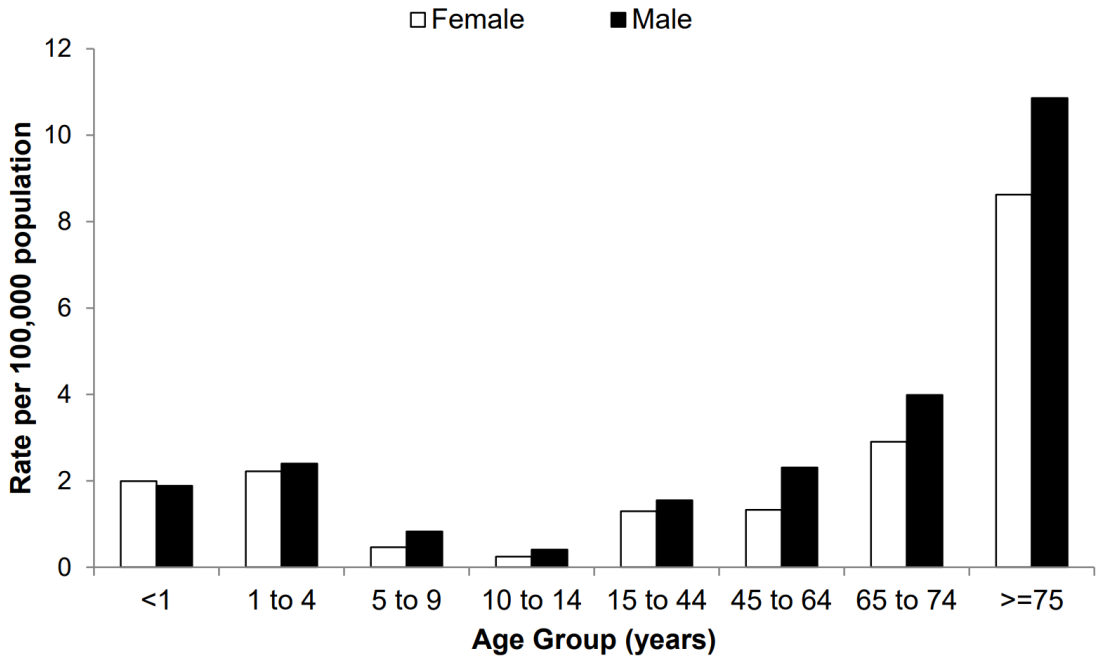


Figure 5.12: Rates of Group A *Streptococcus* bacteraemia in different age groups in 2020 in England. From `file:///C:/Users/david/Desktop/hpr1921_strptcccl-BSI_2020.pdf`.

collected, the use of a background grid is not always crucial but it is rarely perceived as a negative element and for some people it facilitates data interpretation. Also in this

case our findings are in contrast with the point of view of Edward Tufte, according to whom any graphical element that is non-essential should be avoided (Tufte, 1985).

8. **Depending on the dataset in use, expressing data through absolute measures may be more meaningful than using percentages.** This principle is supported by the responses received in studies 1 and 4. Due to the nature of their job, public health workers are particularly interested in the absolute impact of infectious disease in the population. In our work, the *emm*-specific disease burden expressed as number of disease cases was more meaningful than a measure of relative frequency.

9. **Not all visualisations are accessible to everyone.** This principle is supported by the results of study 5 (P-iGAS). Even in a specialised audience, not everyone has the same professional and educational background and this should be considered when presenting a figure. In the case of phylogenetic trees, for example, a thorough caption should be used to describe the figure and make it accessible to everyone.

10. **A constructive discussion with colleagues or members of the target audience should always form part of the design process.** The approach used in this study, which is based on improving visualisations through feedback from the target audience, should be applied as often as possible, even on a smaller scale. Given the paucity of broadly accepted guidelines for data visualisation, it is easy to be biased by our own personal preferences, forgetting that figures are primarily used to communicate data to others.

Chapter 6

General discussion

The approaches and tools used to understand and control the spread of infectious agents have considerably expanded since the very first epidemiological investigation, led by John Snow in 1855 (Snow, 1855). For example, sophisticated phenotyping and molecular techniques, ranging from AST to WGS, have become irreplaceable components in the study of infectious diseases (Weber et al., 1997). Despite the advances in the field of molecular epidemiology, more traditional disease investigation approaches relying on the collection and analysis of population-based epidemiological data are still very much in use because they narrate the story from a different perspective. Only by combining these two different perspectives, the former focusing on the pathogens and the latter on the host populations, can we hope to fully understand and control the impact of infectious diseases (Le and Diep, 2013). This PhD thesis represents an effort to provide a synthesis of these two approaches. The aim of this work was to apply a multi-disciplinary approach to investigate infections caused by *S. pyogenes* and *S. canis* in humans and animals in Scotland in recent years. *Streptococcus pyogenes* is considered a strict human pathogen and it was estimated to be the fifth pathogen with the highest yearly fatality rate worldwide in 2010 (Beaton et al., 2020). *Streptococcus pyogenes* is highly adapted to colonise the upper respiratory tract and skin of humans and a considerable proportion of colonised individuals are asymptomatic (Kaplan, 1980). While invasive *S. pyogenes* infections are not common, they are accompanied by severe clinical manifestations and are associated with a mortality rate as high as 30% in high-income countries (Lamagni et al., 2008a). It should be noted that estimates on the global burden of *S. pyogenes* infections rely on data published in the early 2000s (Ralph and Carapetis,

2012). *Streptococcus canis*, conversely, is a streptococcal pathogen with a much broader host range, having been isolated from a variety of mammalian species (Richards et al., 2012; Numberger et al., 2021). Since dogs and cats appear to be the main hosts of this bacterial species (Fulde and Valentin-Weigand, 2012), *S. canis* is generally regarded as a pathogen of veterinary relevance, although human infections occur and can be severe (Lacave et al., 2016; Taniyama et al., 2017). Even in dogs, cats and other domesticated animals, *S. canis* appears to be a relatively uncommon cause of clinical disease compared to other bacterial pathogens (Wilson et al., 1997; Lamm et al., 2010), despite asymptomatic carriage being commonplace (Lysková et al., 2007b). The true burden of *S. canis* infection in the human and animal community, however, is currently unknown because the majority of medical and veterinary diagnostic laboratories characterise streptococcal isolates only on the basis of their Lancefield antigenic group. Since *S. canis* carries a group G Lancefield antigen, which is also present in strains of *S. dysgalactiae*, it is difficult to estimate accurate prevalence data for this bacterial species (Lam et al., 2007). The perceived low risk that *S. canis* poses to human and animal health has contributed to it remaining a neglected pathogen, which has been understudied and hence poorly-characterised. In the current project, we decided to focus on (a) *S. pyogenes*, due to its public health relevance and (b) on *S. canis*, due to the major knowledge gaps surrounding this species. They are, however, thematically related in that both pathogens are classified as pyogenic streptococci, they result in similar clinical manifestations and they are evolutionarily closely related (Vos et al., 2011).

Data segregation in public health hampers infectious disease control. Enhanced surveillance data were not readily available for iGAS due to the pressure that the COVID-19 pandemic exerted on the Scottish public health system, limiting the conclusions that could be drawn in chapter 3 and leaving questions open about potential epidemiological connections linking *emm*5.23 cases. The hypothetical scenario, outlined at the end of chapter 3, about a putatively undetected chain of transmission events, highlights the importance of facilitating data sharing between public health bodies. Linking isolate typing data and patient-centered enhanced surveillance data would allow the creation of early detection systems that could be used to initiate epidemiological investigations and contain disease outbreaks (Yang et al., 2011). Rapid data sharing between public health and research institutions

has been recognised as an important mechanism enabling a quick response to health emergencies, such as the COVID-19 pandemic (Schwalbe et al., 2020). From a practical standpoint, however, several obstacles stand in the way of a fast and free flow of data between public health organisations. For example, considerable time and resources, that could otherwise be used for data analysis, are required to implement data sharing (Morse, 2007; Lopez, 2010). Moreover, motivational reasons such as the concern of giving away a potential source of scientific credit and opportunities could make public health institutions hesitant to share data collected with their own resources (Pisani and AbouZahr, 2010). The most important barriers to rapid data sharing, however, are those regarding data ownership and privacy legislation (Van Panhuis et al., 2014), which in Europe is regulated by the General Data Protection Regulation (GDPR) (`https://gdpr-info.eu/`). An organisation collecting sensitive data about individuals and communities is also responsible for protecting their privacy. The necessity to anonymise sensitive data and make sure privacy is safeguarded puts additional pressure on public health organisations, making data sharing a delicate matter that requires careful consideration (Van Panhuis et al., 2014). In spite of all the valid reasons that make it difficult to achieve, the major positive outcomes anticipated to be associated with easier access to public health data from different parties would outweigh the risks encountered and the efforts required to facilitate data sharing (Van Panhuis et al., 2014).

Surveillance systems for streptococcal infections need to be improved, particularly in veterinary medicine. Surveillance systems are implemented to monitor the spread and characteristics of diseases in human and animal populations. In Scotland, a different approach to infectious disease surveillance stands out when human and animal healthcare systems are compared. Firstly, human pathogen surveillance, although it relies on the collaboration of different public health and research bodies, is regulated by a single overarching body, the Scottish NHS (`https://www.hps.scot.nhs.uk/web-resources-container/public-health-microbiology-strategy-for-scotland/`). Animal pathogen surveillance, conversely, is fragmented and handled by different organisations. For example, surveillance of zoonotic pathogens is carried out by the UKHSA (`https://www.gov.uk/government/publications`) and PHS

(https://www.hps.scot.nhs.uk/a-to-z-of-topics/zoonoses/), while the Animal and Plant Health Agency (APHA) manages scanning surveillance activities for companion animal, livestock and wildlife diseases (http://apha.defra.gov.uk/vet-gateway/surveillance/scanning/index.htm). APHA surveillance, in turn, relies on the collection of epidemiological data by several organisations and initiatives, such as the Small Animal Veterinary Surveillance Network (SAVSNET) (Radford et al., 2010), VetCompass (https://www.rvc.ac.uk/vetcompass/papers-and-data), the Scottish Rural College (SRUC) (https://www.sruc.ac.uk/veterinary-surveillance/) and the Scottish Government's Centre of Expertise on Animal Disease Outbreaks (EPIC) (Boden et al., 2020). Some of these organisations undertake surveillance by collecting data voluntarily submitted by veterinary practices and diagnostic laboratories. This form of surveillance is subject to several biases, including under-reporting due to the commercial nature of most veterinary diagnostic laboratories and the subsequent reticence to share privately owned data with the public. Another important difference between human and animal disease surveillance is their scope. The UKHSA describes itself as "responsible for protecting every member of every community from the impact of infectious diseases, chemical, biological, radiological and nuclear incidents and other health threats" (https://www.gov.uk/government/organisations/uk-health-security-agency). The APHA, conversely, aims "to safeguard animal and plant health for the benefit of people, the environment and the economy". Subsequently, animal disease surveillance is more targeted to pathogens that (a) affect production animals causing important economical losses (Stärk et al., 2006), (b) affect the horse racing industry (Slater et al., 2017) or (c) represent a zoonotic threat (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1055927/Zoonoses_annual_report_2021.pdf). In the case of *S. pyogenes*, nation-wide statutory surveillance is commonly undertaken in high-income countries for severe forms of disease such as invasive infections and, occasionally, scarlet fever (Efstratiou and Lamagni, 2016). Depending on the country, case metadata and patient information are also collected as part of enhanced surveillance schemes (Efstratiou and Lamagni, 2016). In Scotland,

where invasive infections are the only forms of *S. pyogenes* disease notifiable (https://www.legislation.gov.uk/asp/2008/5/schedules), invasive isolates, case metadata and patient information are collected by SMiRL and PHS (https://www.hps.scot.nhs.uk/a-to-z-of-topics/streptococcal-infections/group-a-streptococcal-infections/). The UKHSA manages *S. pyogenes* surveillance in England, where both invasive infections and scarlet fever cases need to be notified (https://www.gov.uk/government/collections/group-a-streptococcal-infections-guidance-and-data). Surveillance on animal streptococcal infections appears to be generally non-statutory, with an exception being the national surveillance system for *S. agalactiae* bovine mastitis undertaken by the Danish government (Churakov et al., 2021). In Scotland and the rest of the UK, surveillance on animal streptococcal disease is non-statutory and limited to pathogens relevant to the horse racing and food production industry, such as *S. equi*, *S. suis*, *S. agalactiae*, *S. dysgalactiae* and *S. uberis* (http://apha.defra.gov.uk/vet-gateway/surveillance; https://www.gov.uk/government/publications; https://www.hps.scot.nhs.uk). As *S. canis* is most commonly only implicated in companion animal infections, it is rarely considered in any form of surveillance. Both SAVSNET and VetCompass carry out surveillance and research on companion animal diseases using data voluntarily submitted by veterinary practices and diagnostic laboratories (Radford et al., 2010). To date, neither SAVSNET nor VetCompass have published epidemiological data on *S. canis* infections in companion animals in the UK. The lack of epidemiological data on *S. canis* disease is not only due to the scarcity and limitations of the surveillance systems in place, but also to the little attention this pathogen has received from the scientific community. Since research and surveillance demand effort and monetary investment, the prioritization of diseases with large-scale impacts on the economy and animal welfare, such as Foot and Mouth Disease and African Swine Fever, is rightfully applied. More effort should be made, however, to expand our knowledge surrounding the epidemiology of neglected infections such as those by *S. canis*. Voluntarily provided surveillance data like those collected by SAVSNET, for example, should be made more easily available to external researchers, facilitating the disclosure of data regarding non-priority diseases.

A disconnect between human and veterinary healthcare systems limits the application of the One Health vision. A section of chapter 4 discussed the potential for *S. canis* strains to colonise and cause disease in different host species. The results presented provide preliminary evidence that a variety of *S. canis* strains are circulating among different mammalian hosts, supporting the hypothesis that this bacterium can cause zoonotic infections. As such, *S. canis* can be regarded as a One Health pathogen. The concept of One Health was developed to identify any multidisciplinary effort that aims at improving human, animal and environmental health (Gibbs, 2014). A brief review of the published literature on the role of animal pathogens in human health revealed that the application of the One Health approach, although very topical, still faces many challenges. In particular, the implementation of surveillance and control of zoonotic diseases appears to be hindered by a disconnect between veterinary and human health systems (Bhatia, 2019). Practical recommendations to achieve a collaboration between human and veterinary medicine are not provided by the One Health agenda, which has also been criticised for being too broad and vague (Gibbs, 2014). Moreover, due to the high monetary investment they demand, One Health interventions are not generally prioritised by policy-makers and need to be supported by strong evidence of future economical gain in order to be considered (Rabinowitz et al., 2013). The lack of awareness, among the general public, about the One Health concept and a diffuse tendency to view global problems through an anthropocentric lens also impede a full attainment of the One Health view (Gibbs, 2014). If the main priority in human medicine is health protection, national veterinary services are often forced to prioritise economic benefit rather than animal welfare (Stärk et al., 2006). Subsequently, the societal role of veterinary medicine is generally viewed as marginal, making the One Health approach difficult to implement (Zinsstag et al., 2011). Although progress has been made in integrating veterinary sciences into public health (Zinsstag et al., 2009), considerable improvements are still needed (Zinsstag et al., 2011). In the case of bacterial infections, for example, surveillance systems for AMR in human and animal isolates are rarely integrated and harmonised (Silley et al., 2012). The acquisition of AMR by pathogenic bacteria is influenced by antimicrobial use (AMU) in both humans and animals, as well as by antimicrobial disposal in the environment. As such, AMR is regarded as a major One Health concern worldwide (Robinson et al., 2016). In Scotland, the SONAAR report published by

PHS provides an annual overview of AMU and AMR in human and veterinary medicine (`https://www.hps.scot.nhs.uk/web-resources-container`). The SONAAR report represents an important application of the One Health approach in AMU and AMR national surveillance, although it should be noted that the completeness and quality of the human medical dataset is greater than that of the veterinary medical one. Due to a lack of systematic surveillance on the use of drugs in veterinary medicine, SOONAR data on AMU and AMR in animals are limited to voluntary-based information submitted by veterinary practices to SAVSNET. An equalisation and harmonisation of human and veterinary healthcare systems appears necessary to maximise the benefits of the One Health approach. Emerging infectious diseases and AMR spread are both driven by a complex network of interactions that involve human societies, animal populations and environmental factors. Failing to recognise the importance that a One Health approach should have in the management of these global challenges could undermine any effort made to tackle them.

The importance of AMR surveillance is demonstrated by the early detection of *S. pyogenes* isolates with reduced susceptibility to penicillins. One of the reasons why improving AMR surveillance systems is a global priority is that they facilitate an early detection of new resistance mechanisms, as exemplified by the identification of *S. pyogenes* strains with reduced susceptibility to β-lactams by the USA CDC (Vannice et al., 2019). *Streptococcus pyogenes* has traditionally been considered a pathogen fully susceptible to β-lactams because, unlike *S. pneumoniae*, resistance towards this antibiotic class has never been developed and observed. *Streptococcus pyogenes*' inability to be naturally competent has been proposed as an explanation for the lack of β-lactam resistance in this bacterial species (Hanage and Shelburne III, 2020). In recent years, however, concerns over the future development of full resistance have emerged after the identification of single point mutations conferring reduced susceptibility to β-lactams in the *pbp* genes of certain *S. pyogenes* isolates (Vannice et al., 2019; Musser et al., 2020). Large-scale genomic studies on the occurrence of non-synonymous single point mutations in *S. pyogenes pbp* genes revealed that, although geographically spread, these changes are infrequent, occurring in around 2% of the isolates analysed (Hayes et al., 2020; Musser et al., 2020; Beres et al., 2022). Moreover, the majority of non-synonymous mutations detected in *pbp* genes appear to be subject

to negative selection, indicating that they are unlikely to successfully establish in the global *S. pyogenes* population (Beres et al., 2022). In the cohort of Scottish *emm*5.23 WGS analysed in chapter 3, no mutations were found compared to the *pbp* genes of strain Manfredo, which was isolated in the 1950s. The low frequency of occurrence, the signs of negative evolutionary selection and the fact that they are associated with reduced susceptibility but not full resistance suggest that the *pbp* genes single point mutations described are not currently a major public health concern, although they should remain object of close monitoring (Hanage and Shelburne III, 2020). Importantly, a recent study revealed the presence of chimeric *pbp* genes associated with reduced β-lactam susceptibility due to recombination events between *S. pyogenes* and group C and G streptococci (Beres et al., 2022). The authors of that work suggest that, should a resistant *pbp* allele emerge in a streptococcal species that can exchange genetic material with *S. pyogenes*, β-lactam resistance in GAS strains could be acquired by horizontal gene transfer (Beres et al., 2022). The evidence collected so far indicates that, although not a present threat, β-lactam resistance in *S. pyogenes* could occur in future, highlighting the value of AMR surveillance to monitor this potentially alarming scenario.

***Streptococcus pyogenes* and *S. canis* infections can be reduced by understanding their transmission routes.** The results obtained from the genomic investigation of invasive *S. pyogenes emm*5.23 strain in Scotland revealed that isolates with identical core genomes can cause disease in different people, even several months apart. In previous studies the mutation rate of *emm* types 1 and 89 was estimated to be 1.3 and 2.1 SNPs/year, respectively (Nasser et al., 2014; Turner et al., 2015). In the present work, it was impossible to estimate the *emm*5.23-specific mutation rate due to the short time-scale of the study and the low level of polymorphism measured. However, the very limited core genomic diversity detected among the majority of the *emm*5.23 isolates collected in a 2-year time span (2018-2019) appears to be in concordance with mutation rates reported in the literature. Based on phylogenetic tree topology and core genome similarity, 46 of the 54 *emm*5.23 isolates analysed appeared to originate from a relatively recent common ancestor (Pightling et al., 2018). This finding, together with the fact that near-identical isolates caused disease in different individuals several months apart, may indicate that asymptomatic

carriers or unappreciated environmental niches acted as prolonged and common sources of infection. *Streptococcus pyogenes* is traditionally considered a non-environmental pathogen, although biofilm formation that allows survival in the environment for up to four months has been demonstrated (Marks et al., 2014). Prolonged asymptomatic carriage of *S. pyogenes* has been demonstrated in a longitudinal study on school-aged children, showing that isolates of the same *emm* types can be found in the same individuals for up to 8.5 months (mean of 10.8 weeks) (Martin et al., 2004). Unfortunately, the lack of enhanced surveillance data for the *emm*5.23 cases included in this study prevents any insights into possible connections between the affected individuals, making any inference on transmission route purely a matter of speculation. In previous studies, it was hypothesised that the spread of unusual *emm* types could be due to the lack of herd immunity towards those particular strains (Southon et al., 2020). It should be remembered, however, that the first reported case of invasive *emm*5.23 infection in Scotland dated back to 2015, indicating that this strain had been circulating in the country for at least two years prior to the 2018-2019 outbreak. Regardless of how this strain managed to spread in the Scottish population, one can speculate whether the upsurge of invasive disease cases associated with this uncommon *emm* type could have been detected earlier, prompting a public health intervention to contain the outbreak as much as possible. Adopting genomic surveillance for iGAS disease, for example, could allow real-time detection of transmission events and promote targeted interventions (Turner et al., 2017). Other suggested ways to prevent the spread of *S. pyogenes* include increasing awareness about the importance of monitoring even mild infections, which could at some point turn into serious forms of disease (Hikone et al., 2015; Brennan and LeFevre, 2019), improving personal hygiene and ensuring cleanliness of shared facilities, particularly in care homes and hospitals (Avire et al., 2021). A vaccine against *S. pyogenes* is not currently available but many research groups worldwide are working towards this goal (Dale and Walker, 2020). Several vaccine candidates have been considered, including hypervariable and conserved regions of the M protein (Pastural et al., 2020), cell-wall carbohydrate (Van Sorge et al., 2014) and multicomponent formulations of secreted and cell-surface proteins (Rivera-Hernandez et al., 2019). A 30-valent M protein-based vaccine candidate has recently shown promising evidence of protection against the most common GAS strains circulating in North America and Europe, such as *emm*1, *emm*4, *emm*12, *emm*28 and *emm*89, which were consistently among

the ten most common strains associated with iGAS disease in Scotland from 2014 to 2019 (chapter 2) (Pastural et al., 2020). Concerns remain, however, for GAS infections in low and middle-income countries, where a higher *emm* type diversity is observed (Steer et al., 2009). A vaccine candidate including a modified version of the Group A carbohydrate, which is a component of the cell-wall of all strains, failed to induce protection against invasive infection in a mouse model (Rivera-Hernandez et al., 2016). A multicomponent vaccine known as 5CP provided preliminary evidence of protection against both local and systemic disease by targeting the sortase A, C5a peptidase, SpyAD, SpyCEP and SLO proteins (Bi et al., 2019). The genes encoding all five proteins included in 5CP were also found integrated in the genome of strain *emm*5.23 (chapter 3). When a vaccine is finally developed and released, additional measures to prevent GAS infection will include vaccination of at risk individuals and people working in close contact with them, such as carer and nurses. As for *S. canis*, little is known about the ecology of this bacterium and routes of infection. Pyogenic streptococci, such as *S. canis*, are traditionally considered strictly associated with homeothermic hosts (Mundt, 1982), despite recent studies pointing out the ability of some species to survive in the environment for prolonged periods of time (Marks et al., 2014; Jørgensen et al., 2016). Since *S. canis* has been isolated from the rectum of dogs and cats, ongoing deposition in the environment through faeces cannot be ruled out (Lysková et al., 2007b). Studies on the ability of *S. canis* to survive outside the host, however, have not yet been undertaken, thus the significance of indirect transmission of this pathogen is presently unknown. Direct routes of transmission are probably the most common and, based on the results described in chapter 4, inter-species transmission is very likely. Dogs have historically been considered the main host of *S. canis* (Devriese et al., 1986), although carriage of Group G streptococci is not uncommon in asymptomatic people (Haidan et al., 2000; Halperin et al., 2016). To date, however, data on human carriage of *S. canis* have not been published. One may speculate that *S. canis* transmission to humans might occur as a consequence of direct interactions with other humans and not just from companion animals alone. Although the disease burden attributable to *S. canis* in humans appears to be low (Galpérine et al., 2007), infections should be avoided as much as possible particularly by the elderly, immunocompromised and those with pre-existing medical conditions. Prevention measures for *S. canis* infection in people include protecting wounds and other skin injuries from physical contact with pets and, when

suffering from comorbidities, reducing interactions with companion animals. Given the high prevalence of isolation of *S. canis* from the oral cavity of dogs and cats and the reported transmission of this pathogen via animal bites (Takeda et al., 2001; Lysková et al., 2007b; Taniyama et al., 2017), the risk of severe invasive infections by *S. canis* following a dog bite may be reduced by promptly seeking medical assistance and receiving empirical antibiotic treatment.

Next-generation sequencing is opening the doors to a new world, where it's easy to get lost. The use of high-throughput sequencing technologies in the field of microbiology has revolutionised the way in which pathogens are investigated and understood (Schürch et al., 2018). The generation of whole genome sequence data, which was once prohibitively expensive for routinely use (more than £400 per genome), has now become an integral part of research and surveillance due to advancing technology and an associated reduction in cost (ranging from £100 to £40 per genome) (Harris et al., 2013; Quainoo et al., 2017; Schwarze et al., 2018). Since the advent of the WGS revolution, we have been able to monitor the acquisition of point mutations and MGE that influence microbial virulence, fitness and AMR phenotypes (Nasser et al., 2014; Beres et al., 2016). Additionally, we can inspect the evolutionary history of pathogen populations in order to predict their future changes (Crestani et al., 2021). Through the use of WGS data it is also possible to establish molecular links between cases of disease, revealing unknown transmission pathways and infection sources (Chalker et al., 2016). The applications of WGS data analysis are innumerable and benefit both the scientific research and public health communities. This valuable resource, however, is not without its complications. As any powerful engine requires an experienced pilot behind the wheel, the effective analysis of WGS data relies on the abilities of expert bioinformaticians. The majority of computer programs currently available for WGS data analysis function only through the use of the command line, a type of interface unfamiliar to most people with a medical or biological education. Command line software is often difficult to download, install and execute by those with non-specialised informatics skills. Program errors are commonly encountered and virtually impossible to fix for someone without programming experience. Program settings are often non-intuitive and, if used incorrectly, can easily lead to inaccurate output. Even when accurate results are produced,

they are sometimes difficult to interpret. Problems can arise at any step of the data analysis process, making the workflow extremely intricate and delicate. Undertaking this type of analysis may be relatively easy for someone with a background in computer science or a related subject, but those without this educational background need to invest time and sometimes money to acquire this specialised skill set. The access to appropriate training resources is another important barrier to the use of WGS data. While many universities are now offering degrees and diplomas in bioinformatics, high tuition fees and time constraints can be serious issues to graduates who have already heavily invested in prior education and may now be in full-time employment. Online resources such as forums, tutorials and courses are available but these are often incomplete and difficult to fully understand by non-specialists. Due to the incredibly high volume of WGS data produced routinely by public health and research institutions and the need to analyse those data, the involvement of non-bioinformaticians in this sector is expected to grow and thought should be given to how the current shortfall should be addressed. A survey conducted by the ECDC and one conducted by the European Food Safety Authority (EFSA) revealed that the lack of bioinformatics expertise is one of the main barriers limiting the use of WGS analysis in European public health organisations (Revez et al., 2017; García Fierro et al., 2018). A possible way to acquire bioinformatic skills without a formal training, for example, would be through mentoring by experts in the field. Expanding the role of bioinformaticians to incorporate teaching duties in order to train non-experienced data analysts should be considered by organisations that work with WGS data. This approach would allow a reduction in the burden of WGS analysis on bioinformaticians while also improving the performances of students and workers with different backgrounds.

The good, the bad and the ugly of data visualisation. The use of images to communicate important messages predates the beginning of recorded history and has been a crucial element in every human society for millennia (Knight Jr, 2012). As humans, after all, we have evolved to heavily rely on vision to explore and embrace the world (Geldard, 1953). Images instinctively attract our attention, are often easier to interpret than verbal communication and appear to lodge memories better than other stimuli. Some people even claim to process their thoughts via images rather than words (Luck et al., 2008). Given the central role of vision in our everyday life, it is unsurprising that images have become an essential means

of communication in all disciplines, including the various branches of science. Data visualisation in science is particularly important because it allows large and complex datasets to be summarised in a single figure. Instead of reading through pages filled with variables and measurements, we can produce a graph that communicates a complete story in an intuitive and enjoyable way. The array of different options available for visually summarising data is extremely wide and constantly expanding. From simple bar charts and dot plots to elaborate infographics and maps, the discipline of data visualisation offers methods to represent a multitude of datasets. It is striking, however, how little academic guidance is available on the topic of expressing data through figures. Studies on visual perception were able to identify some basic elements that should be considered when producing a figure (Cleveland and McGill, 1985; Chen et al., 2007), but this type of knowledge is rarely integrated in any educational curriculum, making data visualisation an often arbitrary practice that is influenced by authors' preferences. In the case of public health, for example, the use of figures is as important as it is arbitrary. Due to practical and privacy reasons, public health data are commonly handled by specialists and researchers in the field, who do not have formal training in graphic design. This constrains not only the graphical modalities available, but also the quality of implementation. For example, the use of programming languages, such as Python and R, for data manipulation and visualisation require advanced skills and experience that many public health specialists simply do not possess. By their very nature, certain images might be difficult to interpret even by specialised audiences, as demonstrated by the feedback collected on phylogenetic tree interpretation in chapter 5. The obstacles highlighted with regards to the use of data visualisation in public health are commonly neglected but should be addressed in order to improve data communication. More emphasis should be given to data visualisation in every scientific curriculum and more thoughtful discussions should be encouraged in the workplace to promote the adoption of best-practice in data visualisation.

Final thoughts and future directions. A multi-disciplinary approach is required to successfully tackle the complexity of infectious disease epidemiology. In the course of my PhD studies, I have shed new light on two bacterial infections of humans and animals from different angles, providing the following contributions to the field:

- The main characteristics of iGAS disease epidemiology in Scotland were presented

and described in chapter 2, with a particular emphasis on the effects of the COVID-19 pandemic.

- An upsurge of cases of iGAS infections in Scotland associated with the uncommon *emm* type 5.23 was investigated via genomic and transcriptomic analyses revealing a potentially undetected outbreak (chapter 3). The genome of strain *emm*5.23 had not been characterised before this study, adding additional value to the results produced.

- The first publicly available closed and complete genomes of *emm*5.23 isolates were produced as described in chapter 3 (Pagnossin et al., 2021). These high-quality genomes can be used as a reference for future bioinformatic analyses of *S. pyogenes* strains phylogenetically close to the Scottish *emm*5.23 isolates.

- A comprehensive and up to date literature review on *S. canis* was written (chapter 1) and published (Pagnossin et al., 2022). This is the first review centered on this topic to be published.

- The epidemiology of *S. canis* infections in different host species was explored using for the first time WGS data (chapter 4). The results produced allowed a characterisation of AMR, virulence characteristics and population structure of *S. canis* strains, offering strong evidence of the zoonotic potential of this bacterial species.

- Different visualisation techniques used to communicate epidemiological data on iGAS disease in Scotland were explored and optimised in chapter 5. Although other studies on the application of complex visualisation tools in a public health setting have been conducted before (Park et al., 2022), to my knowledge this is the first time the focus has been given to the use of basic graphical elements.

As commonly occurs in research, this work has generated more questions than answers. Each of my thesis chapters should not be seen as the final stop in a journey but as a transit station on a ride with unknown destination. Investigation into the epidemiology of iGAS infections in Scotland should in future be complemented with the availability of enhanced surveillance data that will allow more incisive analytical approaches to be used. This may answer some of the questions raised in chapter 2, such as why the iGAS incidence is particularly high in people aged between 30 and 39 in Scotland and why *emm* types 1, 4 and 12

were drastically affected by the COVID-19 pandemic. Enhanced surveillance data should also be used to investigate transmission events of *emm*5.23 in 2018 and 2019, to help clarify whether cases of *emm*5.23 infections were indeed connected. Transcriptomic analyses of further isolates could also increase our understanding of the *emm*5.23 disease outbreak described in chapter 3. Sequencing additional genomes of *S. canis* isolated from different species would facilitate a medium to large-scale genomic study on *S. canis* evolution and host adaptation, overcoming the limitations of the work presented in chapter 4. Virulence assays on *S. canis* strains carrying different virulence genes could also be undertaken in order to acquire further insight into *S. canis* disease pathogenesis. The work described in chapter 5 could be expanded by including more visualisation types and by recruiting a larger number of participants. Finally, assessing interpretation performance rather than visual preference would be an extremely valuable addition to the work of chapter 5.

As the global population grows, the challenges posed by infectious diseases are expected to increase. Pathogen spread will be facilitated not only by a larger number of susceptible human hosts, but also by resource competition and poverty (Bloom and Cadarette, 2019). Population growth, which is associated with the expansion of urban areas into wildlife habitats, will also facilitate human-wild animal interactions, with potential emergence and spread of zoonotic diseases (Bloom and Cadarette, 2019). More antimicrobial use will be required to treat both human and livestock infections, increasing the chance of AMR acquisition and spread in pathogen populations (Bloom and Cadarette, 2019; Schar et al., 2020; Tiseo et al., 2020). As human societies become more complex, so does the epidemiology of infectious diseases. Any effort made to limit the impact of such multi-faceted problem will require multi-disciplinary solutions. It is then important for experts in this field to find the right balance between specialisation and general understanding of the several disciplines used to investigate and manage infectious diseases. This PhD project was designed and conducted with this principle in mind. The epidemiology of *S. pyogenes* and *S. canis* infections in humans and animals in Scotland was researched using multiple methods ranging from descriptive, molecular and genomic epidemiology techniques to data visualisation and communication. The value of this work resides not only in the original results presented but also in the comprehensive approach that was adopted to produce them.

Appendix A

Supporting information Chapter 2

A.1 Tables and figures

Table A.1: Count of different strains (classified as *emm* types) isolated from normally sterile body sites in Scotland from 2014 to 2021. Two isolates of the same *emm* type coming from the same patient are considered one single strain, while two isolates of different *emm* types from the same patient are counted as two different strains.

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
1.0	74	77	80	56	74	19	34	0	414
89.0	25	31	31	19	34	21	12	4	177
12.0	17	26	23	12	25	17	14	1	135
4.0	9	11	26	12	18	14	7	0	97
76.0	40	44	2	1	0	1	2	1	91
28.0	12	10	13	10	16	15	7	3	86
83.13	0	0	0	0	22	34	24	3	83
3.93	10	2	0	3	51	12	1	0	79
75.0	10	4	16	7	15	6	4	0	62
5.23	0	1	0	0	21	27	8	0	57
3.1	22	15	9	2	4	0	0	0	52
87.0	2	4	4	1	20	15	3	1	50

Table A.1 continued from previous page

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
77.0	1	3	6	3	8	14	7	7	49
6.0	1	9	12	11	13	1	0	0	47
11.0	6	5	9	6	9	5	4	3	47
44.0	2	6	4	12	13	2	3	0	42
94.0	0	3	5	7	12	8	4	3	42
108.1	0	0	1	0	0	13	20	7	41
82.0	1	3	3	3	9	5	2	1	27
2.0	3	6	4	2	7	4	0	0	26
22.0	4	6	2	3	3	0	2	0	20
73.0	0	1	1	0	12	1	4	0	19
9.0	2	1	1	6	5	0	3	0	18
81.0	1	0	1	3	4	4	1	0	14
1.107	0	0	0	2	10	1	0	0	13
6.4	3	2	2	0	5	1	0	0	13
90.2	1	3	1	2	1	2	0	1	11
58.0	0	1	1	3	1	3	1	0	10
12.37	0	4	1	0	2	1	1	1	10
5.3	1	1	2	0	0	2	3	0	9
5.116	0	1	7	0	0	0	0	0	8
118.0	0	0	1	3	4	0	0	0	8
22.1	1	3	3	0	0	0	0	0	7
1.52	6	0	0	0	0	0	0	0	6
5.132	0	0	0	0	2	3	1	0	6
5.44	0	0	0	2	3	0	0	0	5
110.0	0	0	0	4	1	0	0	0	5
1.25	0	1	1	1	1	0	0	0	4
5.133	0	0	3	1	0	0	0	0	4
18.0	0	2	2	0	0	0	0	0	4

Table A.1 continued from previous page

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
85.0	0	0	0	4	0	0	0	0	4
102.2	0	0	0	0	2	1	0	1	4
3.153	0	0	0	0	0	3	0	0	3
3.6	1	0	0	0	0	2	0	0	3
8.0	1	1	0	0	1	0	0	0	3
5.16	0	0	0	0	0	2	1	0	3
5.6	0	0	0	0	1	1	0	1	3
1.19	1	1	0	0	0	0	0	0	2
3.4	0	1	0	0	0	1	0	0	2
4.13	0	1	0	0	1	0	0	0	2
5.165	0	0	0	0	1	1	0	0	2
5.166	0	0	0	0	1	1	0	0	2
6.1	0	1	0	1	0	0	0	0	2
27.0	1	0	0	0	1	0	0	0	2
28.5	1	1	0	0	0	0	0	0	2
50.0	1	0	0	0	1	0	0	0	2
58.7	0	0	0	2	0	0	0	0	2
86.2	0	0	2	0	0	0	0	0	2
103.0	1	0	0	1	0	0	0	0	2
169.3	0	0	1	0	1	0	0	0	2
240.3	0	0	0	0	2	0	0	0	2
12.8	0	0	1	0	0	0	1	0	2
66.1	0	0	0	1	0	0	0	1	2
75.1	0	0	0	0	1	0	1	0	2
108.8	0	0	0	0	0	1	1	0	2
6.125	0	0	0	0	0	0	0	2	2
11.29	0	0	0	0	0	0	0	2	2
102.3	0	0	0	0	0	0	1	1	2

Table A.1 continued from previous page

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
1.24	0	0	1	0	0	0	0	0	1
1.4	0	0	0	0	0	1	0	0	1
1.76	0	0	0	1	0	0	0	0	1
3.138	0	0	0	0	1	0	0	0	1
3.56	0	0	1	0	0	0	0	0	1
3.8	1	0	0	0	0	0	0	0	1
3.94	1	0	0	0	0	0	0	0	1
4.19	0	0	1	0	0	0	0	0	1
5.11	0	1	0	0	0	0	0	0	1
5.111	0	1	0	0	0	0	0	0	1
5.14	0	0	0	1	0	0	0	0	1
5.153	0	0	0	0	1	0	0	0	1
5.167	0	0	0	0	1	0	0	0	1
5.175	0	0	0	0	0	1	0	0	1
5.18	0	0	0	0	0	1	0	0	1
5.5	1	0	0	0	0	0	0	0	1
6.11	0	0	0	0	1	0	0	0	1
8.3	1	0	0	0	0	0	0	0	1
11.1	0	0	0	1	0	0	0	0	1
12.29	0	1	0	0	0	0	0	0	1
14.3	0	1	0	0	0	0	0	0	1
18.39	0	0	1	0	0	0	0	0	1
18.7	0	0	0	1	0	0	0	0	1
27.6	0	0	0	1	0	0	0	0	1
28.14	0	0	0	0	0	1	0	0	1
29.3	0	0	0	0	0	1	0	0	1
31.7	0	1	0	0	0	0	0	0	1
48.1	0	0	0	0	0	1	0	0	1

Table A.1 continued from previous page

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
58.1	0	0	0	0	0	1	0	0	1
66.0	0	0	0	1	0	0	0	0	1
71.0	0	0	0	0	1	0	0	0	1
75.3	1	0	0	0	0	0	0	0	1
80.0	1	0	0	0	0	0	0	0	1
81.1	0	0	1	0	0	0	0	0	1
81.5	0	0	0	0	0	1	0	0	1
84.0	0	1	0	0	0	0	0	0	1
90.5	0	0	0	0	1	0	0	0	1
103.3	0	1	0	0	0	0	0	0	1
106.0	0	0	1	0	0	0	0	0	1
106.6	0	0	0	0	0	1	0	0	1
122.0	0	0	0	1	0	0	0	0	1
148.3	0	1	0	0	0	0	0	0	1
148.4	0	0	0	0	0	1	0	0	1
183.1	0	0	0	0	1	0	0	0	1
225.0	0	0	0	0	1	0	0	0	1
281.4	0	0	0	1	0	0	0	0	1
652.0	0	0	0	0	0	1	0	0	1
5.188	0	0	0	0	0	0	1	0	1
6.122	0	0	0	0	0	0	1	0	1
12.116	0	0	0	0	0	0	1	0	1
18.29	0	0	0	0	0	0	1	0	1
31.8	0	0	0	0	0	0	1	0	1
68.3	0	0	0	0	0	0	1	0	1
108.2	0	0	0	0	0	0	1	0	1
124.2	0	0	0	0	0	0	0	1	1
149.1	0	0	0	0	0	0	1	0	1

Table A.1 continued from previous page

emm types	2014	2015	2016	2017	2018	2019	2020	2021	Total
205.2	0	0	0	0	0	0	0	1	1
Total	267	300	287	214	446	275	185	46	2020

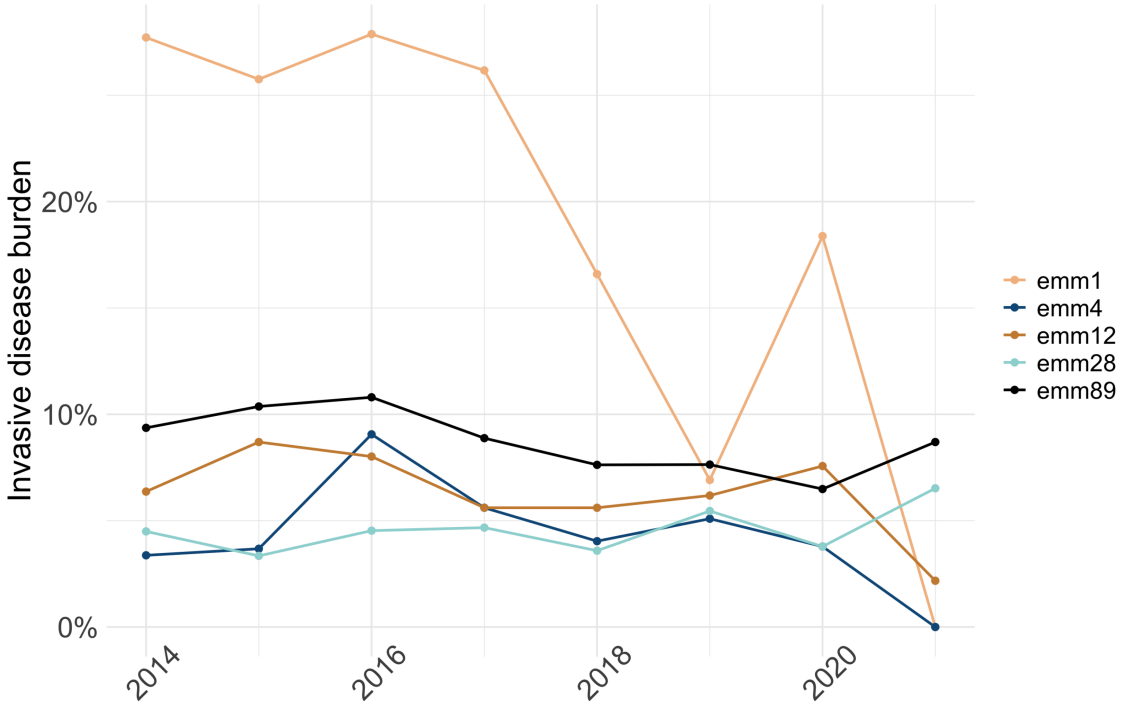


Figure A.1: Frequency of isolation from invasive disease cases of the 5 predominant Group A *Streptococcus* strains in Scotland from 2014 to 2021.

A.2 HSC-PBPP approval letter


NHS
SCOTLAND

Appendix B

Supporting information Chapter 3

B.1 Tables and figures

Table B.1: Metadata of all the *emm*5.23 isolates and relative whole genome sequences included in this study. The colour red is used to highlight the sequences that were excluded from bioinformatic analysis due to low assembly quality and signs of contamination.

ID	Year of isolation	Origin	Site of specimen	Sequencing technology	Assembly method	Genome size	Genome coverage	GC(%)
S.000	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859201	195x	38.42
S.134	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1861813	31x	38.41
S.138	2019	Scotland	Pulmonary tissue	Illumina MiSeq	SPAdes v. 3.11.1	1857380	50x	38.41
S.150	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858626	28x	38.41
S.152	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858467	26x	38.41
S.158	2019	Scotland	Lung	Illumina MiSeq	SPAdes v. 3.11.1	1867764	35x	38.43
S.166	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1866168	22x	38.42
S.169	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857447	72x	38.41
S.179	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1856659	72x	38.41
S.180	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857376	55x	38.41
S.183	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1899867	42x	38.42
S.188	2019	Scotland	Left ankle (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1855476	16x	38.41
S.254	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859365	190x	38.42
S.260	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857859	174x	38.42
S.266	2020	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858306	138x	38.42
S.268	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	2189974	111x	37.64
S.270	2020	Scotland	Ear (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1859278	198x	38.42
S.273	2019	Scotland	Leg (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1875531	44x	38.42
S.28	2020	Scotland	Pleural fluid	Illumina MiSeq	SPAdes v. 3.11.1	1867970	142x	38.42
S.293	2020	Scotland	Ear (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1863423	48x	38.42
S.310	2020	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1893903	45x	38.41

Table B.1 continued from previous page

ID	Year of isolation	Origin	Site of specimen	Sequencing technology	Assembly method	Genome size	Genome coverage	GC(%)
S.314	2020	Scotland	Wound swab (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1876614	54x	38.42
S.315	2020	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1861980	33x	38.41
S.352	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1860325	160x	38.42
S.353	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858697	98x	38.42
S.354	2018	Scotland	Sputum	Illumina MiSeq	SPAdes v. 3.11.1	1858524	198x	38.42
S.358	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857546	229x	38.42
S.359	2018	Scotland	Leg (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1856510	183x	38.42
S.361	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1877977	109x	38.4
S.363	2018	Scotland	Aspirate (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	2015400	203x	39.89
S.367	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859948	219x	38.42
S.368	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858290	260x	38.42
S.370	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857779	194x	38.42
S.371	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859735	131x	38.42
S.376	2018	Scotland	Parastomal hernia (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	1857504	205x	38.42
S.378	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858153	291x	38.42
S.380	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858916	147x	38.42
S.384	2019	Scotland	Throat	Illumina MiSeq	SPAdes v. 3.11.1	1860583	152x	38.42
S.391	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1855278	157x	38.42
S.394	2019	Scotland	Pus (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	3308028	93x	34.95
S.395	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858444	160x	38.41
S.404	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1858589	169x	38.42

Table B.1 continued from previous page

ID	Year of isolation	Origin	Site of specimen	Sequencing technology	Assembly method	Genome size	Genome coverage	GC(%)
S.406	2019	Scotland	Pleural fluid	Illumina MiSeq	SPAdes v. 3.11.1	1861097	176x	38.42
S.408	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859344	100x	38.42
S.426	2015	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1861925	173x	38.42
S.439	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1856271	153x	38.41
S.448	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857116	317x	38.42
S.451	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1859830	117x	38.41
S.453	2018	Scotland	Nose (cutaneous)	Illumina MiSeq	SPAdes v. 3.11.1	3872213	59x	39.21
S.458	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857492	164x	38.42
S.467	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857015	174x	38.41
S.469	2018	Scotland	Mouth	Illumina MiSeq	SPAdes v. 3.11.1	1858209	80x	38.42
S.475	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1877898	141x	38.41
S.476	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1856868	216x	38.42
S.477	2018	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1856724	81x	38.41
S.487	2019	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1857712	281x	38.42
S.867	2020	Scotland	Sputum	Illumina MiSeq	SPAdes v. 3.11.1	3332849	56x	35.41
S.716	2022	Scotland	Blood culture	Illumina MiSeq	SPAdes v. 3.11.1	1853318	28x	38.42
E.837	2012	England		Illumina MiSeq	SPAdes v. 3.11.1	1888373	53x	38.46
E.1305	2013	England		Illumina HiSeq	SPAdes v. 3.11.1	1856330	30x	38.41
E.24945	2014	England		Illumina HiSeq	SPAdes v. 3.11.1	1855472	138x	38.4
E.27709	2014	England		Illumina HiSeq	SPAdes v. 3.11.1	1854136	107x	38.4
E.96984	2014	England		Illumina HiSeq	SPAdes v. 3.11.1	1855108	133x	38.4
E.115332	2015	England		Illumina HiSeq	SPAdes v. 3.11.1	1943769	130x	38.4
E.83606	2015	England		Illumina HiSeq	SPAdes v. 3.11.1	1856795	143x	38.4
E.91955	2015	England		Illumina HiSeq	SPAdes v. 3.11.1	1863147	35x	38.4
E.97017	2015	England		Illumina HiSeq	SPAdes v. 3.11.1	1856095	155x	38.4

Table B.1 continued from previous page

ID	Year of isolation	Origin	Site of specimen	Sequencing technology	Assembly method	Genome size	Genome coverage	GC(%)
E.97024	2015	England		Illumina HiSeq	SPAdes v. 3.11.1	1856391	194x	38.4
E.1123	2016	England		Illumina MiSeq	SPAdes v. 3.11.1	1856865	72x	38.42
E.851	2016	England		Illumina MiSeq	SPAdes v. 3.11.1	1886949	87x	38.47
E.855	2016	England		Illumina MiSeq	SPAdes v. 3.11.1	1891523	59x	38.44
E.834	2017	England		Illumina MiSeq	SPAdes v. 3.11.1	1888135	53x	38.45
E.841	2017	England		Illumina MiSeq	SPAdes v. 3.11.1	1878581	49x	38.45
E.1114	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1858745	36x	38.41
E.1124	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1933751	52x	38.41
E.1127	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1855636	122x	38.42
E.705171	2018	England		Illumina HiSeq	SPAdes v. 3.11.1	1869632	173x	38.4
E.891	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1900733	100x	38.51
E.893	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1918361	575x	38.56
E.897	2018	England		Illumina MiSeq	SPAdes v. 3.11.1	1897659	47x	38.49
E.900	2019	England		Illumina MiSeq	SPAdes v. 3.11.1	1977006	47x	38.59
E.903	2019	England		Illumina MiSeq	SPAdes v. 3.11.1	1945033	127x	38.69
E.904	2019	England		Illumina MiSeq	SPAdes v. 3.11.1	1928021	92x	38.55
E.910	2019	England		Illumina MiSeq	SPAdes v. 3.11.1	1937356	188x	38.62
E.896	2020	England		Illumina MiSeq	SPAdes v. 3.11.1	1919330	112x	38.55
E.913	2020	England		Illumina MiSeq	SPAdes v. 3.11.1	1886280	54x	38.45
E.915	2020	England		Illumina MiSeq	SPAdes v. 3.11.1	1932835	125x	38.57

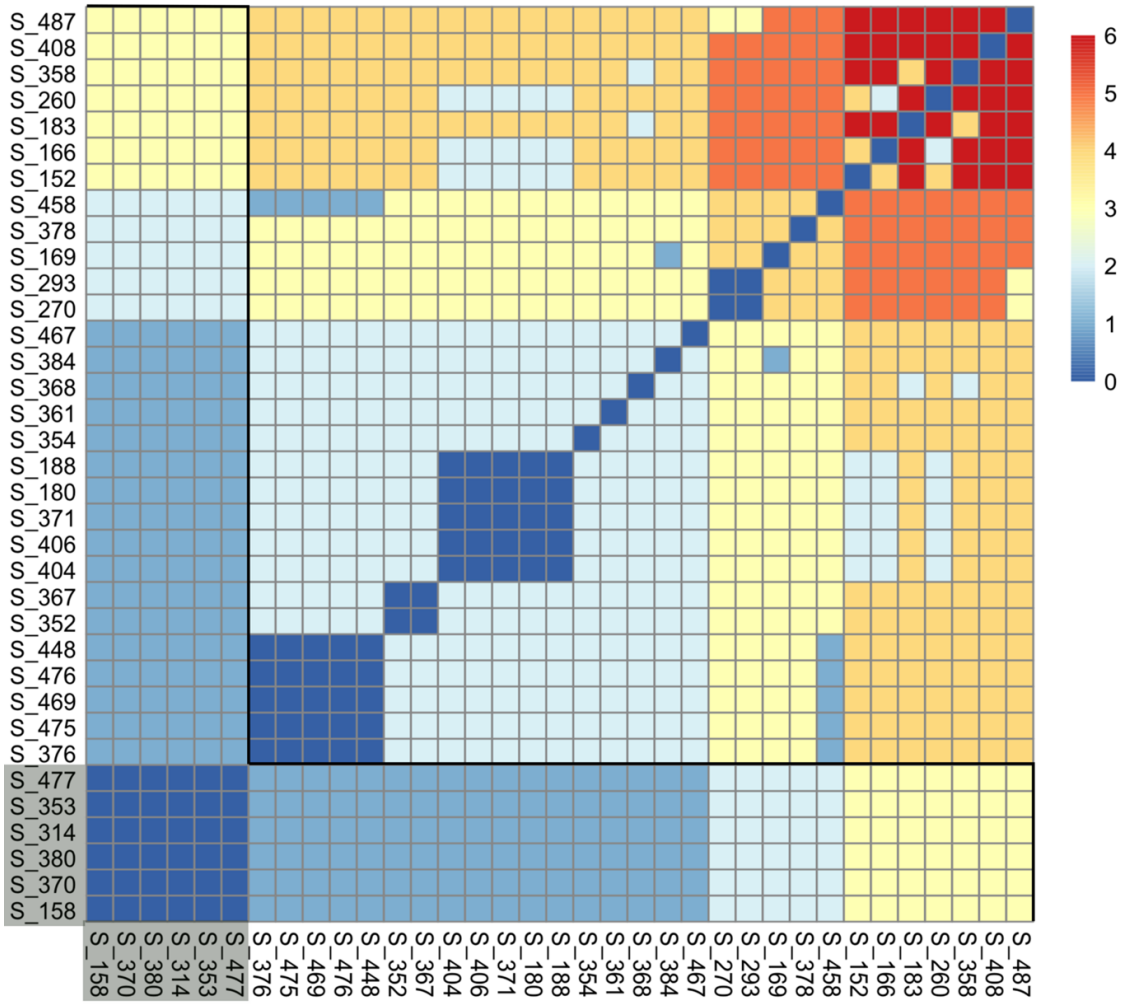


Figure B.1: Pairwise core single nucleotide polymorphism (SNP) distance of the *emm*5.23 isolates having no more than 3 SNPs of difference from the central cluster of the Scottish group (highlighted at the bottom left of the figure). A thick border is used to show the pairwise distance of all isolates compared to the central cluster genotype.

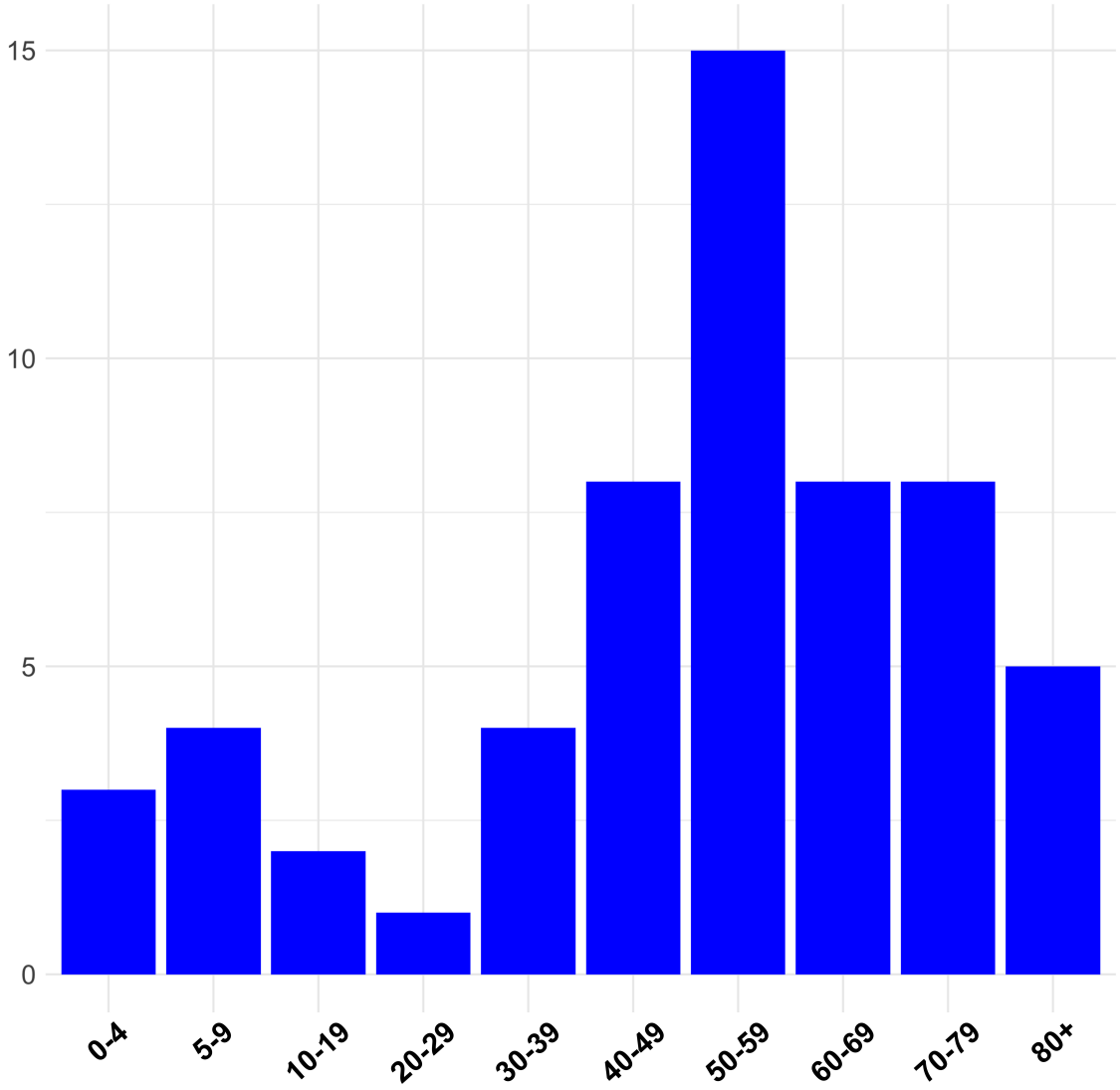


Figure B.2: Absolute frequency of invasive Group A *Streptococcus emm*5.23 cases in different age groups.

B.2 List of commands for bioinformatic analyses

B.2.1 Genome assembly pipeline for short paired-end reads

This pipeline performs the trimming, filtering and assembling of Illumina paired-end reads. Read trimming and filtering is carried out by two scripts that are part of the ConDeTri suite (Smeds and Künstner, 2011), while genome assembly is done using SPAdes (Bankevich et al., 2012).

```
#!/bin/sh

# Create a list with input files (which are zipped fastq files)
   for the loop. The list of files is made only with read 1 of
   each pair of reads. Later the sed command will couple read 1
   and read 2.

ls *1.fastq.gz > list

# Create the work folder (where the files will be processed) and
   the output folder (where the files will be stored).

mkdir Work_folder
mkdir Output
mkdir Output/Contigs
mkdir Output/Scaffolds
mkdir Output/Trim_reads

# This is the main loop of the script and it will move each couple
   of paired end reads in the Work_folder before running three
   different scripts (for trimming, filtering and assembling).

while read fast;
do
```

```
# The sed command is used to name the variables.

varzip1=$(echo $fast)
varzip2=$(echo $fast | sed s/_1.fas/_2.fas/)

# Unzipping the couple of files of interest (read 1 and read 2).

   gzip -d $varzip1
   gzip -d $varzip2

# Define 2 new variables for unzipped files. They are var1 = read
   1 and var2 = read 2.

var1=$(echo $varzip1 | sed s/_1.fastq.gz/_1.fastq/ )
var2=$(echo $varzip2 | sed s/_2.fastq.gz/_2.fastq/ )

# Move unzipped files into Work_folder and define new variable
   "pref" that will be used in the scripts. The prefix variable
   will be the ID of the sequence, without any extension.

mv $var1 Work_folder && mv $var2 Work_folder
cd Work_folder
pref=$(echo $var2 | sed s/_2.fastq/""/)

# First script for trimming the reads (ConDeTri).
echo ""
echo "======Starting ConDeTri for $pref..."
echo ""

perl $HOME/anaconda3/envs/assembly/bin/condetri.pl -fastq1=$var1
   -fastq2=$var2 -prefix=$pref -hq=25 -lq=10 -frac=0.8
   -minle\textit{n}=50 -mh=5 -ml=1 -sc=33
```

```
echo ""
echo "======ConDeTri finished for $pref!"
echo ""

# Remove useless files.

rm *stats ; rm *unpaired.fastq

# Define variables for second script.

trim1=$(ls *trim1*)
trim2=$(ls *trim2*)

# Second script for filtering PCR duplicates (FilterPCRdupl).

echo ""
echo "======Starting FilterPCRdupl for $pref..."
echo ""

perl $HOME/anaconda3/envs/assembly/bin/filterPCRdupl.pl
   -fastq1=$trim1 -fastq2=$trim2 -prefix=$pref -cmp=50

echo ""
echo "======FilterPCRdupl finished for $pref!"
echo ""

# Trimmed reads are zipped and moved to the Trim_reads folder.
   Useless files are removed.

rm *hist
gzip -r *trim*
mv *trim* ../Output/Trim_reads
```

```
# Third script for assembling the reads (SPAdes). Run it and
   remove the useless files.

echo ""
echo "======Starting SPAdes assembly of $pref..."
echo ""

python $HOME/anaconda3/envs/assembly/bin/spades.py -t 12 --careful
   --only-assembler -1 *uniq1.fastq -2 *uniq2.fastq -o ./$pref

rm *uniq*

echo ""
echo "======SPAdes assembly of $pref is finished!"
echo ""

# Go into the folder created by the script, extract 2 files
   (scaffolds and contigs), remove the folder with all the other
   files and move the two files in the Output folder.

echo "======Compressing reads $pref and moving files..."
cd $pref
mv contigs.fasta $pref.fasta
mv $pref.fasta ../../Output/Contigs
mv scaffolds.fasta $pref.fasta
mv $pref.fasta ../../Output/Scaffolds && cd ../ && rm -r $pref

# rezip the input files and move them back to the original folder.

gzip -r *.fastq && mv ./* ../
cd ../
done < list
rm list
```

B.2.2 Oxford Nanopore-read basecalling and genome assembly

This is a list of commands used to convert Oxford Nanopore sequencing output to fastq files (basecalling), separate the resulting files according to the isolate they represent (demultiplexing) and assemble them. Guppy v 3.6.0 (Wick et al., 2019) was used for basecalling and demultiplexing. Genome assembly was performed with Uniclycler v 0.4.8 (Wick et al., 2017).

```
#MinION sequencing output in fast5 format is converted to fastq
   files
guppy_basecaller -r --input_path
   /var/lib/MinKNOW/data/nameofexperiment/
nameofsample/subfolder/fast5folder/ --save_path guppy_nameoflibrary
-c dna_r9.4.1_450bps_fast.cfg

#Long MinION reads in fastq format are demultiplexed
guppy_barcoder --input_path ./ --save_path
nameoflibrary_demultiplexed --barcode_kits SQK-RBK004
   --trim_barcodes

#Reads are merged and filtered
cat *.fastq > nameofsample_combined.fastq

filtlong --target_bases 500000000 nameofsample_combined.fastq.gz |
   gzip > nameofsample_high_quality.fastq.gz

#MinION and Illumina reads are merged to produce a hybrid assembly
unicycler -1 shortread_1.fastq.gz -2 shortread_2.fastq.gz -l
   nameofsample_high_quality.fastq.gz -o
   nameofsample_hybrid_assembly -t 12
```

B.2.3 MGE detection

The following bash script uses SRST2 (Inouye et al., 2014) to align Illumina paired-end reads in fastq format to a published database of GAS MGE, referred to as MGE50+.

```
#!/bin/bash

mkdir MGE-50+;
mkdir MGE-50+/genes;
mkdir MGE-50+/full;

for i in *_1.fastq.gz;

#Capture strain designation, GAS number
do j=`echo $i | cut -d "_" -f 1`;

echo "Starting SRST2 processing of reads for strain $j";

#Run SRST2
srst2 --input_pe $j"_1.fastq.gz" $j"_2.fastq.gz"
   --gene_max_mismatch 20 --min_coverage 90 --max_divergence 10
   --threads 52 --output $j --gene_db
   /Users/davide/databases/MGE-50+.fasta;

#Remove .pileup and .bam files
  rm *.pileup;
  rm *.sorted.bam;
  rm *.log;
  rm *.scores

  mv *_gene*txt MGE-50+/genes;
  mv *_full*txt MGE-50+/full;

done
```

B.2.4 MLST typing, virulence and AMR gene detection

In this section, all commands used to assign MLST types and identify virulence and AMR genes are reported. MLST identification was carried out using SRST2 v0.2.0 (Inouye et al., 2014), while ARIBA v3.1.0 (Hunt et al., 2017) was run to detect virulence and AMR genes.

```
#SRST2 aligns fastq sequences to a GAS MLST database
srst2 --output test --input_pe *.fastq.gz --mlst_db
   Streptococcus_pyogenes.fasta --
   mlst_definitions spyogenes.txt --mlst_delimiter '_' --threads 12

#ARIBA downloads VFDB (for virulence genes) or CARD (for AMR
   genes) database and prepares it for use
ariba getref vfdb_full out.vfdb

ariba prepareref -f out.vfdb.fa -m out.vfdb.tsv\\
   out.vfdb.prepareref

#ARIBA aligns paire-end reads to a gene database (in this case
   VFDB and CARD)
db_dir=/path/to/reference/database

samples=$(ls *1.fastq.gz)

for samp in $samples; do
samp2=${samp//1.fastq/2.fastq}
outdir=$(echo ${samp//.fastq/} | cut -d/ -f2)
ariba run --force $db_dir $samp $samp2 $outdir
done
```

B.2.5 Polymorphism detection and phylogenetic analysis

This section contains information on how to run Snippy v4.4.5 (Seemann, 2015) for polymorphism calling and core SNP alignment and IQ-TREE v2.1.4 (Nguyen et al., 2015) for phylogenetic tree construction. The script snippy-multi, part of the Snippy suite, requires the preparation of an input file to run. The input file needs to be in tab format and consists of three separate columns, the first being a list of sequence Ids and the other two the absolute paths to the files containing the forward and reverse reads for each sequence, respectively. Once the file is ready, it has to be converted into a script that will run Snippy on each sequence.

```
#Generate a script to run snippy-multi
snippy-multi input.tab --ref Reference.gbk --cpus 16 $>$ runme.sh

sh ./runme.sh

#Produce a core SNP alignment masking MGE regions
snippy-core --mask excluded.bed --ref ref.fa snippy1 snippy2
   snippy3 ...

#Construct a maximum-likelihood phylogenetic tree with IQ-TREE
   using the best-fit substitution model function
iqtree -s core.aln -m MFP -b 100
```

Appendix C

Supporting information Chapter 4

C.1 Tables and figures

Table C.1: Metadata for all the *Streptococcus canis* isolates included in this study.

ID	Host	Year	Origin	Source	Sequencing technology	Assembly method	Genome coverage	Genome size (bp)	Accession number
B.FSLZ3	Bovine	1999	USA	Milk	454 FLX; Illumina GA2	Newbler v. 1.1.03.24; Velvet v. April-2008	23x	2267856	ASM26830v2
B.NCTC12191	Bovine	1900/1988	UK	Not specified	PacBio RS	Not specified	100x	2084744	42912_C01
C.1085	Cat	2020	UK	Urine	Illumina MiSeq	SPAdes v. 3.11.1	104x	2138017	
C.2080	Cat	2003	UK	Lung	Illumina MiSeq	SPAdes v. 3.11.1	50x	2126595	
C.280	Cat	2011	UK	Vaginal swab	Illumina MiSeq	SPAdes v. 3.11.1	44x	2456517	
C.284	Cat	2020	UK	Vaginal swab	Illumina MiSeq	SPAdes v. 3.11.1	32x	2430554	
C.63	Cat	2020	UK	Abdominal fluid	Illumina MiSeq	SPAdes v. 3.11.1	156x	2248939	
C.70	Cat	2020	UK	Abdominal fluid	Illumina MiSeq	SPAdes v. 3.11.1	43x	2244311	
C.FU1	Cat	2017	Japan	Pus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.1	196x	2061753	ASM981169v1
C.FU53	Cat	2017	Japan	Nasal cavity	Illumina MiSeq	CLC Genomics Workbench v. 6.5.1	695x	1901156	ASM1072169v1
C.FU6	Cat	2017	Japan	Pus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.8	400x	2203489	ASM981171v1
D.1069	Dog	2020	UK	Abscess	Illumina MiSeq	SPAdes v. 3.11.1	123x	1944887	
D.1070	Dog	2020	UK	Wound	Illumina MiSeq	SPAdes v. 3.11.1	113x	1971917	
D.1071	Dog	2021	UK	Urine	Illumina MiSeq	SPAdes v. 3.11.1	145x	1982953	
D.1080	Dog	2021	UK	Urine	Illumina MiSeq	SPAdes v. 3.11.1	195x	2134716	
D.1084	Dog	2021	UK	Urine	Illumina MiSeq	SPAdes v. 3.11.1	58x	1972289	
D.1098	Dog	2020	UK	Joint fluid	Illumina MiSeq	SPAdes v. 3.11.1	110x	2003361	
D.1099	Dog	2021	UK	Urine	Illumina MiSeq	SPAdes v. 3.11.1	268x	1867957	
D.2060	Dog	2006	UK	Wound	Illumina MiSeq	SPAdes v. 3.11.1	42x	2176835	
D.2067	Dog	2005	UK	Joint fluid	Illumina MiSeq	SPAdes v. 3.11.1	88x	2046307	
D.2079	Dog	2006	UK	Wound	Illumina MiSeq	SPAdes v. 3.11.1	42x	2182156	
D.2088	Dog	2002	UK	Cutaneous disharge	Illumina MiSeq	SPAdes v. 3.11.1	166x	2135647	

Table C.1 continued from previous page

ID	Host	Year	Origin	Source	Sequencing technology	Assembly method	Genome coverage	Genome size (bp)	Accession number
D.2110	Dog	2006	UK	Wound	Illumina MiSeq	SPAdes v. 3.11.1	102x	2053832	
D.257	Dog	2020	UK	Abscess	Illumina MiSeq	SPAdes v. 3.11.1	22x	2186353	
D.262	Dog	2009	UK	Joint fluid	Illumina MiSeq	SPAdes v. 3.11.1	15x	2026182	
D.266	Dog	2007	UK	Intestinal content	Illumina MiSeq	SPAdes v. 3.11.1	109x	2186682	
D.33	Dog	2016	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	112x	2068191	
D.35	Dog	2018	UK	Aborted fetus	Illumina MiSeq	SPAdes v. 3.11.1	191x	2238031	
D.40	Dog	2019	UK	Cornea	Illumina MiSeq	SPAdes v. 3.11.1	119x	1972417	
D.54	Dog	2020	UK	Pin tract swab	Illumina MiSeq	SPAdes v. 3.11.1	210x	2074552	
D.56	Dog	2020	UK	Joint fluid	Illumina MiSeq	SPAdes v. 3.11.1	168x	2241102	
D.82	Dog	2020	UK	Nose/oropharynx	Illumina MiSeq	SPAdes v. 3.11.1	18x	1969067	
D.B700072	Dog	2017	UK	Corneal ulcer	PacBio RS	Not specified	50x	2106614	5645STDY7765767
D.FMV2238	Dog	2002	Portugal	Ear canal	Illumina MiSeq	Not specified	68x	2408692	2238-02
D.FU129	Dog	2017	Japan	Pus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.3	433x	2028761	ASM981173v1
D.FU149	Dog	2019	Japan	Blood	Illumina MiSeq	CLC Genomics Workbench v. 12.0	284x	2108133	ASM1276785v1
D.FU29	Dog	2017	Japan	Uterus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.4	457x	2029450	ASM1072157v1
D.FU93	Dog	2017	Japan	Pus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.5	373x	2042355	ASM1072181v1
D.FU97	Dog	2017	Japan	Pus	Illumina MiSeq	CLC Genomics Workbench v. 6.5.6	614x	2032437	ASM1072183v1
D.HL_100	Dog	2018	South Korea	Urine	Illumina MiSeq and MinION	Unicycler v. 0.4.8	170x	2178238	ASM973855v1
D.HL_77_1	Dog	2018	South Korea	Ear canal	Illumina MiSeq and MinION	Unicycler v. 0.4.8	1691.1x	2265274	ASM1099391v2

Table C.1 continued from previous page

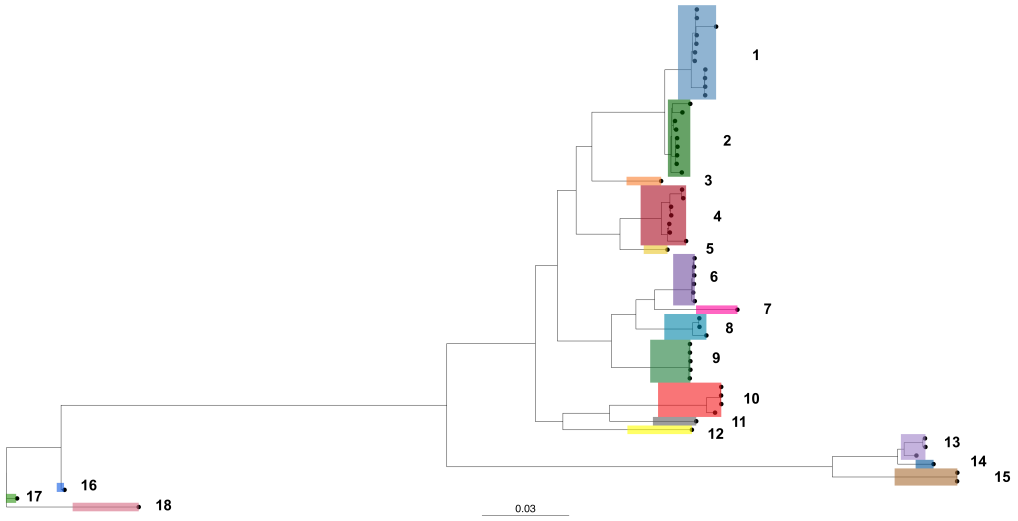
ID	Host	Year	Origin	Source	Sequencing technology	Assembly method	Genome coverage	Genome size (bp)	Accession number
D.HL_77_2	Dog	2018	South Korea	Ear canal	Illumina MiSeq and MinION	Unicycler v. 0.4.8	1222.8x	2157617	ASM1099384v2
D.HL_98_2	Dog	2018	South Korea	Nasal cavity	Illumina MiSeq and MinION	Unicycler v. 0.4.8	1808.8x	2176257	ASM1099386v2
H.1361	Human	2017	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	19x	2082116	
H.1384	Human	2012	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	30x	2016358	
H.1385	Human	2013	UK	Peripheral catheter	Illumina MiSeq	SPAdes v. 3.11.1	26x	2027329	
H.1386	Human	2019	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	26x	1951537	
H.1408	Human	2020	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	36x	2232105	
H.1414	Human	2019	UK	Leg swab	Illumina MiSeq	SPAdes v. 3.11.1	26x	1962353	
H.1419	Human	2019	UK	Foot swab	Illumina MiSeq	SPAdes v. 3.11.1	25x	1938228	
H.1434	Human	2018	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	39x	2252712	
H.1445	Human	2015	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	27x	2139594	
H.1462	Human	2019	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	49x	2014597	
H.1499	Human	2021	UK	Blood	Illumina MiSeq	SPAdes v. 3.11.1	81x	2130800	
H.G361	Human	2006	Germany	Vaginal swab	Illumina GAIIx	SPAdes v. June-2009	100x	2045931	ASM223493v1
H.OT1	Human	2012	Japan	Blood	Illumina MiSeq	CLC Genomics Workbench v. 6.5.2	427x	2030366	ASM654032v2
H.TA4	Human	2016	Japan	Blood	Illumina MiSeq	CLC Genomics Workbench v. 6.5.7	181x	2251539	ASM277055v1
NCTC6198	Animal	2018	UK	Not specified	PacBio RS	Not specified	100x	2143748	42197_B02
S.59	Seal	2011	UK	Lung	Illumina MiSeq	SPAdes v. 3.11.1	117x	2096801	

Table C.2: Minimum inhibitory concentration values (μg/mL) to a panel of antibiotics commonly used to treat Gram positive infections of the 39 *Streptococcus canis* isolates tested with broth microdilutions in this study. Red values represent phenotypic resistance according to the EUCAST breakpoint values v 12.0.

WGS id	Ampicillin	Amoxicillin	Clindamycin	Ceftriaxone	Cefotaxime	Doxycycline	Erythromycin	Levofloxacin	Meropenem	Moxifloxacin	Oxacillin	Penicillin G	Tetracycline	Vancomycin
D.2088	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
C.2080	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	>4	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	>4	≤1
D.2067	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.5	≤0.125	≤0.03125	≤0.5	≤1
D.2079	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.2060	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
D.2110	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.5	≤0.125	≤0.03125	=2	≤1
C.280	≤0.0625	≤0.125	>8	≤0.0625	≤0.0625	≤0.5	=4	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
S.59	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.33	≤0.0625	=0.5	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
D.35	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	>4	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	>4	≤1
D.40	≤0.0625	=0.25	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	=2	≤1
D.54	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.56	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	>4	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	>4	≤1
C.63	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	>4	≤0.125	=0.5	≤0.125	=0.25	≤0.125	=0.0625	>4	≤1
D.266	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.262	≤0.0625	≤0.125	≤0.125	=0.125	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=2	≤1
D.257	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.5	≤0.125	≤0.03125	≤0.5	≤1
D.82	≤0.0625	≤0.125	=0.5	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
C.284	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	=0.25	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
C.70	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
C.1085	=0.125	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.1069	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=2	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
D.1070	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.5	≤0.125	≤0.03125	=2	≤1
D.1098	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=2	≤0.125	=0.5	=0.5	≤0.03125	=1	≤1
D.1080	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.1099	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
D.1084	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=2	≤1
D.1071	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.5	≤0.125	≤0.03125	≤0.5	≤1
H.1384	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=0.5	≤0.125	=0.5	≤0.125	≤0.03125	=1	≤1
H.1385	≤0.0625	≤0.125	>8	≤0.0625	≤0.0625	>4	>8	=1	≤0.125	=0.25	≤0.125	=0.0625	>4	≤1
H.1445	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	=0.25	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
H.1361	≤0.0625	≤0.125	=0.5	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
H.1434	=0.125	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	=0.25	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
H.1386	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	=1	=0.25	=1	≤0.125	=0.5	=0.25	=0.0625	=4	≤1

Table C.2 continued from previous page

WGS id	Ampicillin	Amoxicillin	Clindamycin	Ceftriaxone	Cefotaxime	Doxycycline	Erythromycin	Levofloxacin	Meropenem	Moxifloxacin	Oxacillin	Penicillin G	Tetracycline	Vancomycin
H.1462	≤0.0625	≤0.125	>8	≤0.0625	≤0.0625	>4	>8	=1	≤0.125	=0.25	≤0.125	≤0.03125	>4	≤1
H.1419	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	=1	≤1
H.1414	≤0.0625	≤0.125	=0.25	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.25	≤0.125	≤0.03125	≤0.5	≤1
H.1408	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	>4	≤0.125	=0.5	≤0.125	=0.25	≤0.125	≤0.03125	>4	≤1
H.1499	≤0.0625	≤0.125	≤0.125	≤0.0625	≤0.0625	≤0.5	≤0.125	=1	≤0.125	=0.5	≤0.125	≤0.03125	=1	≤1



CGS cluster	Within-cluster pairwise SNP distance
1	Min:22 Max:1586 Avg:606
2	Min:120 Max:1238 Avg:579
4	Min:61 Max:1904 Avg:896
6	Min:73 Max:225 Avg:154
8	Min:47 Max:847 Avg:578
9	Min:57 Max:97 Avg:83
10	Min:69 Max:958 Avg:514
13	Min:196 Max:1368 Avg:972
15	Min:20 Max:20 Avg:20

CGS cluster	Between-cluster pairwise SNP distance
1	1624
2	5769
3	5924
4	3741
5	7690
6	4713
7	6059
8	6135
9	10791
10	7033
11	8786
12	21543
13	2191
14	8505
15	25676
16	2749
17	5470
18	

Pairwise SNP distance - whole tree

Min:20
Max:26610
Avg:10676

Figure C.1: *Streptococcus canis* maximum likelihood core single nucleotide polymorphism (SNP) phylogeny displaying the eighteen core genome SNP (CGS) clusters identified with TreeCluster using a threshold value of 0.017. For clusters composed of more than one isolate, a summary of the within-cluster pairwise SNP distance is provided. The pairwise SNP distance between the isolates closest to the tree root within each of two adjacent clusters is also reported as a measure of between-cluster distance.

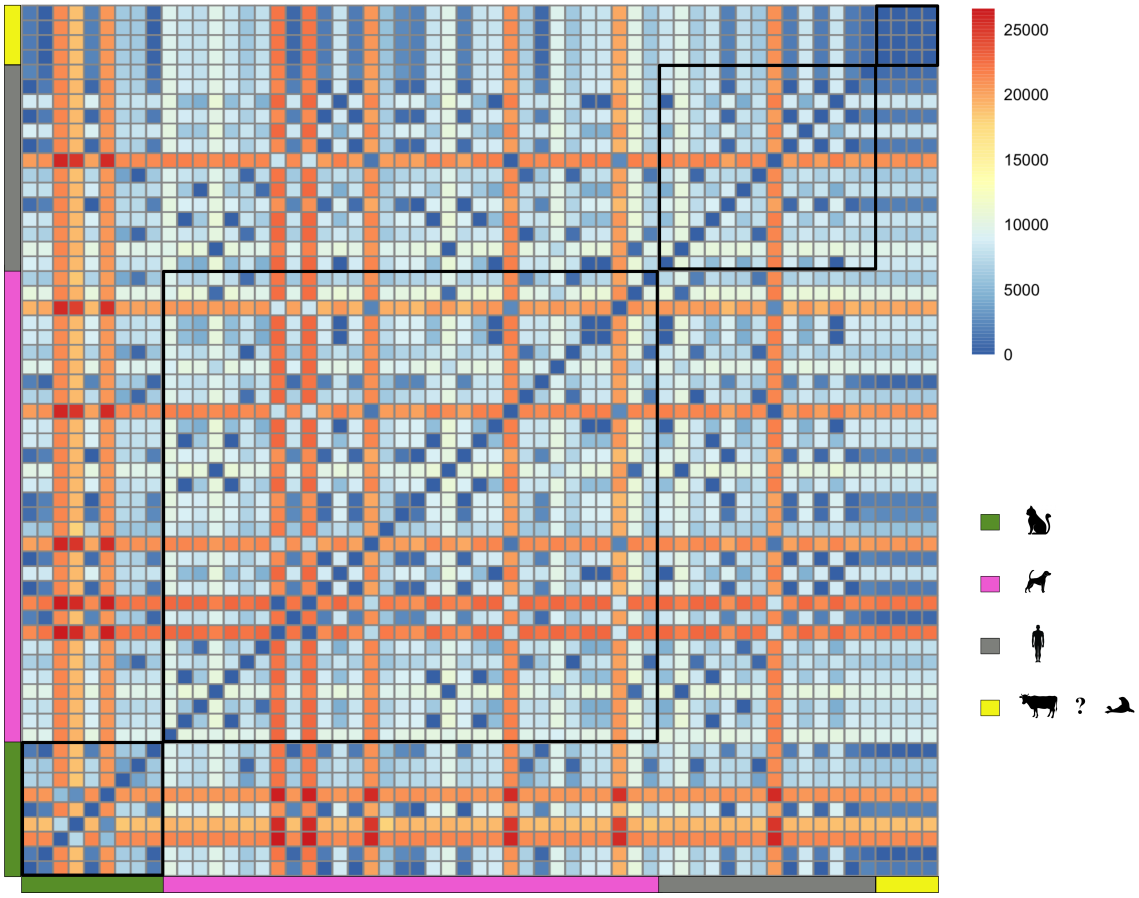


Figure C.2: Pairwise core single nucleotide polymorphism distances of the 59 *Streptococcus canis* whole genome sequences analyzed, arranged according to the host they were isolated from. Black borders highlight pairwise distances between isolates from the same hosts or category of hosts.

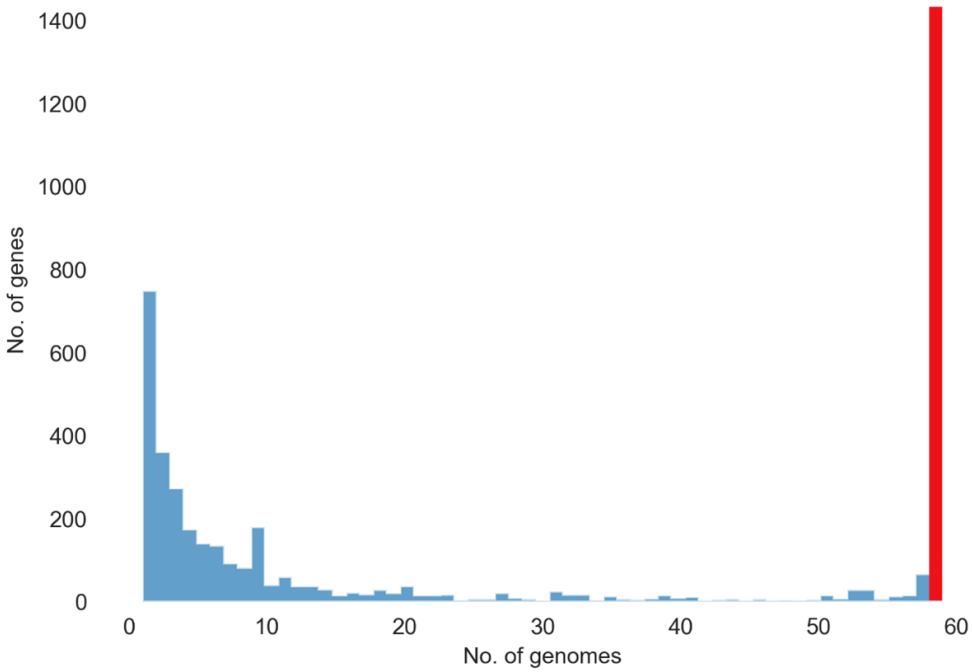


Figure C.3: Distribution of the 4426 *Streptococcus canis* genes detected across the whole genome sequences included in this study. Core genes are indicated in red.

C.2 List of commands for bioinformatic analyses

C.2.1 Local BLAST search

This script uses BLASTn (Camacho et al., 2009) to align fasta files (in this case assembled *S. canis* genomes) to a database of genes (in this case the VFDB full gene database, downloaded on the 06/10/2021).

```
#!/usr/local/bin/bash

mkdir output

for file in *.fasta;
do tag=${file%.fasta};
blastn -db ./db/VFDB_setB_nt.fas -query
   /Users/davide/Desktop/S.canis/virulence_factors/blast_new/"$tag".fasta
   -perc_identity 20 -outfmt 5 -out ./output/"$tag".xml
done
```

C.2.2 MLST, TreeCluster and AU test

The following commands were used to assign MLST types and CGS types with mlst (`https://github.com/tseemann/mlst`) and TreeCluster (Balaban et al., 2019), respectively. Below, the command used to perform the AU statistical test (Shimodaira, 2002) to compare tree topologies with IQTREE (Nguyen et al., 2015) is reported too.

```
#mlst assigns MLST types to WGS in fasta format
mlst *.fasta > mlst.csv

#TreeCluster finds clusters of isolates with less than 1000
   pairwise core SNP of difference in the phylogenetic tree
TreeCluster.py -i core.aln.treefile -t 0.017 -m max

#The unconstrained phylogenetic tree and the two constrained
```

```
   phylogenies are first concatenated into a single file and then
   tested using the AU test
cat core.aln.treefile core_MLST_constrained.treefile
   core_SCM_constrained.treefile > phylogenies.treels

iqtree -s core.aln -m TVM+F+ASC+R2 -z phylogenies.treels -n 0 -zb
   10000 -au
```

C.2.3 Pangenome analysis and accessory genome network

The commands reported in this section were used to annotate *S. canis* genome assemblies using Prokka (Seemann, 2014), generate a *S. canis* pangenome with Panaroo (Tonkin-Hill et al., 2020) and carry out a pan-GWAS analysis using Scoary (Brynildsrud et al., 2016).

```
#Prokka annotates assemblies
mkdir annotations
run_prokka -i *.fasta -o annotations

#Panaroo generates a pangenome using the annotated genomes
   produced by Prokka
mkdir results
panaroo -i *.gff -o results --clean-mode strict

#Scoary performs a pan-GWAS analysis looking for associations
   between genomic traits and host species
scoary -g gene_presence_absence_roary.csv -t
   hostgroup_membership.csv -p 1E-5 -c BH --no_pairwise
```

C.2.4 Accessory genome network

The commands shown below were used to create a network based on the presence/absence of accessory genes in the 59 *S. canis* WGS analysed.

```
#GraPPLE re-formats the gene_presence_absence_roary.csv produced
   by Panaroo
gene_matrix_to_binary.py -i gene_presence_absence_roary.csv -o
   output --start_col 15 --delimiter ,

#GraPPLE generate a network based on presence/absence of genes
python pw_similarity.py -i binary_presc_absc.tsv -o example1 -r
   "isolates" -s "jaccard" -f 0.715 -t 2

#Metadata (specifically concerning host species) are added to the
   network
metadata_to_layout.py -l acc_gene_dist_isols_pw_sim.layout -m
   gene_info.tsv -r "copy" -s headers.txt
```

Appendix D

Supporting information Chapter 5

D.1 R codes

In this section, all pieces of R code used to produce the visualizations assessed in Chapter 5 are reported.

D.1.1 Yearly incidence of iGAS disease

The following pieces of code were written to produce Figures 5.1A and B:

```
incidence<-read.csv("iGAS_incidence.csv")
A<-ggplot(incidence, aes(x = Year, y = Incidence, fill=as.factor(Year))) +
geom_histogram(stat = "identity") +
scale_x_continuous(breaks = c(2014,2015,2016,2017,2018,2019)) +
scale_y_continuous(limits = c(0, 8.5)) +
scale_fill_manual(values= c("royalblue2", "royalblue2", "royalblue2",
"royalblue2", "royalblue2", "royalblue2")) +
theme_bw()+
labs(x = "", y = "Cases per 100,000 people") +
theme(panel.grid = element_blank(),
legend.position = "none",
axis.text = element_text(size = 20),
axis.title = element_text(size = 20, face = "plain"),
```

```
plot.title = element_text(size = 20, hjust = 0.5, face = "bold"))


    B<-ggplot(incidence, aes(x=Year, y=Incidence,
group=Disease, color=Disease)) +
geom_line() +
geom_point() +
theme_minimal() +
ylab("Cases per 100.000 people") +
xlab(NULL) +
theme(axis.text = element_text(size = 20),
axis.title = element_text(size = 20, face = "plain"),
legend.text = element_text(size=20), legend.title =
element_text(size=20)) +
expand_limits(y=0)
```

D.1.2 Monthly incidence of iGAS disease

The following pieces of code were written to produce Figures 5.2A, B and C:

```
  months <- read.csv("date_coll.csv")
dates <- months$Date
i <- incidence(as.Date(dmy(dates)), interval = "1 month")
b <- make_breaks(i, n_breaks = 12, labels_week = FALSE)

A <- plot(i, color="navyblue", xlab = "",
ylab = "Confirmed new cases") +
scale_x_incidence(i, n_breaks = 12, labels_week = FALSE) +
scale_x_date(breaks=b$breaks, labels = date_format("%d %b %Y")) +
theme_bw() + theme(panel.background = element_rect(fill = "white"),
axis.text.x = element_text(angle = 45, hjust = 1, size = 18),
axis.text.y = element_text(size=18),
legend.title = element_blank(),
```

```
axis.title=element_text(size=20))

i <- incidence(as.Date(dmy(dates)), interval = "1 month", group = months$Year)
b <- make_breaks(i, n_breaks = 12, labels_week = FALSE)
pal <- wes_palette("Darjeeling1", 6, type = "continuous")

B <- plot(i, color= pal, xlab = "", ylab = "Confirmed new cases") +
scale_x_incidence(i, n_breaks = 12, labels_week = FALSE) +
scale_x_date(breaks=b$breaks,
labels = date_format("%d %b %Y")) +
theme(panel.background = element_rect(fill = "white"),
axis.text.x = element_text(angle = 45, hjust = 1, size = 18),
axis.text.y = element_text(size=18),
legend.title = element_blank(),
axis.title=element_text(size=20),
legend.text = element_text(size=16))

month_cases<-read.csv("month_cases.csv")
month_cases %>% mutate(Month = fct_relevel(Month,
"Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
month_cases$Month<- factor(month_cases$Month, levels=c
("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"))
month_cases$Year<-as.factor(month_cases$Year)

C <- ggplot(month_cases, aes(x=Month, y=Cases, group=Year, color=Year))+
geom_line() + geom_point() +
scale_color_manual(values= c("#FF0000","#32806E", "#91A737", "#F49C00",
"#D98F2A", "#5BBCD6")) +
theme_minimal() +
ylab("Confirmed new cases") +
xlab(NULL)+
```

```
theme(axis.text = element_text(size = 18),
axis.title = element_text(size=20),
legend.text = element_text(size=16),
legend.title = element_blank())
```

D.1.3 Incidence of iGAS disease in different age groups

The following pieces of code were written to produce Figures 5.3A, B and C:

```
agegroups<-read.csv("age_incidence.csv")

A <- ggplot(agegroups, aes(x=Age, y=Incidence)) +
geom_boxplot(aes(fill=Age)) + theme_bw() +
scale_fill_manual(values=wes_palette("Zissou1",
10, type = "continuous")) + labs(x="\n Age groups",
y= "Cases per 100.000 people \n") +
theme(axis.text = element_text(size = 18),
axis.title = element_text(size = 20, face = "plain"),
panel.grid = element_blank(),
plot.margin = unit(c(1,1,1,1), units = , "cm"),
legend.position = "none")

agegroups$Year<-as.factor(agegroups$Year)
p <- agegroups %>%
mutate(Age = fct_relevel(Age,
"0-4", "5-9", "10-19",
"20-29", "30-39", "40-49",
"50-59", "60-69", "70-79", "≥80"))
agegroups$Age<- factor(agegroups$Age,
levels = c("0-4", "5-9", "10-19", "20-29",
"30-39", "40-49", "50-59", "60-69", "70-79", "≥80"))
```

```
B <- ggplot(p, aes(x=Age, y=Incidence, fill=Year))+
geom_boxplot(fill="white", outlier.shape = NA, colour='gray48') +
geom_dotplot(binaxis='y', stackdir = 'center', binpositions = 'all',
stackgroups = TRUE, stackratio=1.5, dotsize = 0.6, binwidth=0.5) +
theme_classic() +
labs(x="\n Age groups", y= "Cases per 100.000 people \n") +
theme(axis.text = element_text(size = 18),
axis.title = element_text(size = 20, face = "plain"),
legend.key.size = unit(0.8, 'cm'),
legend.key.height = unit(0.8, 'cm'),
legend.key.width = unit(0.8, 'cm'),
legend.title = element_blank(),
legend.text = element_text(size=18))

C <- ggplot(p, aes(x=Age, y=Incidence)) +
geom_dotplot(binaxis='y', stackdir = 'center', binpositions = 'all',
stackgroups = TRUE, stackratio=1.5, dotsize = 0.6, binwidth=0.5) +
theme_bw() + labs(x="\n Age groups", y= "Cases per 100.000 people \n") +
theme(axis.text = element_text(size = 18),
axis.title = element_text(size = 20, face = "plain"),
legend.key.size = unit(0.8, 'cm'),
legend.key.height = unit(0.8, 'cm'),
legend.key.width = unit(0.8, 'cm'),
legend.title = element_text(size=12),
legend.text = element_text(size=18))
```

D.1.4 *Emm* type-specific invasive disease burden

The following pieces of code were written to produce Figures 5.4A, B and C:

```
  emm<-read.csv("emm_types.csv")
emm %>%
mutate(emm_type = fct_relevel(emm_type,
"emm1", "emm89", "emm12", "emm4", "emm76", "emm3.93",
"emm28", "emm75", "emm83.13", "emm3.1", "emm5.23",
"emm6", "emm87", "emm11", "emm44", "Other"))

emm$emm_type<- factor(emm$emm_type, levels=c ("emm1", "emm89", "emm12",
"emm4", "emm76", "emm3.93", "emm28", "emm75", "emm83.13", "emm3.1",
"emm5.23", "emm6", "emm87", "emm11", "emm44", "Other"))

emm$year<-as.factor(emm$year)

A <- ggplot(emm, aes(x=emm_type, y=cases, fill=year)) +
geom_bar(position="dodge", stat="identity") +
scale_fill_manual(values = wes_palette("Zissou1", 6, type = "continuous")) +
facet_grid(year~.) + theme_bw() +
theme(legend.position="none",
axis.text.x = element_text(angle = 45, hjust = 1, size = 18),
axis.text.y = element_text(size=16),
axis.title=element_text(size=20),
strip.text.y = element_text(size = 16)) +
labs(x = "", y = "Number of cases")

emm2<-read.csv("emm2.csv")
emm2 %>%
mutate(emm_type = fct_relevel(emm_type,
"emm1", "emm89", "emm12", "emm4", "emm76", "emm3.93",
"emm28", "emm75", "emm83.13", "emm3.1",
"emm5.23", "emm6", "emm87", "emm11", "emm44", "Other"))
```

```
emm2$emm_type<- factor(emm2$emm_type,
levels=c ("emm1", "emm89", "emm12","emm4",
"emm76", "emm3.93", "emm28", "emm75", "emm83.13",
"emm3.1", "emm5.23", "emm6", "emm87", "emm11", "emm44", "Other"))

emm2$year<-as.factor(emm2$year)

emmNew <- emm2 %>% group_by(year, emm_type) %>%
summarize(count = n()) %>%
mutate(pct = count/sum(count))

B <- ggplot(emmNew, aes(emm_type, pct)) +
geom_bar(stat = 'identity') +
scale_y_continuous(labels = scales::percent) +
facet_grid(year~.) + theme_classic() +
theme(legend.position="none",
axis.text.x = element_text(angle = 45, hjust = 1, size = 18),
axis.text.y = element_text(size=15),
axis.title=element_text(size=20),
strip.text.y = element_text(size = 16)) +
labs(x = "", y = "Relative frequency")

C <- ggplot(emmNew, aes(x=year, y=pct, fill=factor(emm_type))) +
geom_bar(stat = "identity", width = 0.7) +
geom_col(position = position_stack(reverse = TRUE)) +
scale_y_continuous(labels=scales::percent) +
scale_fill_manual(values=c("dodgerblue2", "#E31A1C",
"green4", "#6A3D9A", "#FF7F00", "black", "#FB9A99", "gold1",
"skyblue2", "palegreen2", "#CAB2D6", "#FDBF6F", "gray70",
"khaki2", "maroon", "orchid1")) +
labs(x = "", y = "Relative frequency", fill = "") +
```

```
guides(fill = guide_legend(reverse=TRUE)) +
theme_classic() + theme(panel.grid = element_blank(),
axis.text.x = element_text(angle = 45, hjust = 1, size = 18),
axis.text.y = element_text(size=16),
axis.title.y = element_text(size=20),
legend.title = element_text(size=14), legend.text = element_text(size=16))
```

D.1.5 Online survey

In the following pages a copy of the online survey used in this study is reported.

Assessment of different visualisation options to describe the epidemiology of invasive Group A Streptococcus (iGAS) disease in Scotland

In each of the following sections you will be presented with a few different figures that display one dataset. You will be asked a series of multiple choice and short open questions to appraise the available options. Your feedback will be later used to produce a new figure for each dataset and to draw conclusions regarding data visualisation in the public health sector.

*Required

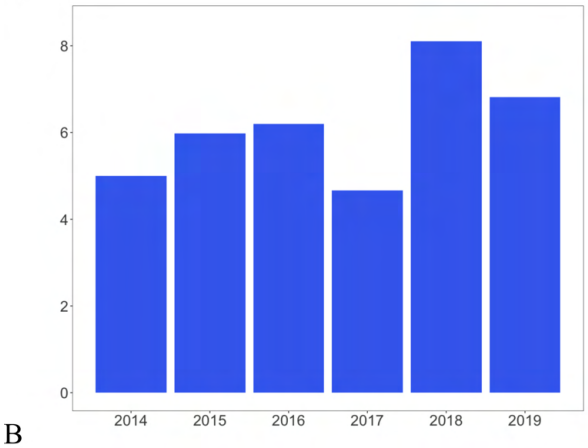
1. Before starting, we would like to know which one of the following disciplines better describes your day-to-day occupation *

Mark only one oval.

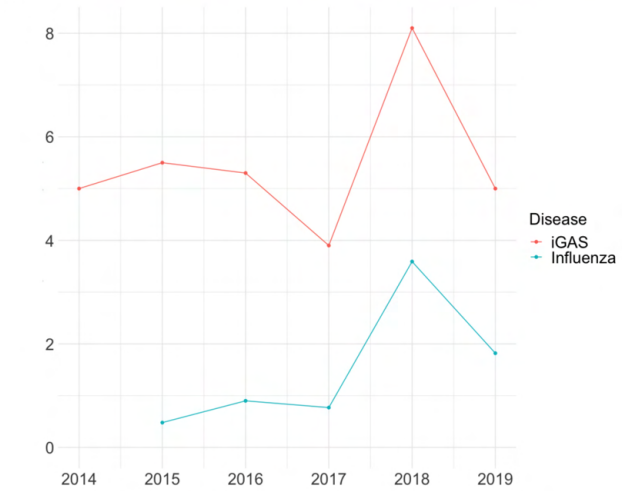
- ◯ Public health
- ◯ Clinical microbiology
- ◯ Microbiology laboratory
- ◯ Infectious disease
- ◯ Other: _____

Yearly incidence of iGAS disease in Scotland

A



B

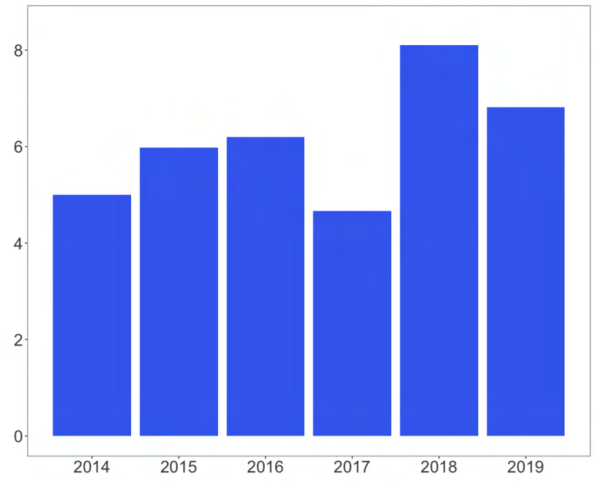


Yearly incidence of iGAS disease in Scotland

A (Bar chart) - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people.

B (Line chart) - Pink line - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people. Blue line –Mean weekly incidence of confirmed hospitalised cases of influenza expressed as cases per 100,000 people in the flu season from 2015 to 2019. Influenza data are collected by the USISS (UK Severe Influenza Surveillance Systems) sentinel scheme "which is a sentinel network of acute trusts in England who report weekly aggregate numbers on laboratory confirmed influenza hospital admissions at all levels of care".

Yearly incidence of iGAS disease in Scotland - Bar chart



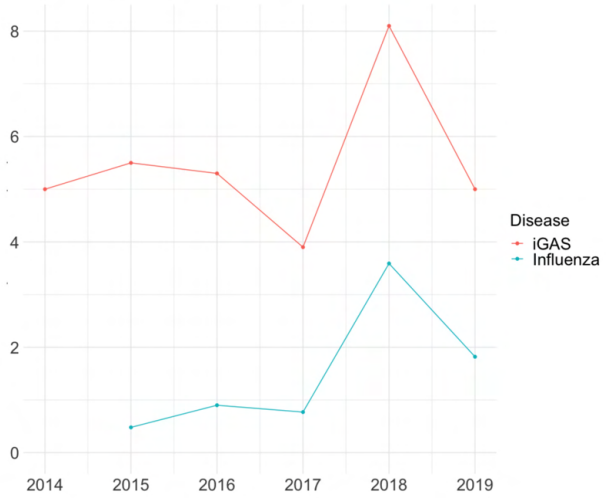
Bar chart
Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people.

2. This figure is purposefully plain (monochromic, blank background, only one variable displayed). Which one of the following statements better describes this figure from your point of view? *

Mark only one oval.

- ◯ Simple and effective
- ◯ Basic and unremarkable
- ◯ Other: _____

Yearly incidence of iGAS disease in Scotland - Line chart



Line chart
Pink line - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people. Blue line –Mean weekly incidence of confirmed hospitalised cases of influenza expressed as cases per 100,000 people in the flu season from 2015 to 2019. Influenza data are collected by the USISS (UK Severe Influenza Surveillance Systems) sentinel scheme "which is a sentinel network of acute trusts in England who report weekly aggregate numbers on laboratory confirmed influenza hospital admissions at all levels of care".

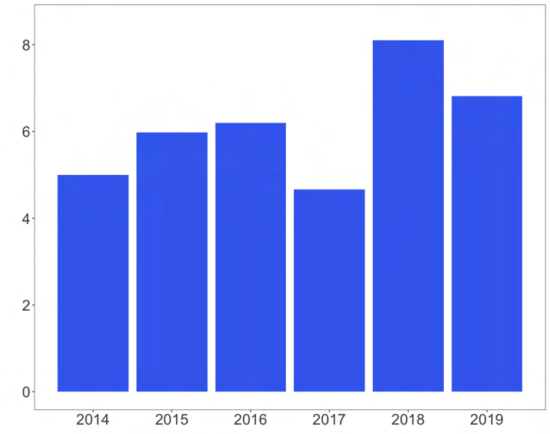
3. Incidence data regarding hospitalisation due to Influenza are displayed as a means of comparison for iGAS disease incidence. Do you find this choice helpful to contextualise iGAS epidemiological data or distracting? *

Mark only one oval.

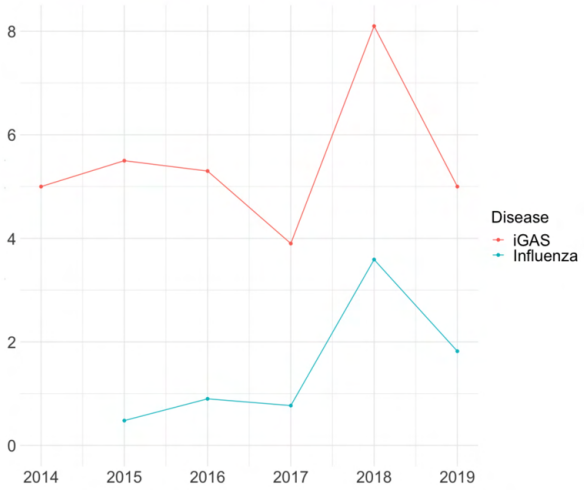
- ◯ Helpful
- ◯ Distracting
- ◯ Other: _____

Yearly incidence of iGAS disease in Scotland - Figures combined

A



B



Yearly incidence of iGAS disease in Scotland

A - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100.000 people.

B - Pink line - Yearly incidence of iGAS disease in the Scottish population from 2014 to 2019, expressed as cases per 100,000 people. Blue line –Mean weekly incidence of confirmed hospitalised cases of influenza expressed as cases per 100,000 people in the flu season from 2015 to 2019. Influenza data are collected by the USISS (UK Severe Influenza Surveillance Systems) sentinel scheme "which is a sentinel network of acute trusts in England who report weekly aggregate numbers on laboratory confirmed influenza hospital admissions at all levels of care".

4. For the bar chart, name one change that could improve it *

_____

_____

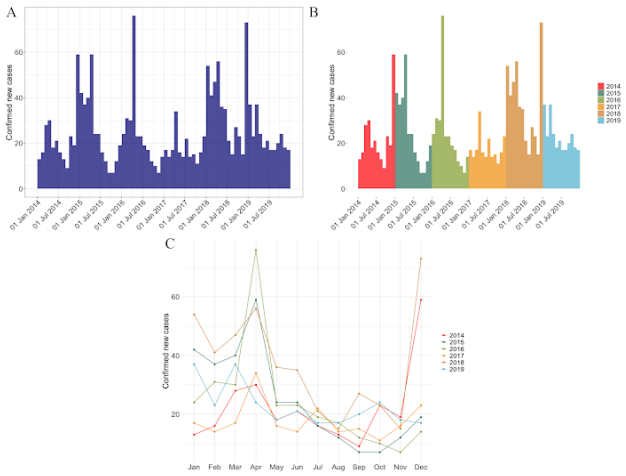
_____

_____

_____

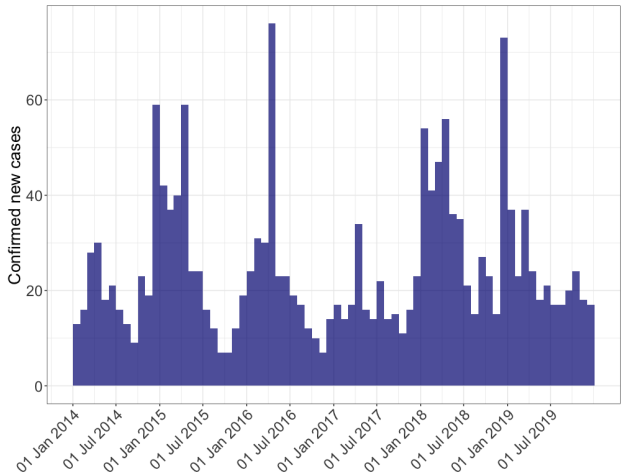
5. For the line chart, name one change that could improve it *

Monthly burden of iGAS disease



Monthly burden of iGAS disease
A (Monochromic histogram), B (Coloured histogram), C (Line chart) - Monthly burden of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

Monthly burden of iGAS disease - Monochromic histogram



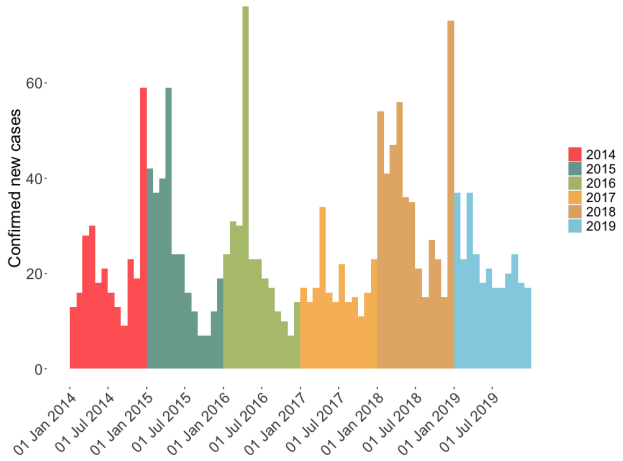
Monochromic histogram
Monthly burden of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

6. Is the background grid a useful or distractive element in this figure? *

Mark only one oval.

- ◯ Useful
- ◯ Distractive
- ◯ Other: _____

Monthly burden of iGAS disease - Coloured histogram



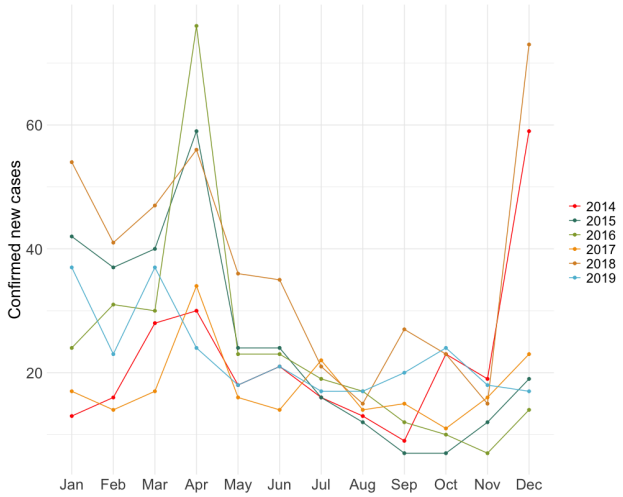
Coloured histogram
Monthly burden of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

7. Compared to the monochromic histogram, colours were used here to mark the passing of time. Would you consider this an improvement from the monochromic version? *

Mark only one oval.

- ◯ Yes
- ◯ No

Monthly burden of iGAS disease - Line chart



Line chart

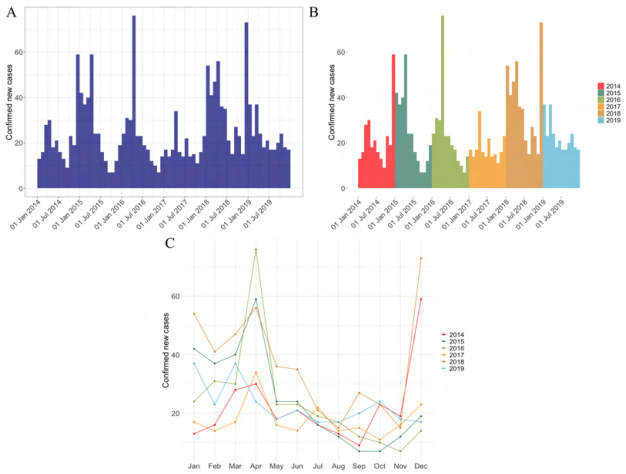
Monthly burden of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

8. Line chart - Which one of the following statements better reflects your opinion on this figure? *

Mark only one oval.

- ◯ It is a good way to compare the seasonal trends of iGAS disease across the six years considered
- ◯ It is hard to detect the disease burden in a specific point in time

Monthly burden of iGAS disease - Combined figures



Monthly burden of iGAS disease

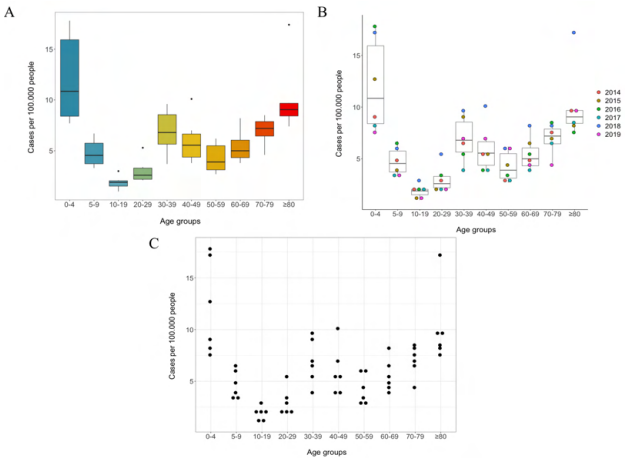
A (Monochromic histogram), B (Coloured histogram), C (Line chart) - Monthly burden of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of confirmed new cases per month.

9. Did you notice any seasonal trend in iGAS disease? if so, which visualisation better displays it? *

Mark only one oval.

- ◯ A
- ◯ B
- ◯ C

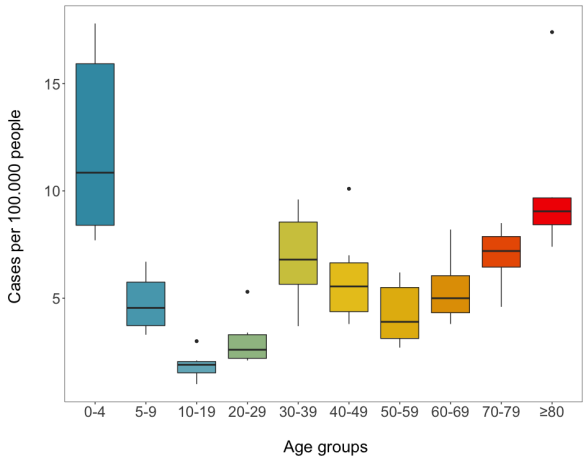
Incidence of iGAS disease in different age groups



Incidence of iGAS disease in different age groups

A (Box plot), B (Dot/box plot), C (Dot plot) - Age-specific incidence of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

Incidence of iGAS disease in different age groups - Box plot



Box plot

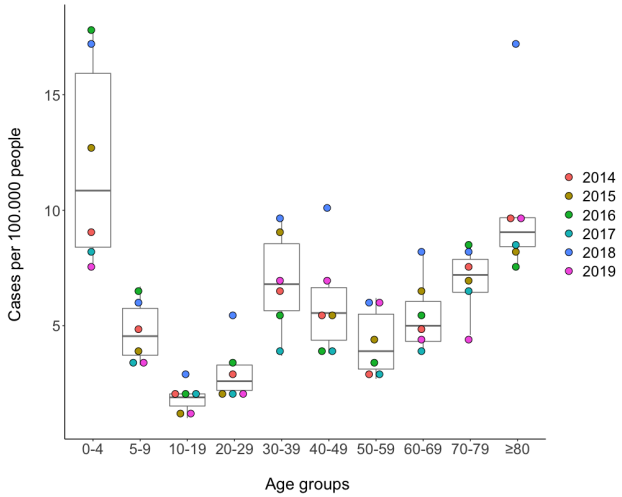
Age-specific incidence of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

10. In this figure colours are used purely for aesthetical reasons (they are not supposed to communicate any message). Which of the following statements do you agree with? *

Mark only one oval.

- ◯ The use of colours attracted my attention and helped me to focus on the data displayed
- ◯ The use of colours is superfluous and distracted me from the data displayed

Incidence of iGAS disease in different age groups - Dot/box plot



Dot/box plot

Age-specific incidence of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

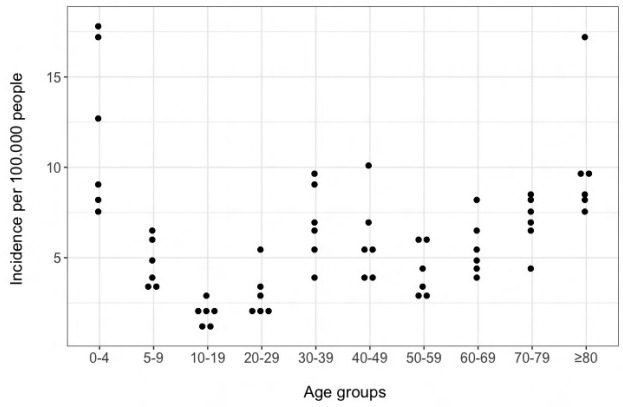
11. In this figure you can visualise the age-specific incidence of iGAS disease for each of the six years considered. Do you consider this additional detail relevant? *

Mark only one oval.

- ◯ Yes, it is important to know the age-specific incidence in each year considered
- ◯ Yes, but the colours make the figure hard to interpret
- ◯ No, what really matters is the average age-specific incidence

Incidence of iGAS disease in different age groups - Dot plot

Dot plot



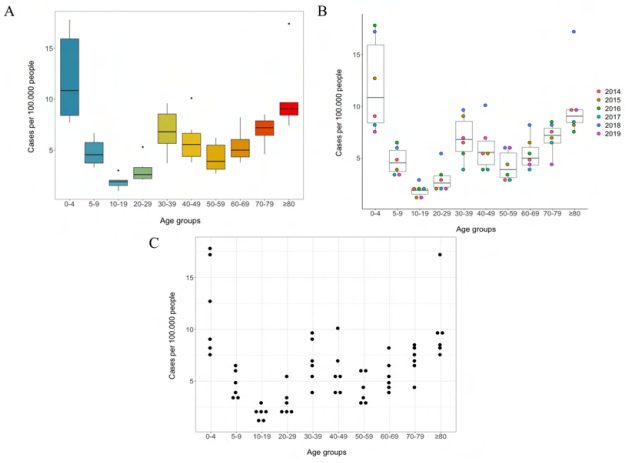
Dot plot
Age-specific incidence of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

12. Dot plot - In this figure simplicity is prioritised over details. Each dot represents an age-specific measurement in one of the six years considered but it is impossible to tell which year it refers to. Which of the following descriptions reflect your opinion about this figure? You can choose more than one answer *

Tick all that apply.

- ☐ I like it - in data communication less is always more
- ☐ It is incomplete - it is communicating only part of the available data
- ☐ It is too plain - I don't feel compelled to give it my attention
- ☐ It is simple and easy to interpret - I find it more effective than a busy figure with a lot of details

Incidence of iGAS disease in different age groups - Figures combined



Incidence of iGAS disease in different age groups

A (Box plot), B (Dot/box plot), C (Dot plot) - Age-specific incidence of iGAS disease in the Scottish population from 2014 to 2019 expressed as number of cases per 100.000 people per year.

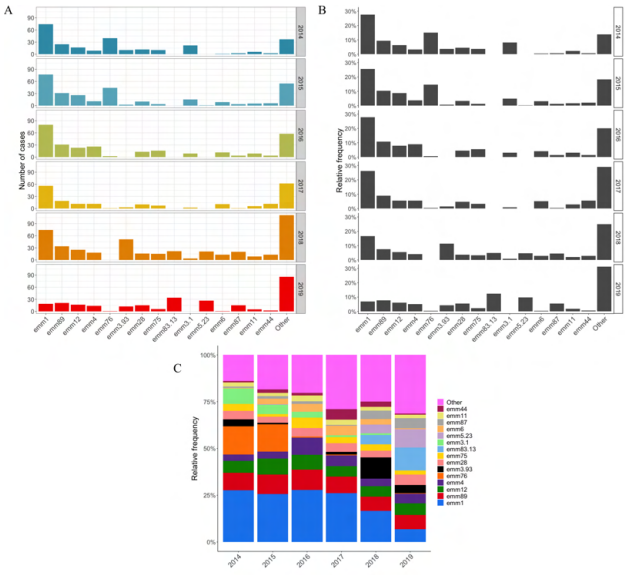
13. Which of the three options better shows the difference in iGAS disease incidence across age groups? *

Mark only one oval.

- A
- B
- C

Emm type-specific invasive disease burden

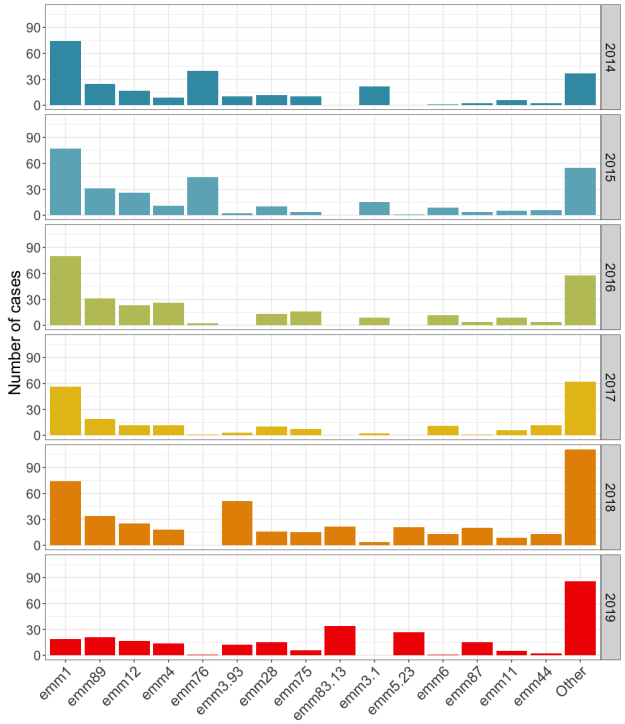
GAS strains are divided in emm types, which can be considered sub-populations within the species



Emm type-specific invasive disease burden

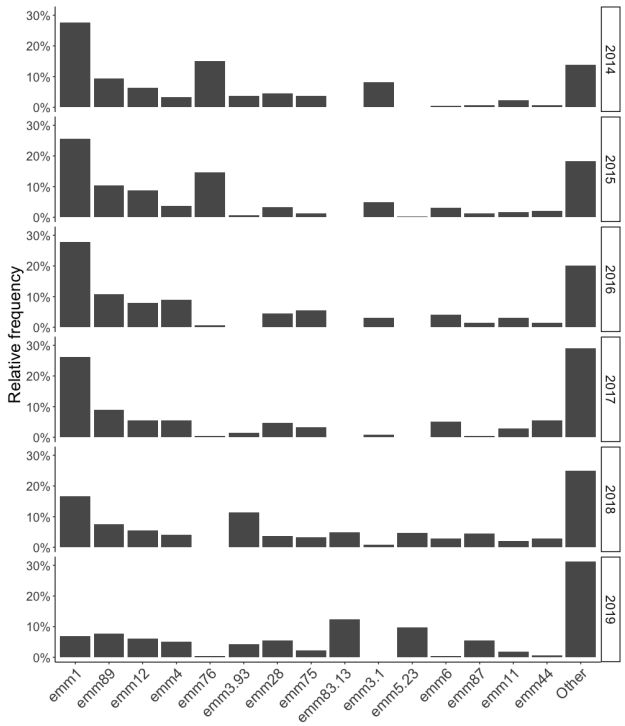
A (Coloured bar chart) - Absolute frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.
B (Black and white bar chart), C (Stacked bar chart) - Relative frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.

Emm type-specific invasive disease burden - Bar charts



Coloured bar chart

Absolute frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.



Black and white bar chart
Relative frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.

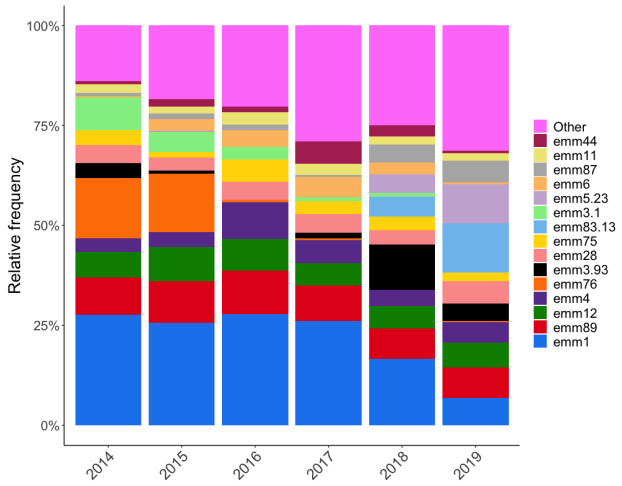
14. The emm-specific disease burden is expressed in a different way in each figure. Which option do you find more appropriate for this dataset? *

Mark only one oval.

- Absolute frequency (coloured bar chart)
- Relative frequency (black and white bar chart)

15. Focussing on the background and colour choice, which one of the two do you prefer? Can you provide a brief explanation for your choice? *

Emm type-specific invasive disease burden - Stacked bar chart

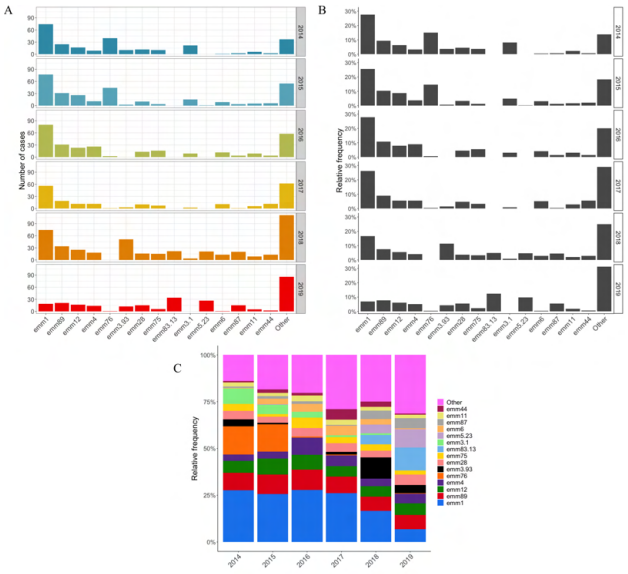


Stacked bar chart
Relative frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.

16. Name one positive attribute about this figure. *

17. Name one negative attribute about this figure. *

Emm type-specific invasive disease burden - Figures combined



Emm type-specific invasive disease burden

A (Coloured bar chart) - Absolute frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.
B (Black and white bar chart), C (Stacked bar chart) - Relative frequency of isolation of the 15 most common emm types circulating in Scotland from 2014 to 2019.

18. Overall, which one of the three options better represent the fluctuations in emm-specific disease burden? *

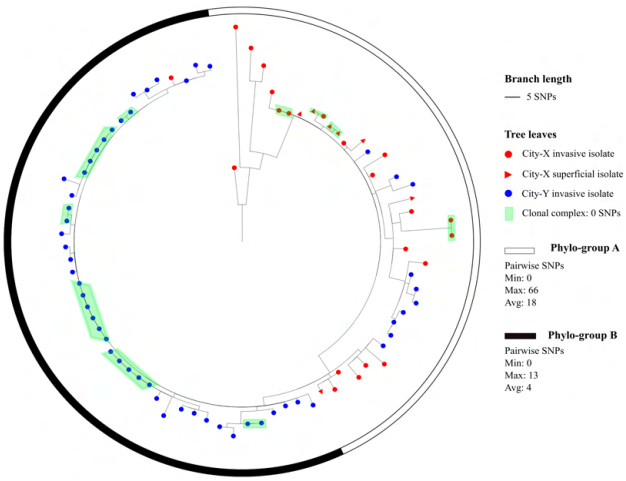
Mark only one oval.

- ◯ Coloured bar chart
- ◯ Black and white bar chart
- ◯ Stacked bar chart

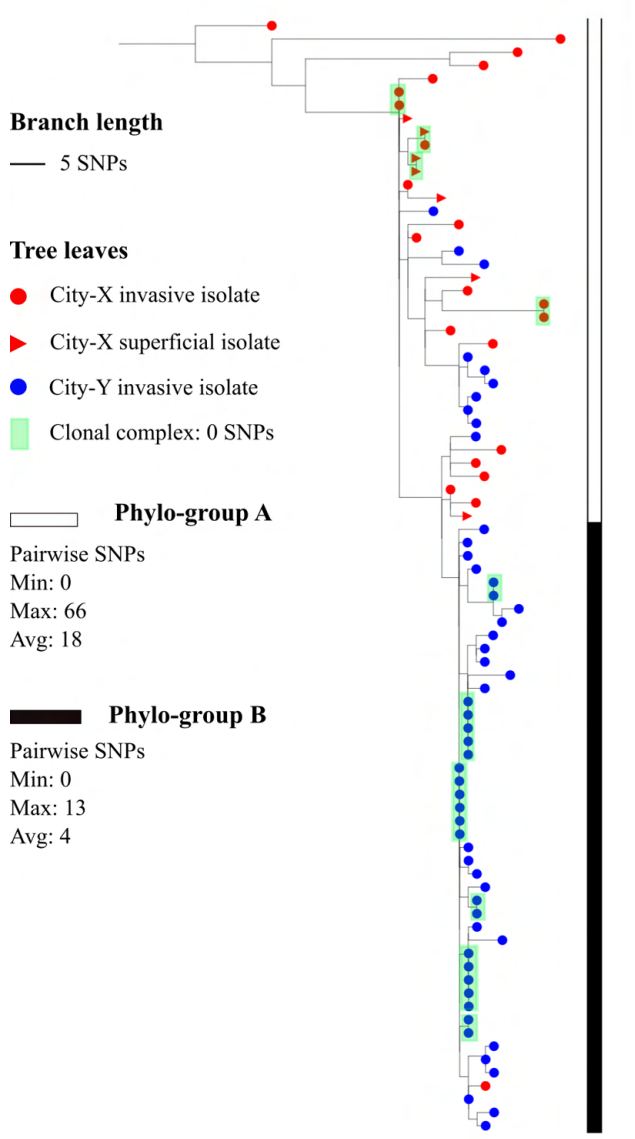
Phylogeny

The following core single nucleotide polymorphisms (SNPs) phylogenetic trees represent the evolutionary relatedness of GAS isolates belonging to the same emm type, emm 5.23. All the isolates were collected in city-X and city-Y between 2012 and 2021. For ease of interpretation, the part of the tree dominated by city-X isolates is called "Phylo-group A" and the one dominated by city-Y isolates is called "Phylo-group B".

Circular tree



Rectangular tree



19. Which one of the trees do you find more captivating? *

Mark only one oval.

- Circular tree
- Rectangular tree

20. Which one of the two trees do you find easier to read and interpret? *

Mark only one oval.

- Circular tree
- Rectangular tree

21. Do you think both trees display exactly the same information? If not could you tell the difference between them? *

References

Acheson, D. et al. (1988). Public health in england: the report of the committee of inquiry into the future development of the public health function. *London: The Stationary Office*, pages 23–34.

Agniswamy, J., Lei, B., Musser, J. M., and Sun, P. D. (2004). Insight of host immune evasion mediated by two variants of group a streptococcus mac protein. *Journal of Biological Chemistry*, 279(50):52789–52796.

Ajdic, D., Ferretti, J. J., et al. (1999). The rgg gene of streptococcus pyogenes nz131 positively influences extracellular spe b production. *Infection and Immunity*, 67(4):1715–1722.

Åkesson, P., Sjöholm, A. G., and Björck, L. (1996). Protein sic, a novel extracellular protein of streptococcus pyogenes interfering with complement function. *Journal of Biological Chemistry*, 271(2):1081–1088.

Alcock, B. P., Raphenya, A. R., Lau, T. T., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., et al. (2020). Card 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*, 48(D1):D517–D525.

Amsallem, M., Iung, B., Bouleti, C., Armand-Lefevre, L., Eme, A.-L., Touati, A., Kirsch, M., Duval, X., and Vahanian, A. (2014). First reported human case of native mitral infective endocarditis caused by streptococcus canis. *Canadian Journal of Cardiology*, 30(11):1462–e1.

Anders, S., Pyl, P. T., and Huber, W. (2015). Htseq—a python framework to work with high-throughput sequencing data. *bioinformatics*, 31(2):166–169.

Andrews, S. (2010). Fastqc: a quality control tool for high throughput sequence data.

Anselin, L., Syabri, I., and Kho, Y. (2010). Geoda: an introduction to spatial data analysis. In *Handbook of applied spatial analysis*, pages 73–89. Springer.

Aparicio, M. and Costa, C. J. (2015). Data visualization. *Communication design quarterly review*, 3(1):7–11.

Arai, K., Hirakata, Y., Yano, H., Kanamori, H., Endo, S., Hirotani, A., Abe, Y., Nagasawa, M., Kitagawa, M., Aoyagi, T., et al. (2011). Emergence of fluoroquinolone-resistant streptococcus pyogenes in japan by a point mutation leading to a new amino acid substitution. *Journal of antimicrobial chemotherapy*, 66(3):494–498.

Aruni, A. W., Mishra, A., Dou, Y., Chioma, O., Hamilton, B. N., and Fletcher, H. M. (2015). Filifactor alocis–a new emerging periodontal pathogen. *Microbes and infection*, 17(7):517–530.

Arvand, M., Hoeck, M., Hahn, H., and Wagner, J. (2000). Antimicrobial resistance in streptococcus pyogenes isolates in berlin. *Journal of Antimicrobial Chemotherapy*, 46(4):621–624.

Athey, T. B., Teatero, S., Sieswerda, L. E., Gubbay, J. B., Marchand-Austin, A., Li, A., Wasserscheid, J., Dewar, K., McGeer, A., Williams, D., et al. (2016). High incidence of invasive group a streptococcus disease caused by strains of uncommon emm types in thunder bay, ontario, canada. *Journal of clinical microbiology*, 54(1):83–92.

Avire, N. J., Whiley, H., and Ross, K. (2021). A review of streptococcus pyogenes: Public health risk factors, prevention and control. *Pathogens*, 10(2):248.

AvRuskin, G. A., Jacquez, G. M., Meliker, J. R., Slotnick, M. J., Kaufmann, A. M., and Nriagu, J. O. (2004). Visualization and exploratory analysis of epidemiologic data using a novel space time information system. *International Journal of Health Geographics*, 3(1):1–10.

Azzam, T., Evergreen, S., Germuth, A. A., and Kistler, S. J. (2013). Data visualization and evaluation. *New Directions for Evaluation*, 2013(139):7–32.

Bah, T. (2007). *Inkscape: guide to a vector drawing program*. prentice hall press.

Balaban, M., Moshiri, N., Mai, U., Jia, X., and Mirarab, S. (2019). Treecluster: Clustering biological sequences using phylogenetic trees. *PloS one*, 14(8):e0221068.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477.

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., and Brooks, C. (2010). Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2573–2582.

Beall, B., Facklam, R., and Thompson, T. (1996). Sequencing emm-specific pcr products for routine and accurate typing of group a streptococci. *Journal of clinical microbiology*, 34(4):953–958.

Beaton, A., Kamalembo, F. B., Dale, J., Kado, J. H., Karthikeyan, G., Kazi, D. S., Longenecker, C. T., Mwangi, J., Okello, E., Ribeiro, A. L. P., et al. (2020). The american heart association's call to action for reducing the global burden of rheumatic heart disease: a policy statement from the american heart association. *Circulation*, 142(20):e358–e368.

Beckert, S., Kreikemeyer, B., and Podbielski, A. (2001). Group a streptococcal rofa gene is involved in the control of several virulence genes and eukaryotic cell attachment and internalization. *Infection and immunity*, 69(1):534–537.

Ben, A. N., Herwald, H., Sjöbring, U., Renné, T., Müller-Esterl, W., and Björck, L. (1997). Absorption of kininogen from human plasma by streptococcus pyogenes is followed by the release of bradykinin. *The Biochemical journal*, 326(Pt 3):657–660.

Ben Zakour, N. L., Venturini, C., Beatson, S. A., and Walker, M. J. (2012). Analysis of a streptococcus pyogenes puerperal sepsis cluster by use of whole-genome sequencing. *Journal of clinical microbiology*, 50(7):2224–2228.

Benfield, T., Jensen, J., and Nordestgaard, B. (2007). Influence of diabetes and hyperglycaemia on infectious disease hospitalisation and outcome. *Diabetologia*, 50(3):549–554.

Beres, S. B., Kachroo, P., Nasser, W., Olsen, R. J., Zhu, L., Flores, A. R., de la Riva, I., Paez-Mayorga, J., Jimenez, F. E., Cantu, C., et al. (2016). Transcriptome remodeling contributes to epidemic disease caused by the human pathogen streptococcus pyogenes. *MBio*, 7(3).

Beres, S. B., Olsen, R. J., Saavedra, M. O., Ure, R., Reynolds, A., Lindsay, D. S., Smith, A. J., and Musser, J. M. (2017). Genome sequence analysis of emm89 streptococcus pyogenes strains causing infections in scotland, 2010–2016. *Journal of medical microbiology*, 66(12):1765.

Beres, S. B., Sylva, G. L., Barbian, K. D., Lei, B., Hoff, J. S., Mammarella, N. D., Liu, M.-Y., Smoot, J. C., Porcella, S. F., Parkins, L. D., et al. (2002). Genome sequence of a serotype m3 strain of group a streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proceedings of the National Academy of Sciences*, 99(15):10078–10083.

Beres, S. B., Zhu, L., Pruitt, L., Olsen, R. J., Faili, A., Kayal, S., and Musser, J. M. (2022). Integrative reverse genetic analysis identifies polymorphisms contributing to decreased antimicrobial agent susceptibility in streptococcus pyogenes. *Mbio*, 13(1):e03618–21.

Berggård, K., Johnsson, E., Morfeldt, E., Persson, J., Stålhammar-Carlemalm, M., and Lindahl, G. (2001). Binding of human c4bp to the hypervariable region of m protein: a molecular mechanism of phagocytosis resistance in streptococcus pyogenes. *Molecular microbiology*, 42(2):539–551.

Bergmann, S., Eichhorn, I., Kohler, T. P., Hammerschmidt, S., Goldmann, O., Rohde, M., and Fulde, M. (2017). Scm, the m protein of streptococcus canis binds immunoglobulin g. *Frontiers in cellular and infection microbiology*, 7:80.

Berkley, J. A., Lowe, B. S., Mwangi, I., Williams, T., Bauni, E., Mwarumba, S., Ngetsa, C., Slack, M. P., Njenga, S., Hart, C. A., et al. (2005). Bacteremia among children admitted to a rural hospital in kenya. *New England Journal of Medicine*, 352(1):39–47.

Bert, F. and Lambert-Zechovsky, N. (1997). Septicemia caused by streptococcus canis in a human. *Journal of clinical microbiology*, 35(3):777–779.

Bertin, J. and Barbut, M. (1967). Sémiologie graphique. les diagrammes, les réseaux, les cartes paris.

Bertini, E. and Santucci, G. (2006). Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110.

Bessell, P. R., Auty, H. K., Roberts, H., McKendrick, I. J., Bronsvoort, B. M. d. C., and Boden, L. A. (2020). A tool for prioritizing livestock disease threats to scotland. *Frontiers in veterinary science*, 7:223.

Bessen, D. E. (2009). Population biology of the human restricted pathogen, streptococcus pyogenes. *Infection, Genetics and Evolution*, 9(4):581–593.

Bessen, D. E., McShan, W. M., Nguyen, S. V., Shetty, A., Agrawal, S., and Tettelin, H. (2015). Molecular epidemiology and genomics of group a streptococcus. *Infection, Genetics and Evolution*, 33:393–418.

Betschel, S. D., Borgia, S. M., Barg, N. L., Low, D. E., and De Azavedo, J. C. (1998). Reduced virulence of group a streptococcal tn916 mutants that do not produce streptolysin s. *Infection and immunity*, 66(4):1671–1679.

Bhakdi, S., Tranum-Jensen, J., and Sziegoleit, A. (1985). Mechanism of membrane damage by streptolysin-o. *Infection and immunity*, 47(1):52–60.

Bhatia, R. (2019). Implementation framework for one health approach. *The Indian journal of medical research*, 149(3):329.

Bi, S., Xu, M., Zhou, Y., Xing, X., Shen, A., and Wang, B. (2019). A multicomponent vaccine provides immunity against local and systemic infections by group a streptococcus across serotypes. *MBio*, 10(6):e02600–19.

Bingen, E., Denamur, E., Lambert-Zechovsky, N., Boissinot, C., Brahimi, N., Aujard, Y., Blot, P., and Elion, J. (1992a). Mother-to-infant vertical transmission and cross-colonization of streptococcus pyogenes confirmed by dna restriction fragment length polymorphism analysis. *Journal of infectious diseases*, 165(1):147–150.

Bingen, E., Denamur, E., Lambert-Zechovsky, N., Braimi, N., El Lakany, M., and Elion, J. (1992b). Dna restriction fragment length polymorphism differentiates recurrence from relapse in treatment failures of streptococcus pyogenes pharyngitis. *Journal of medical microbiology*, 37(3):162–164.

Bisno, A. L., Gerber, M. A., Gwaltney Jr, J. M., Kaplan, E. L., and Schwartz, R. H. (2002). Practice guidelines for the diagnosis and management of group a streptococcal pharyngitis. *Clinical infectious diseases*, pages 113–125.

Bisno, A. L. and Stevens, D. L. (1996). Streptococcal infections of skin and soft tissues. *New England Journal of Medicine*, 334(4):240–246.

Blagden, S., Watts, V., Verlander, N., and Pegorie, M. (2020). Invasive group a streptococcal infections in north west england: epidemiology, risk factors and fatal infection. *Public Health*, 186:63–70.

Blanton, J. D., Manangan, A., Manangan, J., Hanlon, C. A., Slate, D., and Rupprecht, C. E. (2006). Development of a gis-based, real-time internet mapping tool for rabies surveillance. *International journal of health geographics*, 5(1):1–8.

Bloom, D. E. and Cadarette, D. (2019). Infectious disease threats in the twenty-first century: strengthening the global response. *Frontiers in immunology*, 10:549.

Boden, L. A., Voas, S., Mellor, D., and Auty, H. (2020). Epic, scottish government's centre of expertise in animal disease outbreaks: A model for provision of risk-based evidence to policy. *Frontiers in veterinary science*, 7:119.

Bohach, G. A., Fast, D. J., Nelson, R. D., and Schlievert, P. M. (1990). Staphylococcal and streptococcal pyrogenic toxins involved in toxic shock syndrome and related illnesses. *Critical reviews in microbiology*, 17(4):251–272.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Brenciani, A., Bacciaglia, A., Vecchi, M., Vitali, L. A., Varaldo, P. E., and Giovanetti, E. (2007). Genetic elements carrying erm (b) in streptococcus pyogenes and association with tet (m) tetracycline resistance gene. *Antimicrobial agents and chemotherapy*, 51(4):1209–1216.

Brenciani, A., Bacciaglia, A., Vignaroli, C., Pugnaloni, A., Varaldo, P. E., and Giovanetti, E. (2010). ϕm46. 1, the main streptococcus pyogenes element carrying mef (a) and tet (o) genes. *Antimicrobial agents and chemotherapy*, 54(1):221–229.

Brennan, M. R. and LeFevre, F. (2019). Necrotizing fasciitis: Infection identification and management. *Nursing2020 Critical Care*, 14(1):6–11.

Breskin, A., Adimora, A. A., and Westreich, D. (2017). Women and hiv in the united states. *PLoS One*, 12(2):e0172367.

Bricker, A. L., Cywes, C., Ashbaugh, C. D., and Wessels, M. R. (2002). Nad+-glycohydrolase acts as an intracellular toxin to enhance the extracellular survival of group a streptococci. *Molecular microbiology*, 44(1):257–269.

Briko, N., Filatov, N., Zhuravlev, M., Lytkina, I., Ezhlova, E., Brazhnikov, A., Tsapkova, N., and Malyshev, N. (2003). Epidemiological pattern of scarlet fever in recent years. *Zhurnal mikrobiologii, epidemiologii, i immunobiologii*, (5):67–72.

Brinton, W. C. (1939). *Graphic presentation*. Brinton associates.

Brouwer, S., Barnett, T. C., Rivera-Hernandez, T., Rohde, M., and Walker, M. J. (2016). Streptococcus pyogenes adhesion and colonization. *FEBS letters*, 590(21):3739–3757.

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with scoary. *Genome biology*, 17(1):1–9.

Buache, P. (1752). Essai de géographie physique. *Mémoires de l'Académie royale des sciences*, pages 399–416.

Bubba, L., Bundle, N., Kapatai, G., Daniel, R., Balasegaram, S., Anderson, C., Chalker, V., Lamagni, T., Brown, C., Ready, D., et al. (2019). Genomic sequencing of a national emm66 group a streptococci (gas) outbreak among people who inject drugs and the homeless community in england and wales, january 2016–may 2017. *Journal of Infection*, 79(5):435–443.

Buchanan, J. T., Simpson, A. J., Aziz, R. K., Liu, G. Y., Kristian, S. A., Kotb, M., Feramisco, J., and Nizet, V. (2006). Dnase expression allows the pathogen group a streptococcus to escape killing in neutrophil extracellular traps. *Current Biology*, 16(4):396–400.

Bundle, N., Bubba, L., Coelho, J., Kwiatkowska, R., Cloke, R., King, S., Rajan-Iyer, J., Courtney-Pillinger, M., Beck, C. R., Hope, V., et al. (2017). Ongoing outbreak of invasive and non-invasive disease due to group a streptococcus (gas) type emm66 among homeless and people who inject drugs in england and wales, january to december 2016. *Eurosurveillance*, 22(3):30446.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1–9.

Carapetis, J., Gardiner, D., Currie, B., and Mathews, J. D. (1995). Multiple strains of streptococcus pyogenes in skin sores of aboriginal australians. *Journal of clinical microbiology*, 33(6):1471–1472.

Carapetis, J., Steer, A., Mulholland, E., and Weber, M. (2005). The global burden of group a streptococcal diseases. *The Lancet infectious diseases*, 5(11):685–694.

Carlsson, F., Berggård, K., Stålhammar-Carlemalm, M., and Lindahl, G. (2003). Evasion of phagocytosis through cooperation between two ligand-binding regions in streptococcus pyogenes m protein. *Journal of Experimental Medicine*, 198(7):1057–1068.

Carvalho-Castro, G. A., Silva, J. R., Paiva, L. V., Custódio, D. A., Moreira, R. O., Mian, G. F., Prado, I. A., Chalfun-Junior, A., and Costa, G. M. (2017). Molecular epidemiology of streptococcus agalactiae isolated from mastitis in brazilian dairy herds. *brazilian journal of microbiology*, 48(3):551–559.

Castro, S. A. and Dorfmueller, H. C. (2021). A brief review on group a streptococcus pathogenesis and vaccine development. *Royal Society open science*, 8(3):201991.

Castronovo, D. A., Chui, K. K., and Naumova, E. N. (2009). Dynamic maps: a visual-analytic methodology for exploring spatio-temporal disease patterns. *Environmental Health*, 8(1):1–9.

Caswell, C. C., Oliver-Kozup, H., Han, R., Lukomska, E., and Lukomski, S. (2010). Scl1, the multifunctional adhesin of group a streptococcus, selectively binds cellular fibronectin and laminin, and mediates pathogen internalization by human cells. *FEMS microbiology letters*, 303(1):61–68.

CDC (2018). M protein gene (*emm*) typing. `https://www.cdc.gov/streplab/groupa-strep/emm-background.html`.

CDC (2020). Abcs report: Group a streptococcus, 2018. `https://www.cdc.gov/abcs/reports-findings/survreports/gas18.html`.

Chaffer, M., Friedman, S., Saran, A., and Younis, A. (2005). An outbreak of streptococcus canis mastitis in a dairy herd in israel. *New Zealand Veterinary Journal*, 53(4):261–264.

Chalker, V. J., Smith, A., Al-Shahib, A., Botchway, S., Macdonald, E., Daniel, R., Phillips, S., Platt, S., Doumith, M., Tewolde, R., et al. (2016). Integration of genomic and other epidemiologic data to investigate and control a cross-institutional outbreak of streptococcus pyogenes. *Emerging infectious diseases*, 22(6):973.

Chang, A., Khemlani, A., Kang, H., and Proft, T. (2011). Functional analysis of streptococcus pyogenes nuclease a (spna), a novel group a streptococcal virulence factor. *Molecular microbiology*, 79(6):1629–1642.

Chen, Hardle, W. K., and Unwin, A. (2007). *Handbook of data visualization*. Springer Science & Business Media.

Chen, H., Chen, W., Mei, H., Liu, Z., Zhou, K., Chen, W., Gu, W., and Ma, K.-L. (2014). Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1683–1692.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., and Jin, Q. (2005). Vfdb: a reference database for bacterial virulence factors. *Nucleic acids research*, 33(suppl_1):D325–D328.

Chen, M., Wang, W., Tu, L., Zheng, Y., Pan, H., Wang, G., Chen, Y., Zhang, X., Zhu, L., Chen, J., et al. (2017). An emm 5 group a streptococcal outbreak among workers in a factory manufacturing telephone accessories. *Frontiers in microbiology*, 8:1156.

Chen, M., Yao, W., Wang, X., Li, Y., Chen, M., Wang, G., Zhang, X., Pan, H., Hu, J., and Zeng, M. (2012). Outbreak of scarlet fever associated with emm12 type group a streptococcus in 2011 in shanghai, china. *The Pediatric infectious disease journal*, 31(9):e158–e162.

Chiang-Ni, C., Wu, A.-B., Liu, C.-C., Chen, K.-T., Lin, Y.-S., Chuang, W.-J., Fang, H.-Y., and Wu, J.-J. (2011). Emergence of uncommon emm types of streptococcus pyogenes among adult patients in southern taiwan. *Journal of Microbiology, Immunology and Infection*, 44(6):424–429.

Churakov, M., Katholm, J., Rogers, S., Kao, R. R., and Zadoks, R. N. (2021). Assessing potential routes of streptococcus agalactiae transmission between dairy herds using national surveillance, animal movement data and molecular typing. *Preventive Veterinary Medicine*, 197:105501.

Churchward, G. (2007). The two faces of janus: virulence gene regulation by covr/s in group a streptococci. *Molecular microbiology*, 64(1):34–41.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Clagett, M. et al. (1968). Nicole oresme and the medieval geometry of qualities and motions: A treatise on the uniformity and difformity of intensities known as tractatus de configurationibus qualitatum et motuum. *Publications in medieval science*.

Clark, R. C. and Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. john Wiley & sons.

Clausen, P. T., Zankari, E., Aarestrup, F. M., and Lund, O. (2016). Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *Journal of Antimicrobial Chemotherapy*, 71(9):2484–2488.

Cleary, P. P., Kaplan, E., Livdahl, C., and Skjold, S. (1988). Dna fingerprints of streptococcus pyogenes are m type specific. *Journal of infectious diseases*, 158(6):1317–1323.

Cleary, P. P., McLandsborough, L., Ikeda, L., Cue, D., Krawczak, J., and Lam, H. (1998). High-frequency intracellular infection and erythrogenic toxin a expression undergo phase variation in m1 group a streptococci. *Molecular microbiology*, 28(1):157–167.

Cleary, P. P., Schlievert, P., Handley, J., Kim, M., Hauser, A., Kaplan, E., and Wlazlo, A. (1992). Clonal basis for resurgence of serious streptococcus pyogenes disease in the 1980s. *The Lancet*, 339(8792):518–521.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833.

Cleveland, W. S. and McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500.

Cleveland, W. S. and McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society: Series A (General)*, 150(3):192–210.

Cloutier, G., D'allaire, S., Martinez, G., Surprenant, C., Lacouture, S., and Gottschalk, M. (2003). Epidemiology of streptococcus suis serotype 5 infection in a pig herd with and without clinical disease. *Veterinary microbiology*, 97(1-2):135–151.

Cole, J. J., Faydaci, B. A., McGuinness, D., Shaw, R., Maciewicz, R. A., Robertson, N. A., and Goodyear, C. S. (2021). Searchlight: automated bulk rna-seq exploration and visualisation using dynamically generated r scripts. *BMC bioinformatics*, 22(1):1–21.

Coleman, J. L. and Benach, J. L. (1999). Use of the plasminogen activation system by microorganisms. *Journal of Laboratory and Clinical Medicine*, 134(6):567–576.

Collin, M., Svensson, M. D., Sjöholm, A. G., Jensenius, J. C., Sjöbring, U., and Olsén, A. (2002). Endos and speb from streptococcus pyogenes inhibit immunoglobulin-mediated opsonophagocytosis. *Infection and immunity*, 70(12):6646–6651.

Collins, C., Penn, G., and Carpendale, S. (2009). Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1009–1016.

Committee for the Study of the Future of Public Health, I. o. M. (1988). The future of public health.

Commodari, E. and La Rosa, V. L. (2020). Adolescents in quarantine during covid-19 pandemic in italy: perceived health risk, beliefs, psychological experiences and expectations for the future. *Frontiers in psychology*, 11:559951.

Commons, R. J., Smeesters, P. R., Proft, T., Fraser, J. D., Robins-Browne, R., and Curtis, N. (2014). Streptococcal superantigens: categorization and clinical associations. *Trends in molecular medicine*, 20(1):48–62.

Cordery, R., Purba, A. K., Begum, L., Mills, E., Mosavie, M., Vieira, A., Jauneikaite, E., Leung, R. C., Siggins, M. K., Ready, D., et al. (2022). Frequency of transmission, asymptomatic shedding, and airborne spread of streptococcus pyogenes in schoolchildren exposed to scarlet fever: a prospective, longitudinal, multicohort, molecular epidemiological, contact-tracing study in england, uk. *The Lancet Microbe*, 3(5):e366–e375.

Cornax, I., Zulk, J., Olson, J., Fulde, M., Nizet, V., and Patras, K. A. (2021). Novel models of streptococcus canis colonization and disease reveal modest contributions of m-like (scm) protein. *Microorganisms*, 9(1):183.

Courtney, H. S., Hasty, D. L., and Dale, J. B. (2006). Anti-phagocytic mechanisms of streptococcus pyogenes: binding of fibrinogen to m-related protein. *Molecular microbiology*, 59(3):936–947.

Crestani, C., Forde, T. L., Lycett, S. J., Holmes, M. A., Fasth, C., Persson-Waller, K., and Zadoks, R. N. (2021). The fall and rise of group b streptococcus in dairy cattle: reintroduction due to human-to-cattle host jumps? *Microbial genomics*, 7(9).

Creti, R. (2017). Have group a and b streptococcal infections become neglected diseases in europe?

Croll, D. and McDonald, B. A. (2012). The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS pathogens*, 8(4):e1002608.

Cunningham, M. W. (2000). Pathogenesis of group a streptococcal infections. *Clinical microbiology reviews*, 13(3):470–511.

Curtis, S., Tanna, A., Russell, H., Efstratiou, A., Paul, J., Cubbon, M., and Sriskandan, S. (2007). Invasive group a streptococcal infection in injecting drug users and non-drug users in a single uk city. *Journal of Infection*, 54(5):422–426.

Cywes, C. and Wessels, M. R. (2001). Group a streptococcus tissue invasion by cd44-mediated cell signalling. *Nature*, 414(6864):648.

Dale, J. B. and Walker, M. J. (2020). Update on group a streptococcal vaccine development. *Current opinion in infectious diseases*, 33(3):244.

Dale, J. B., Washburn, R. G., Marques, M. B., and Wessels, M. R. (1996). Hyaluronate capsule and surface m protein in resistance to opsonization of group a streptococci. *Infection and immunity*, 64(5):1495–1501.

Dallas, S. D., McGee, L., Limbago, B., Patel, J. B., McElmeel, M. L., Fulcher, L. C., Lonsway, D. R., and Jorgensen, J. H. (2013). Development of doxycycline mic and disk diffusion interpretive breakpoints and revision of tetracycline breakpoints for streptococcus pneumoniae. *Journal of clinical microbiology*, 51(6):1798–1802.

Dalton, T. L. and Scott, J. R. (2004). Covs inactivates covr and is required for growth under conditions of general stress in streptococcus pyogenes. *Journal of bacteriology*, 186(12):3928–3937.

Danchin, M. H., Rogers, S., Kelpie, L., Selvaraj, G., Curtis, N., Carlin, J. B., Nolan, T. M., and Carapetis, J. R. (2007). Burden of acute sore throat and group a streptococcal pharyngitis in school-aged children and their families in australia. *Pediatrics*, 120(5):950–957.

Danino, D., Ben-Shimol, S., Van der Beek, B. A., Givon-Lavi, N., Avni, Y. S., Greenberg, D., Weinberger, D. M., and Dagan, R. (2021). Decline in pneumococcal disease in young children during the coronavirus disease 2019 (covid-19) pandemic in israel associated with suppression of seasonal respiratory viruses, despite persistent pneumococcal carriage: A prospective cohort study. *Clinical Infectious Diseases*.

Darenberg, J., Luca-Harari, B., Jasir, A., Sandgren, A., Pettersson, H., Schalén, C., Norgren, M., Romanus, V., Norrby-Teglund, A., and Normark, B. H. (2007). Molecular and clinical characteristics of invasive group a streptococcal infection in sweden. *Clinical infectious diseases*, 45(4):450–458.

Davies, H. D., McGeer, A., Schwartz, B., Green, K., Cann, D., Simor, A. E., Low, D. E., and Group, O. G. A. S. S. (1996). Invasive group a streptococcal infections in ontario, canada. *New England Journal of Medicine*, 335(8):547–554.

Davies, M. R., Holden, M. T., Coupland, P., Chen, J. H., Venturini, C., Barnett, T. C., Zakour, N. L. B., Tse, H., Dougan, G., Yuen, K.-Y., et al. (2015). Emergence of scarlet fever streptococcus pyogenes emm12 clones in hong kong is associated with toxin acquisition and multidrug resistance. *Nature genetics*, 47(1):84.

Davies, M. R., McIntyre, L., Mutreja, A., Lacey, J. A., Lees, J. A., Towers, R. J., Duchêne, S., Smeesters, P. R., Frost, H. R., Price, D. J., et al. (2019). Atlas of group a streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nature genetics*, 51(6):1035–1043.

Dégi, J., Cristina, R., et al. (2011). Clinical prevalence of streptococcus canis in cats, with otitis externa in western region of romania. *Lucrari Stiintifice-Universitatea de Stiinte Agricole a Banatului Timisoara, Medicina Veterinara*, 44(2):282–284.

Desai, M., Efstratiou, A., George, R., and Stanley, J. (1999). High-resolution genotyping of streptococcus pyogenesserotype m1 isolates by fluorescent amplified-fragment length polymorphism analysis. *Journal of clinical microbiology*, 37(6):1948–1952.

Descartes, R. (1637). *Discours de la méthode. La Géométrie*.

Devriese, L., Colque, J. C., De Herdt, P., and Haesebrouck, F. (1992). Identification and composition of the tonsillar and anal enterococcal and streptococcal flora of dogs and cats. *Journal of Applied Bacteriology*, 73(5):421–425.

Devriese, L. A., Hommez, J., Kilpper-Bälz, R., and Schleifer, K.-H. (1986). Streptococcus canis sp. nov.: a species of group g streptococci from animals. *International Journal of Systematic and Evolutionary Microbiology*, 36(3):422–425.

DeWinter, L., Low, D., and Prescott, J. (1999). Virulence of streptococcus canis from canine streptococcal toxic shock syndrome and necrotizing fasciitis. *Veterinary microbiology*, 70(1-2):95–110.

DeWinter, L. and Prescott, J. (1999). Relatedness of streptococcus canis from canine streptococcal toxic shock syndrome and necrotizing fasciitis. *Canadian journal of veterinary research*, 63(2):90.

Dey, S. K., Rahman, M. M., Siddiqi, U. R., and Howlader, A. (2020). Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis approach. *Journal of medical virology*, 92(6):632–638.

Dominkovics, P., Granell, C., Pérez-Navarro, A., Casals, M., Orcau, À., and Caylà, J. A. (2011). Development of spatial density maps based on geoprocessing web services: application to tuberculosis incidence in barcelona, spain. *International journal of health geographics*, 10(1):1–14.

Dowell, S. F., Whitney, C. G., Wright, C., Rose Jr, C. E., and Schuchat, A. (2003). Seasonal patterns of invasive pneumococcal disease. *Emerging infectious diseases*, 9(5):574.

du Carla-Boniface, M. (1782). Expression des nivellements; ou, méthode nouvelle pour marquer sur les cartes terrestres et marines les hauteurs et les configurations du terrain.

Edwards, R. J., Taylor, G. W., Ferguson, M., Murray, S., Rendell, N., Wrigley, A., Bai, Z., Boyle, J., Finney, S. J., Jones, A., et al. (2005). Specific c-terminal cleavage and inactivation of interleukin-8 by invasive disease isolates of streptococcus pyogenes. *The Journal of infectious diseases*, 192(5):783–790.

Efstratiou, A. (2000). Group a streptococci in the 1990s. *Journal of Antimicrobial Chemotherapy*, 45(suppl_1):3–12.

Efstratiou, A. and Lamagni, T. (2016). Epidemiology of streptococcus pyogenes.

Eibl, C., Baumgartner, M., Urbantke, V., Sigmund, M., Lichtmannsperger, K., Wittek, T., and Spergser, J. (2021). An outbreak of subclinical mastitis in a dairy herd caused by a novel streptococcus canis sequence type (st55). *Animals*, 11(2):550.

Elliott, M. and Rensink, R. (2015). Interference in the perception of two-population scatterplots. *Journal of Vision*, 15(12):893–893.

Enache, A. E., Mitchell, C., Kafarnik, C., and Waller, A. S. (2020). Streptococcus canis multilocus sequence typing in a case series of dogs with ulcerative keratitis. *Veterinary ophthalmology*, 23(2):252–258.

Engelthaler, D. M., Valentine, M., Bowers, J., Pistole, J., Driebe, E. M., Terriquez, J., Nienstadt, L., Carroll, M., Schumacher, M., Ormsby, M. E., et al. (2016). Hypervirulent emm59 clone in invasive group a streptococcus outbreak, southwestern united states. *Emerging infectious diseases*, 22(4):734.

Enright, M. C., Spratt, B. G., Kalia, A., Cross, J. H., and Bessen, D. E. (2001). Multilocus sequence typing of streptococcus pyogenes and the relationships between emm type and clone. *Infection and immunity*, 69(4):2416–2427.

Facklam, R., Beall, B., Efstratiou, A., Fischetti, V., Johnson, D., Kaplan, E., Kriz, P., Lovgren, M., Martin, D., Schwartz, B., et al. (1999). emm typing and validation of provisional m types for group a streptococci. *Emerging infectious diseases*, 5(2):247.

Fagan, P. K., Reinscheid, D., Gottschalk, B., and Chhatwal, G. S. (2001). Identification and characterization of a novel secreted immunoglobulin binding protein from group a streptococcus. *Infection and immunity*, 69(8):4851–4857.

Falkenhorst, G., Bagdonaite, J., Lisby, M., Madsen, S., Lambertsen, L., Olsen, K., and Mølbak, K. (2008). Outbreak of group a streptococcal throat infection: don't forget to ask about food. *Epidemiology & Infection*, 136(9):1165–1171.

Falugi, F., Zingaretti, C., Pinto, V., Mariani, M., Amodeo, L., Manetti, A. G., Capo, S., Musser, J. M., Orefici, G., Margarit, I., et al. (2008). Sequence variation in group a streptococcus pili and association of pilus backbone types with lancefield t serotypes. *The Journal of infectious diseases*, 198(12):1834–1841.

Fay, K., Onukwube, J., Chochua, S., Schaffner, W., Cieslak, P., Lynfield, R., Muse, A., Smelser, C., Harrison, L. H., Farley, M., et al. (2021). Patterns of antibiotic nonsusceptibility among invasive group a streptococcus infections—united states, 2006–2017. *Clinical Infectious Diseases*, 73(11):1957–1964.

Feng, W., Liu, M., Chen, D. G., Yiu, R., Fang, F. C., and Lei, B. (2016). Contemporary pharyngeal and invasive emm1 and invasive emm12 group a streptococcus isolates exhibit similar in vivo selection for covrs mutants in mice. *PLoS One*, 11(9):e0162742.

Ferreira, J. and Zwinderman, A. (2006). On the benjamini–hochberg method. *The Annals of Statistics*, 34(4):1827–1849.

Few, S. and Edge, P. (2007). Save the pies for dessert. *Visual business intelligence newsletter*, pages 1–14.

Fiedler, T., Köller, T., and Kreikemeyer, B. (2015). Streptococcus pyogenes biofilms—formation, biology, and clinical relevance. *Frontiers in cellular and infection microbiology*, 5:15.

Fiedler, T., Sugareva, V., Patenge, N., and Kreikemeyer, B. (2010). Insights into streptococcus pyogenes pathogenesis from transcriptome studies. *Future microbiology*, 5(11):1675–1694.

Freeman, T. C., Horsewell, S., Patir, A., Harling-Lee, J., Regan, T., Shih, B. B., Prendergast, J., Hume, D. A., and Angus, T. (2020). Graphia: A platform for the graph-based visualisation and analysis of complex data. *bioRxiv*.

Frick, I.-M., Åkesson, P., Rasmussen, M., Schmidtchen, A., and Björck, L. (2003). Sic, a secreted protein of streptococcus pyogenesthat inactivates antibacterial peptides. *Journal of Biological Chemistry*, 278(19):16561–16566.

Friendly, M. (2008). A brief history of data visualization. In *Handbook of data visualization*, pages 15–56. Springer.

Fukushima, Y., Takahashi, T., Goto, M., Yoshida, H., and Tsuyuki, Y. (2020a). Novel diverse sequences of the streptococcus canis m-like protein (scm) gene and their prevalence in diseased companion animals: association of their alleles with sequence types. *Journal of Infection and Chemotherapy*, 26(9):908–915.

Fukushima, Y., Tsuyuki, Y., Goto, M., Yoshida, H., and Takahashi, T. (2020b). Novel quinolone nonsusceptible streptococcus canis strains with point mutations in quinolone resistance-determining regions and their related factors. *Japanese journal of infectious diseases*, 73(3):242–249.

Fukushima, Y., Yoshida, H., Goto, M., Tsuyuki, Y., and Takahashi, T. (2018). Prevalence and diversity of m-like protein (scm) gene in streptococcus canis isolates from diseased companion animals in japan: implication of scm allele. *Veterinary microbiology*, 225:120–124.

Fulde, M., Rohde, M., Hitzmann, A., Preissner, K. T., Nitsche-Schmitz, D. P., Nerlich, A., Chhatwal, G. S., and Bergmann, S. (2011a). Scm, a novel m-like protein from streptococcus canis, binds (mini)-plasminogen with high affinity and facilitates bacterial transmigration. *Biochemical Journal*, 434(3):523–535.

Fulde, M., Rohde, M., Polok, A., Preissner, K. T., Chhatwal, G. S., and Bergmann, S. (2013). Cooperative plasminogen recruitment to the surface of streptococcus canis via m protein and enolase enhances bacterial survival. *MBio*, 4(2).

Fulde, M. and Valentin-Weigand, P. (2012). Epidemiology and pathogenicity of zoonotic streptococci. *Host-pathogen interactions in streptococcal diseases*, pages 49–81.

Fulde, M., Willenborg, J., de Greeff, A., Benga, L., Smith, H. E., Valentin-Weigand, P., and Goethe, R. (2011b). Argr is an essential local transcriptional regulator of the arcabc operon in streptococcus suis and is crucial for biological fitness in an acidic environment. *Microbiology*, 157(2):572–582.

Galdas, P. M., Cheater, F., and Marshall, P. (2005). Men and health help-seeking behaviour: literature review. *Journal of advanced nursing*, 49(6):616–623.

Galpérine, T., Cazorla, C., Blanchard, E., Boineau, F., Ragnaud, J.-M., and Neau, D. (2007). Streptococcus canis infections in humans: retrospective study of 54 patients. *Journal of Infection*, 55(1):23–26.

García, A., Fox, J. G., and Besser, T. E. (2010). Zoonotic enterohemorrhagic escherichia coli: a one health perspective. *Ilar Journal*, 51(3):221–232.

García Fierro, R., Thomas-Lopez, D., Deserio, D., Liebana, E., Rizzi, V., and Guerra, B. (2018). Outcome of ec/efsa questionnaire (2016) on use of whole genome sequencing (wgs) for food-and waterborne pathogens isolated from animals, food, feed and related environmental samples in eu/efta countries. Technical report, Wiley Online Library.

Gardiner, D., Hartas, J., Currie, B., Mathews, J. D., Kemp, D. J., and Sriprakash, K. (1995). Vir typing: a long-pcr typing method for group a streptococci. *Genome Research*, 4(5):288–293.

Geldard, F. A. (1953). The human senses.

Gherardi, G., Vitali, L. A., and Creti, R. (2018). Prevalent emm types among invasive gas in europe and north america since year 2000. *Frontiers in public health*, 6:59.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–1455.

Gibbs, E. P. J. (2014). The evolution of one health: a decade of progress and challenges for the future. *Veterinary Record*, 174(4):85–91.

Giovanetti, E., Brenciani, A., Lupidi, R., Roberts, M. C., and Varaldo, P. E. (2003). Presence of the tet (o) gene in erythromycin-and tetracycline-resistant strains of streptococcus pyogenes and linkage with either the mef (a) or the erm (a) gene. *Antimicrobial agents and chemotherapy*, 47(9):2844–2849.

Gleicher, M., Correll, M., Nothelfer, C., and Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 19(12):2316–2325.

Goessens, W., Lemmens-den Toom, N., Hageman, J., Hermans, P., Sluijter, M., De Groot, R., and Verbrugh, H. (2000). Evaluation of the vitek 2 system for susceptibility testing of streptococcus pneumoniae isolates. *European Journal of Clinical Microbiology and Infectious Diseases*, 19(8):618–622.

Gossain, S., Chalker, V., Kapatai, G., Coelho, J., Mohamed, H., Martin, K., and Tahir, M. (2016). A protracted outbreak of invasive group a streptococcal infection at a uk long-term facility investigated using whole genome sequencing. In *Open Forum Infectious Diseases*, volume 3. Oxford University Press.

GOV.UK (2020). Group a streptococcal infections: activity during the 2019 to 2020 season. `https://www.gov.uk/government/publications/group-a-streptococcal-infections-activity-during-the-2019-to-2020`

Graham, M. R., Smoot, L. M., Migliaccio, C. A. L., Virtaneva, K., Sturdevant, D. E., Porcella, S. F., Federle, M. J., Adams, G. J., Scott, J. R., and Musser, J. M. (2002). Virulence control in group a streptococcus by a two-component gene regulatory system: global expression profiling and in vivo infection modeling. *Proceedings of the National Academy of Sciences*, 99(21):13855–13860.

Gramazio, C. C., Schloss, K. B., and Laidlaw, D. H. (2014). The relation between visualization size, grouping, and user performance. *IEEE transactions on visualization and computer graphics*, 20(12):1953–1962.

Grebe, T. and Hakenbeck, R. (1996). Penicillin-binding proteins 2b and 2x of streptococcus pneumoniae are primary resistance determinants for different classes of beta-lactam antibiotics. *Antimicrobial agents and chemotherapy*, 40(4):829–834.

Griffith, F. (1934). The serological classification of streptococcus pyogenes. *Epidemiology & Infection*, 34(4):542–584.

Gryllos, I., Levin, J. C., and Wessels, M. R. (2003). The csrr/csrs two-component system of group a streptococcus responds to environmental mg2+. *Proceedings of the National Academy of Sciences*, 100(7):4227–4232.

Guerrero, A., Stornelli, M. C., Jurado, S. B., Giacoboni, G., Sguazza, G. H., de la Sota, R. L., and Stornelli, M. A. (2018). Vaginal isolation of beta-haemolytic streptococcus from bitches with and without neonatal deaths in the litters. *Reproduction in domestic animals*, 53(3):609–616.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075.

Guy, R., Williams, C., Irvine, N., Reynolds, A., Coelho, J., Saliba, V., Thomas, D., Doherty, L., Chalker, V., Von Wissmann, B., et al. (2014). Increase in scarlet fever notifications in the united kingdom, 2013/2014. *Eurosurveillance*, 19(12):20749.

Haidan, A., Talay, S. R., Rohde, M., Sriprakash, K. S., Currie, B. J., and Chhatwal, G. S. (2000). Pharyngeal carriage of group c and group g streptococci and acute rheumatic fever in an aboriginal population. *The Lancet*, 356(9236):1167–1169.

Halperin, T., Levine, H., Korenman, Z., Burstein, S., Amber, R., Sela, T., and Valinsky, L. (2016). Molecular characterization and antibiotic resistance of group g streptococci in israel: comparison of invasive, non-invasive and carriage isolates. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(10):1649–1654.

Hamilton, S. M., Bayer, C. R., Stevens, D. L., Lieber, R. L., and Bryant, A. E. (2008). Muscle injury, vimentin expression, and nonsteroidal anti-inflammatory drugs predispose to cryptic group a streptococcal necrotizing infection. *The Journal of infectious diseases*, 198(11):1692–1698.

Hammond-Collins, K., Strauss, B., Barnes, K., Demczuk, W., Domingo, M.-C., Lamontagne, M.-C., Lu, D., Martin, I., and Tepper, M. (2019). Group a streptococcus outbreak in a canadian armed forces training facility. *Military medicine*, 184(3-4):e197–e204.

Hanage, W. P. and Shelburne III, S. A. (2020). Streptococcus pyogenes with reduced susceptibility to β-lactams: how big an alarm bell?

Harkness, G. A., Bentley, D. W., Mottley, M., and Lee, J. (1992). Streptococcus pyogenes outbreak in a long-term care facility. *American journal of infection control*, 20(3):142–148.

Harris, S. R., Cartwright, E. J., Török, M. E., Holden, M. T., Brown, N. M., Ogilvy-Stuart, A. L., Ellington, M. J., Quail, M. A., Bentley, S. D., Parkhill, J., et al. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant staphylococcus aureus: a descriptive study. *The Lancet infectious diseases*, 13(2):130–136.

Hassan, A. A., Akineden, Ö., and Usleber, E. (2005). Identification of streptococcus canis isolated from milk of dairy cows with subclinical mastitis. *Journal of clinical microbiology*, 43(3):1234–1238.

Hayes, A., Lacey, J. A., Morris, J. M., Davies, M. R., and Tong, S. Y. (2020). Restricted sequence variation in streptococcus pyogenes penicillin binding proteins. *Msphere*, 5(2):e00090–20.

Henningham, A., Barnett, T. C., Maamary, P. G., and Walker, M. J. (2012). Pathogenesis of group a streptococcal infections. *Discovery medicine*, 13(72):329–342.

Herrera, A. L., Huber, V. C., and Chaussee, M. S. (2016). The association between invasive group a streptococcal diseases and viral respiratory tract infections. *Frontiers in microbiology*, 7:342.

Hikone, M., Kobayashi, K.-i., Washino, T., Ota, M., Sakamoto, N., Iwabuchi, S., and Ohnishi, K. (2015). Streptococcal toxic shock syndrome secondary to group a streptococcus vaginitis. *Journal of Infection and Chemotherapy*, 21(12):873–876.

Hitzmann, A., Bergmann, S., Rohde, M., Chhatwal, G., and Fulde, M. (2013). Identification and characterization of the arginine deiminase system of streptococcus canis. *Veterinary microbiology*, 162(1):270–277.

Holden, M. T., Scott, A., Cherevach, I., Chillingworth, T., Churcher, C., Cronin, A., Dowd, L., Feltwell, T., Hamlin, N., Holroyd, S., et al. (2007). Complete genome of acute rheumatic fever-associated serotype m5 streptococcus pyogenes strain manfredo. *Journal of bacteriology*, 189(4):1473–1477.

Holmes, N. and Heller, S. (2006). *Nigel Holmes: on information design*. Jorge Pinto Books Inc.

Hondorp, E. R., Hou, S. C., Hause, L. L., Gera, K., Lee, C.-E., and McIver, K. S. (2013). Pts phosphorylation of mga modulates regulon expression and virulence in the group a streptococcus. *Molecular microbiology*, 88(6):1176–1193.

Hookey, J., Saunders, N., Clewley, J., Efstratiou, A., and George, R. (1996). Virulence regulon polymorphism in group a streptococci revealed by long pcr and implications for epidemiological and evolutionary studies. *Journal of medical microbiology*, 45(4):285–293.

Hsieh, Y.-C. and Huang, Y.-C. (2011). Scarlet fever outbreak in hong kong, 2011. *Journal of Microbiology, Immunology and Infection*, 44(6):409–411.

Hunt, M., Mather, A. E., Sánchez-Busó, L., Page, A. J., Parkhill, J., Keane, J. A., and Harris, S. R. (2017). Ariba: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial genomics*, 3(10).

Iglauer, F., Kunstỳr, I., Mörstedt, R., Farouq, H., Wullenweber, M., and Damsch, S. (1991). Streptococcus canis arthritis in a cat breeding colony. *Journal of experimental animal science*, 34(2):59–65.

Igwe, E. I., Shewmaker, P. L., Facklam, R. R., Farley, M. M., Van Beneden, C., and Beall, B. (2003). Identification of superantigen genes spem, ssa, and smez in invasive strains of beta-hemolytic group c and g streptococci recovered from humans. *FEMS microbiology letters*, 229(2):259–264.

Imöhl, M., Fitzner, C., Perniciaro, S., and van der Linden, M. (2017). Epidemiology and distribution of 10 superantigens among invasive streptococcus pyogenes disease in germany from 2009 to 2014. *PLoS One*, 12(7):e0180757.

Imöhl, M., Reinert, R. R., Ocklenburg, C., and Van Der Linden, M. (2010). Epidemiology of invasive streptococcus pyogenes disease in germany during 2003–2007. *FEMS Immunology & Medical Microbiology*, 58(3):389–396.

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J., and Holt, K. E. (2014). Srst2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11):90.

Ioannidis, Y. (2003). The history of histograms (abridged). In *Proceedings 2003 VLDB Conference*, pages 19–30. Elsevier.

Ivens, P., Matthews, D., Webb, K., Newton, J., Steward, K., Waller, A., Robinson, C., and Slater, J. (2011). Molecular characterisation of 'strangles' outbreaks in the uk: the use of m-protein typing of streptococcus equi ssp. equi. *Equine veterinary journal*, 43(3):359–364.

Jackson, S. J., Steer, A. C., and Campbell, H. (2011). Systematic review: estimation of global burden of non-suppurative sequelae of upper respiratory tract infection: rheumatic fever and post-streptococcal glomerulonephritis. *Tropical Medicine & International Health*, 16(1):2–11.

Jagadish, H. V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K. C., and Suel, T. (1998). Optimal histograms with quality guarantees. In *VLDB*, volume 98, pages 24–27.

Jayarao, B., Gillespie, B., Lewis, M., Dowlen, H., and Oliver, S. (1999). Epidemiology of streptococcus uberis intramammary infections in a dairy herd. *Journal of Veterinary Medicine, Series B*, 46(7):433–442.

Ji, Y., McLandsborough, L., Kondagunta, A., and Cleary, P. P. (1996). C5a peptidase alters clearance and trafficking of group a streptococci by infected mice. *Infection and immunity*, 64(2):503–510.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., et al. (2016). Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, page gkw1004.

Johnson, A. F. and LaRock, C. N. (2021). Antibiotic treatment, mechanisms for failure, and adjunctive therapies for infections by group a streptococcus. *Frontiers in Microbiology*, 12.

Johnson, D. R., Kaplan, E. L., Bicova, R., Havlicek, J., Havlickova, H., Kritz, P., Motlova, J., Sramek, J., Organization, W. H., et al. (1996). Laboratory diagnosis of group a streptococcal infections.

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. *Wellcome open research*, 3.

Jolley, K. A. and Maiden, M. C. (2010). Bigsdb: scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11(1):1–11.

Jørgensen, H., Nordstoga, A., Sviland, S., Zadoks, R., Sølverød, L., Kvitle, B., and Mørk, T. (2016). Streptococcus agalactiae in the environment of bovine dairy herds–rewriting the textbooks? *Veterinary microbiology*, 184:64–72.

Kachroo, P., Eraso, J. M., Beres, S. B., Olsen, R. J., Zhu, L., Nasser, W., Bernard, P. E., Cantu, C. C., Saavedra, M. O., Arredondo, M. J., et al. (2019). Integrated analysis of population genomics, transcriptomics and virulence provides novel insights into streptococcus pyogenes pathogenesis. *Nature genetics*, 51(3):548–559.

Kadioglu, A., Weiser, J. N., Paton, J. C., and Andrew, P. W. (2008). The role of streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nature Reviews Microbiology*, 6(4):288–301.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589.

Kaplan, E. L. (1980). The group a streptococcal upper respiratory tract carrier state: an enigma. *The Journal of pediatrics*, 97(3):337–345.

Karthikeyan, G. and Guilherme, L. (2018). Acute rheumatic fever. *The Lancet*, 392(10142):161–174.

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of dna sequences with mafft. In *Bioinformatics for DNA sequence analysis*, pages 39–64. Springer.

Kaufhold, A., Podbielski, A., Baumgarten, G., Blokpoel, M., Top, J., and Schouls, L. (1994). Rapid typing of group a streptococci by the use of dna amplification and non-radioactive allele-specific oligonucleotide probes. *FEMS microbiology letters*, 119(1-2):19–25.

Kaufhold, A., Podbielski, A., Johnson, D. R., Kaplan, E., and Lütticken, R. (1992). M protein gene typing of streptococcus pyogenes by nonradioactively labeled oligonucleotide probes. *Journal of clinical microbiology*, 30(9):2391–2397.

Kawabata, S., Tamura, Y., Murakami, J., Terao, Y., Nakagawa, I., and Hamada, S. (2002). A novel, anchorless streptococcal surface protein that binds to human immunoglobulins. *Biochemical and biophysical research communications*, 296(5):1329–1333.

Kelleher, C. and Wagener, T. (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software*, 26(6):822–827.

Kemble, S. K., Westbrook, A., Lynfield, R., Bogard, A., Koktavy, N., Gall, K., Lappi, V., DeVries, A. S., Kaplan, E., and Smith, K. E. (2013). Foodborne outbreak of group a streptococcus pharyngitis associated with a high school dance team banquet—minnesota, 2012. *Clinical infectious diseases*, 57(5):648–654.

Kittang, B., Bruun, T., Langeland, N., Mylvaganam, H., Glambek, M., and Skrede, S. (2011). Invasive group a, c and g streptococcal disease in western norway: virulence gene profiles, clinical features and outcomes. *Clinical Microbiology and Infection*, 17(3):358–364.

Knight Jr, V. J. (2012). *Iconographic method in New World prehistory*. Cambridge University Press.

Kreikemeyer, B., Boyle, M. D., Buttaro, B. A., Heinemann, M., and Podbielski, A. (2001). Group a streptococcal growth phase-associated virulence factor regulation by a novel operon (fas) with homologies to two-component-type regulators requires a small rna molecule. *Molecular microbiology*, 39(2):392–406.

Kreikemeyer, B., McIver, K. S., and Podbielski, A. (2003). Virulence factor regulation and regulatory networks in streptococcus pyogenes and their impact on pathogen–host interactions. *Trends in microbiology*, 11(5):224–232.

Kristian, S. A., Datta, V., Weidenmaier, C., Kansal, R., Fedtke, I., Peschel, A., Gallo, R. L., and Nizet, V. (2005). D-alanylation of teichoic acids promotes group a streptococcus antimicrobial peptide resistance, neutrophil survival, and epithelial cell invasion. *Journal of Bacteriology*, 187(19):6719–6725.

Król, J., Twardoń, J., Mrowiec, J., Podkowik, M., Dejneka, G., Dębski, B., Nowicki, T., and Zalewski, W. (2015). Streptococcus canis is able to establish a persistent udder infection in a dairy herd. *Journal of dairy science*, 98(10):7090–7096.

Kruger, E. F., Byrne, B. A., Pesavento, P., Hurley, K. F., Lindsay, L. L., and Sykes, J. E. (2010). Relationship between clinical manifestations and pulsed-field gel profiles of streptococcus canis isolates from dogs and cats. *Veterinary microbiology*, 146(1-2):167–171.

Krum, R. (2013). *Cool infographics: Effective communication with data visualization and design*. John Wiley & Sons.

Krzywinski, M. and Altman, N. (2014). Visualizing samples with box plots.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). Mega x: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35(6):1547–1549.

Lacave, G., Coutard, A., Troché, G., Augusto, S., Pons, S., Zuber, B., Laurent, V., Amara, M., Couzon, B., Bédos, J.-P., et al. (2016). Endocarditis caused by streptococcus canis: an emerging zoonosis? *Infection*, 44(1):111–114.

Lagace-Wiens, P. R., Rubinstein, E., and Gumel, A. (2010). Influenza epidemiology—past, present, and future. *Critical care medicine*, 38:e1–e9.

Lam, M. M., Clarridge, J. E., Young, E., and Mizuki, S. (2007). The other group g streptococcus: increased detection of streptococcus canis ulcer infections in dog owners. *Journal of clinical microbiology*, 45(7):2327–2329.

Lamagni, T., Darenberg, J., Luca-Harari, B., Siljander, T., Efstratiou, A., Henriques-Normark, B., Vuopio-Varkila, J., Bouvet, A., Creti, R., Ekelund, K., et al. (2008a). Epidemiology of severe streptococcus pyogenes disease in europe. *Journal of clinical microbiology*, 46(7):2359–2367.

Lamagni, T., Efstratiou, A., Vuopio-Varkila, J., Jasir, A., and Schalen, C. (2005). The epidemiology of severe streptococcus pyogenes associated disease in europe. *Euro surveillance: bulletin Europeen sur les maladies transmissibles= European communicable disease bulletin*, 10(9):9–10.

Lamagni, T., Neal, S., Keshishian, C., Alhaddad, N., George, R., Duckworth, G., Vuopio-Varkila, J., and Efstratiou, A. (2008b). Severe streptococcus pyogenes infections, united kingdom, 2003–2004. *Emerging infectious diseases*, 14(2):202.

Lamagni, T., Neal, S., Keshishian, C., Hope, V., George, R., Duckworth, G., Vuopio-Varkila, J., and Efstratiou, A. (2008c). Epidemic of severe streptococcus pyogenes infections in injecting drug users in the uk, 2003–2004. *Clinical microbiology and infection*, 14(11):1002–1009.

Lamagni, T., Tyrrell, G., Lovgren, M., Siljander, T., Lyytikäinen, O., Vuopio-Varkila, J., Van Beneden, C., Martin, D., and Efstratiou, A. (2009). Seasonal patterns of invasive streptococcus pyogenes disease in the northern hemisphere: P1534. *Clinical Microbiology & Infection*, 15(4).

Lamb, L. E., Sriskandan, S., and Tan, L. K. (2015). Bromine, bear-claw scratch fasciotomies, and the eagle effect: management of group a streptococcal necrotising fasciitis and its association with trauma. *The Lancet Infectious Diseases*, 15(1):109–121.

Lamm, C. G., Ferguson, A., Lehenbauer, T. W., and Love, B. (2010). Streptococcal infection in dogs: a retrospective study of 393 cases. *Veterinary pathology*, 47(3):387–395.

Lancefield, R. C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *Journal of experimental medicine*, 57(4):571–595.

Lancefield, R. C. (1962). Current knowledge of type-specific m antigens of group a streptococci. *The Journal of Immunology*, 89(3):307–313.

Lappin, E. and Ferguson, A. J. (2009). Gram-positive toxic shock syndromes. *The Lancet infectious diseases*, 9(5):281–290.

Laupland, K. B., Ross, T., Church, D., and Gregson, D. (2006). Population-based surveillance of invasive pyogenic streptococcal infection in a large canadian region. *Clinical microbiology and infection*, 12(3):224–230.

Le, V. T. M. and Diep, B. A. (2013). Selected insights from application of whole genome sequencing for outbreak investigations. *Current opinion in critical care*, 19(5):432.

Le Breton, Y., Belew, A. T., Valdes, K. M., Islam, E., Curry, P., Tettelin, H., Shirtliff, M. E., El-Sayed, N. M., and McIver, K. S. (2015). Essential genes in the core genome of the human pathogen streptococcus pyogenes. *Scientific reports*, 5(1):1–13.

Lederman, Z., Leskes, H., and Brosh-Nissimov, T. (2020). One health and streptococcus canis in the emergency department: A case of cellulitis and bacteremia in an immunocompromised patient treated with etanercept. *The Journal of emergency medicine*, 58(3):e129–e132.

Lee, I. P. A. and Andam, C. P. (2022). Frequencies and characteristics of genome-wide recombination in streptococcus agalactiae, streptococcus pyogenes, and streptococcus suis. *Scientific reports*, 12(1):1–11.

Lefébure, T., Richards, V. P., Lang, P., Pavinski-Bitar, P., and Stanhope, M. J. (2012). Gene repertoire evolution of streptococcus pyogenes inferred from phylogenomic analysis with streptococcus canis and streptococcus dysgalactiae. *PloS one*, 7(5):e37607.

Lei, B., DeLeo, F. R., Hoe, N. P., Graham, M. R., Mackie, S. M., Cole, R. L., Liu, M., Hill, H. R., Low, D. E., Federle, M. J., et al. (2001). Evasion of human innate and acquired immunity by a bacterial homolog of cd11b that inhibits opsonophagocytosis. *Nature medicine*, 7(12):1298.

Leonard, A., Wright, A., Saavedra-Campos, M., Lamagni, T., Cordery, R., Nicholls, M., Domoney, C., Sriskandan, S., and Balasegaram, S. (2019). Severe group a streptococcal infections in mothers and their newborns in london and the south east, 2010–2016: assessment of risk and audit of public health management. *BJOG: An International Journal of Obstetrics & Gynaecology*, 126(1):44–53.

Lepoutre, A., Doloy, A., Bidet, P., Leblond, A., Perrocheau, A., Bingen, E., Trieu-Cuot, P., Bouvet, A., Poyart, C., Lévy-Bruhl, D., et al. (2011). Epidemiology of invasive streptococcus pyogenes infections in france in 2007. *Journal of clinical microbiology*, 49(12):4094–4100.

Letunic, I. and Bork, P. (2021). Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, 49(W1):W293–W296.

Levy, M., Johnson, C. G., and Kraa, E. (2003). Tonsillopharyngitis caused by foodborne group a streptococcus: a prison-based outbreak. *Clinical infectious diseases*, 36(2):175–182.

Ligozzi, M., Bernini, C., Bonora, M. G., De Fatima, M., Zuliani, J., and Fontana, R. (2002). Evaluation of the vitek 2 system for identification and antimicrobial susceptibility testing of medically relevant gram-positive cocci. *Journal of clinical microbiology*, 40(5):1681–1686.

Lim, S. C., Knight, D., and Riley, T. V. (2019). Clostridium difficile and one health. *Clinical Microbiology and Infection*.

Lindsay, D., Brown, A., Scott, K., Denham, B., Thom, L., Rundell, G., Ure, R., Jones, B., and Smith, A. (2016). Circulating emm types of streptococcus pyogenes in scotland: 2011-2015. *Journal of medical microbiology*, 65(10):1229.

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). Vfdb 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic acids research*, 47(D1):D687–D692.

Loof, T. G., Deicke, C., and Medina, E. (2014). The role of coagulation/fibrinolysis during streptococcus pyogenes infection. *Frontiers in cellular and infection microbiology*, 4:128.

Lopez, A. D. (2010). Sharing data for public health: where is the vision? *Bulletin of the World Health Organization*, 88(6):467–467.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21.

Luca-Harari, B., Darenberg, J., Neal, S., Siljander, T., Strakova, L., Tanna, A., Creti, R., Ekelund, K., Koliou, M., Tassios, P. T., et al. (2009). Clinical and microbiological characteristics of severe streptococcus pyogenes disease in europe. *Journal of clinical microbiology*, 47(4):1155–1165.

Luca-Harari, B., Ekelund, K., Van Der Linden, M., Staum-Kaltoft, M., Hammerum, A. M., and Jasir, A. (2008). Clinical and epidemiological aspects of invasive streptococcus pyogenes infections in denmark during 2003 and 2004. *Journal of clinical microbiology*, 46(1):79–86.

Luck, S. J., Hollingworth, A., et al. (2008). *Visual memory*. OUP USA.

Lukomski, S., Nakashima, K., Abdi, I., Cipriano, V. J., Shelvin, B. J., Graviss, E. A., and Musser, J. M. (2001). Identification and characterization of a second extracellular collagen-like protein made by group astreptococcus: Control of production at the level of translation. *Infection and immunity*, 69(3):1729–1738.

Lynskey, N. N., Jauneikaite, E., Li, H. K., Zhi, X., Turner, C. E., Mosavie, M., Pearson, M., Asai, M., Lobkowicz, L., Chow, J. Y., et al. (2019). Emergence of dominant toxigenic m1t1 streptococcus pyogenes clone during increased scarlet fever activity in england: a population-based molecular epidemiological study. *The Lancet Infectious Diseases*, 19(11):1209–1218.

Lyon, W. R., Gibson, C. M., and Caparon, M. G. (1998). A role for trigger factor and an rgg-like regulator in the transcription, secretion and processing of the cysteine proteinase of streptococcus pyogenes. *The EMBO journal*, 17(21):6263–6275.

Lysková, P., Vydržalová, M., Královcová, D., Mazurová, J., et al. (2007a). Identification and antimicrobial susceptibility of bacteria and yeasts isolated from healthy dogs and dogs with otitis externa. *Journal of Veterinary Medicine Series A*, 54(10):559–563.

Lysková, P., Vydržalová, M., Královcová, D., Mazurová, J., et al. (2007b). Prevalence and characteristics of streptococcus canis strains isolated from dogs and cats. *Acta Veterinaria Brno*, 76(4):619–625.

Maddocks, S. E., Wright, C. J., Nobbs, A. H., Brittan, J. L., Franklin, L., Strömberg, N., Kadioglu, A., Jepson, M. A., and Jenkinson, H. F. (2011). Streptococcus pyogenes antigen i/ii-family polypeptide aspa shows differential ligand-binding properties and mediates biofilm formation. *Molecular microbiology*, 81(4):1034–1049.

Magee, J., Hindmarch, J., Burnett, I., and Pease, A. (1989). Epidemiological typing of streptococcus pyogenes by pyrolysis mass spectrometry. *Journal of medical microbiology*, 30(4):273–278.

Magee, J., Hindmarch, J., and Nicol, C. (1991). Typing of streptococcus pyogenes by pyrolysis mass spectrometry. *Journal of medical microbiology*, 35(5):304–306.

Magiorakos, A.-P., Srinivasan, A., Carey, R., Carmeli, Y., Falagas, M., Giske, C., Harbarth, S., Hindler, J., Kahlmeter, G., Olsson-Liljequist, B., et al. (2012). Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clinical microbiology and infection*, 18(3):268–281.

Mahida, N., Beal, A., Trigg, D., Vaughan, N., and Boswell, T. (2014). Outbreak of invasive group a streptococcus infection: contaminated patient curtains and cross-infection on an ear, nose and throat ward. *Journal of Hospital Infection*, 87(3):141–144.

Mahmmod, Y., Klaas, I., Katholm, J., Lutton, M., and Zadoks, R. (2015). Molecular epidemiology and strain-specific characteristics of streptococcus agalactiae at the herd and cow level. *Journal of dairy science*, 98(10):6913–6924.

Malbruny, B., Werno, A. M., Murdoch, D. R., Leclercq, R., and Cattoir, V. (2011). Cross-resistance to lincosamides, streptogramins a, and pleuromutilins due to the lsa (c) gene in streptococcus agalactiae ucn70. *Antimicrobial Agents and Chemotherapy*, 55(4):1470–1474.

Mališová, B., Šantavỳ, P., Lovečková, Y., Hladkỳ, B., Kotásková, I., Pol, J., Lonskỳ, V., Němec, P., and Freiberger, T. (2019). Human native endocarditis caused by streptococcus canis—a case report. *Apmis*, 127(1):41–44.

Manetti, A. G., Zingaretti, C., Falugi, F., Capo, S., Bombaci, M., Bagnoli, F., Gambellini, G., Bensi, G., Mora, M., Edwards, A. M., et al. (2007). Streptococcus pyogenes pili promote pharyngeal cell adhesion and biofilm formation. *Molecular microbiology*, 64(4):968–983.

Margolis, E., Yates, A., and Levin, B. R. (2010). The ecology of nasal colonization of streptococcus pneumoniae, haemophilus influenzae and staphylococcus aureus: the role of competition and interactions with host's immune response. *BMC microbiology*, 10(1):1–11.

Marks, L. R., Reddinger, R. M., and Hakansson, A. P. (2014). Biofilm formation enhances fomite survival of streptococcus pneumoniae and streptococcus pyogenes. *Infection and immunity*, 82(3):1141–1146.

Marshall, C. S., Cheng, A. C., Markey, P. G., Towers, R. J., Richardson, L. J., Fagan, P. K., Scott, L., Krause, V. L., and Currie, B. J. (2011). Acute post-streptococcal glomerulonephritis in the northern territory of australia: a review of 16 years data and comparison with the literature. *The American journal of tropical medicine and hygiene*, 85(4):703–710.

Martin, J. M., Green, M., Barbadora, K. A., and Wald, E. R. (2004). Group a streptococci among school-aged children: clinical characteristics and the carrier state. *Pediatrics*, 114(5):1212–1219.

Matsuu, A., Kanda, T., Sugiyama, A., Murase, T., and Hikasa, Y. (2007). Mitral stenosis with bacterial myocarditis in a cat. *Journal of Veterinary Medical Science*, 69(11):1171–1174.

Mayorga, A. and Gleicher, M. (2013). Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, 19(9):1526–1538.

McCrorie, A., Donnelly, C., and McGlade, K. (2016). Infographics: healthcare communication for the digital age. *The Ulster medical journal*, 85(2):71.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16.

McGuire, A., Krysa, N., and Mann, S. (2021). Hair of the dog? periprosthetic joint infection with streptococcus canis. *Arthroplasty today*, 8:53–56.

McIver, K. S. (2009). Stand-alone response regulators controlling global virulence networks in streptococcus pyogenes. In *Bacterial Sensing and Signaling*, volume 16, pages 103–119. Karger Publishers.

McIver, K. S. and Scott, J. R. (1997). Role of mga in growth phase regulation of virulence genes of the group a streptococcus. *Journal of Bacteriology*, 179(16):5178–5187.

McNamara, D. R., Tleyjeh, I. M., Berbari, E. F., Lahr, B. D., Martinez, J. W., Mirzoyev, S. A., and Baddour, L. M. (2007). Incidence of lower-extremity cellulitis: a population-based study in olmsted county, minnesota. In *Mayo Clinic Proceedings*, volume 82, pages 817–821. Elsevier.

Medina, E., Schulze, K., Chhatwal, G. S., and Guzmán, C. A. (2000). Nonimmune interaction of the sfbi protein of streptococcus pyogenes with the immunoglobulin gf (ab') 2fragment. *Infection and immunity*, 68(8):4786–4788.

Microsoft Corporation (2021). Microsoft excel.

Miyoshi-Akiyama, T., Takamatsu, D., Koyanagi, M., Zhao, J., Imanishi, K., and Uchiyama, T. (2005). Cytocidal effect of streptococcus pyogenes on mouse neutrophils in vivo and the critical role of streptolysin s. *The Journal of infectious diseases*, 192(1):107–116.

Molesworth, A. M., Cuevas, L. E., Connor, S. J., Morse, A. P., and Thomson, M. C. (2003). Environmental risk and meningitis epidemics in africa. *Emerging infectious diseases*, 9(10):1287.

Molinari, G., Rohde, M., Talay, S. R., Chhatwal, G. S., Beckert, S., and Podbielski, A. (2001). The role played by the group a streptococcal negative regulator nra on bacterial interactions with epithelial cells. *Molecular microbiology*, 40(1):99–114.

Monto, A. S. (2002). Epidemiology of viral respiratory infections. *The American journal of medicine*, 112(6):4–12.

Morens, D. M., Taubenberger, J. K., and Fauci, A. S. (2008). Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness. *The Journal of infectious diseases*, 198(7):962–970.

Morse, S. A., Johnson, S., Biddle, J., and Roberts, M. (1986). High-level tetracycline resistance in neisseria gonorrhoeae is result of acquisition of streptococcal tetm determinant. *Antimicrobial agents and chemotherapy*, 30(5):664–670.

Morse, S. S. (2007). Global infectious disease surveillance and health intelligence. *Health Affairs*, 26(4):1069–1077.

Mundt, J. O. (1982). The ecology of the streptococci. *Microbial ecology*, 8(4):355–369.

Musser, J. M., Beres, S. B., Zhu, L., Olsen, R. J., Vuopio, J., Hyyryläinen, H.-L., Gröndahl-Yli-Hannuksela, K., Kristinsson, K. G., Darenberg, J., Henriques-Normark, B., et al. (2020). Reduced in vitro susceptibility of streptococcus pyogenes to β-lactam antibiotics associated with mutations in the pbp2x gene is geographically widespread. *Journal of Clinical Microbiology*, 58(4).

Musser, J. M., Gray, B., Schlievert, P., and Pichichero, M. (1992). Streptococcus pyogenes pharyngitis: characterization of strains by multilocus enzyme genotype, m and t protein serotype, and pyrogenic exotoxin gene probing. *Journal of clinical microbiology*, 30(3):600–603.

Musser, J. M. and Shelburne, S. A. (2009). A decade of molecular pathogenomic analysis of group a streptococcus. *The Journal of clinical investigation*, 119(9):2455–2463.

Nakagawa, I., Kurokawa, K., Yamashita, A., Nakata, M., Tomiyasu, Y., Okahashi, N., Kawabata, S., Yamazaki, K., Shiba, T., Yasunaga, T., et al. (2003). Genome sequence of an m3 strain of streptococcus pyogenes reveals a large-scale genomic rearrangement in invasive strains and new insights into phage evolution. *Genome research*, 13(6a):1042–1055.

Nasser, W., Beres, S. B., Olsen, R. J., Dean, M. A., Rice, K. A., Long, S. W., Kristinsson, K. G., Gottfredsson, M., Vuopio, J., Raisanen, K., et al. (2014). Evolutionary pathway to increased virulence and epidemic group a streptococcus disease derived from 3,615 genome sequences. *Proceedings of the National Academy of Sciences*, 111(17):E1768–E1776.

Navarro, V. J., Axelrod, P. I., Pinover, W., Hockfield, H. S., and Kostman, J. R. (1993). A comparison of streptococcus pyogenes (group a streptococcal) bacteremia at an urban

and a suburban hospital: the importance of intravenous drug use. *Archives of internal medicine*, 153(23):2679–2684.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274.

Nielsen, H. U. K., Hammerum, A. M., Ekelund, K., Bang, D., Pallesen, L. V., and Frimodt-Møller, N. (2004). Tetracycline and macrolide co-resistance in streptococcus pyogenes: co-selection as a reason for increase in macrolide-resistant s. pyogenes? *Microbial drug resistance*, 10(3):231–238.

Numberger, D., Siebert, U., Fulde, M., and Valentin-Weigand, P. (2021). Streptococcal infections in marine mammals. *Microorganisms*, 9(2):350.

Nyberg, P., Rasmussen, M., and Björck, L. (2004). α2-macroglobulin-proteinase complexes protect streptococcus pyogenes from killing by the antimicrobial peptide ll-37. *Journal of Biological Chemistry*, 279(51):52820–52823.

O'Connor, S., Waite, M., Duce, D., O'Donnell, A., and Ronquillo, C. (2020). Data visualization in health care: The florence effect.

Ohtaki, H., Ohkusu, K., Ohta, H., Miyazaki, T., Yonetamari, J., Usui, T., Mori, I., Ito, H., Ishizuka, T., and Seishima, M. (2013). A case of sepsis caused by streptococcus canis in a dog owner: a first case report of sepsis without dog bite in japan. *Journal of Infection and Chemotherapy*, 19(6):1206–1209.

Olafsdottir, L., Erlendsdóttir, H., Melo-Cristino, J., Weinberger, D., Ramirez, M., Kristinsson, K., and Gottfredsson, M. (2014). Invasive infections due to streptococcus pyogenes: seasonal variation of severity and clinical characteristics, iceland, 1975 to 2012. *Eurosurveillance*, 19(17):20784.

O'Loughlin, R. E., Roberson, A., Cieslak, P. R., Lynfield, R., Gershman, K., Craig, A., Albanese, B. A., Farley, M. M., Barrett, N. L., Spina, N. L., et al. (2007). The epidemiology of invasive group a streptococcal infection and potential vaccine implications: United states, 2000–2004. *Clinical Infectious Diseases*, 45(7):853–862.

Österlund, A., Popa, R., Nikkilä, T., Scheynius, A., and Engstrand, L. (1997). Intracellular reservoir of streptococcus pyogenes in vivo: a possible explanation for recurrent pharyngotonsillitis. *The Laryngoscope*, 107(5):640–647.

Otten, J. J., Cheng, K., and Drewnowski, A. (2015). Infographics and public policy: using data visualization to convey complex information. *Health Affairs*, 34(11):1901–1907.

Pagnossin, D., Smith, A., Oravcova, K., and Weir, W. (2022). Streptococcus canis, the underdog of the genus. *Veterinary Microbiology*, page 109524.

Pagnossin, D., Smith, A., Weir, W., Crestani, C., Lindsay, D., Ure, R., and Oravcova, K. (2021). Complete genome sequences of three invasive strains of streptococcus pyogenes subtype emm 5.23 isolated in scotland. *Microbiology Resource Announcements*, 10(15):e00101–21.

Park, S., Bekemeier, B., Flaxman, A., and Schultz, M. (2022). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*, 47(2):175–193.

Parkinson, N., Robin, C., Newton, J., Slater, J., and Waller, A. (2011). Molecular epidemiology of strangles outbreaks in the uk during 2010. *Veterinary Record*, 168(25):666–666.

Parnaby, M. G. and Carapetis, J. R. (2010). Rheumatic fever in indigenous australian children. *Journal of paediatrics and child health*, 46(9):527–533.

Pastural, É., McNeil, S. A., MacKinnon-Cameron, D., Ye, L., Langley, J. M., Stewart, R., Martin, L. H., Hurley, G. J., Salehi, S., Penfound, T. A., et al. (2020). Safety and immunogenicity of a 30-valent m protein-based group a streptococcal vaccine in healthy adult volunteers: A randomized, controlled phase i study. *Vaccine*, 38(6):1384–1392.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1):3–1.

Pesavento, P., Bannasch, M., Bachmann, R., Byrne, B. A., and Hurley, K. (2007). Fatal streptococcus canis infections in intensively housed shelter cats. *Veterinary Pathology*, 44(2):218–221.

Pesola, A., Sihvonen, R., Lindholm, L., and Pätäri-Sampo, A. (2015). Clindamycin resistant emm33 streptococcus pyogenes emerged among invasive infections in helsinki metropolitan area, finland, 2012 to 2013. *Eurosurveillance*, 20(18):21117.

Pierce, R. and Chick, H. (2013). Workplace statistical literacy for teachers: Interpreting box plots. *Mathematics Education Research Journal*, 25(2):189–205.

Pightling, A. W., Pettengill, J. B., Luo, Y., Baugher, J. D., Rand, H., and Strain, E. (2018). Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. *Frontiers in microbiology*, 9:1482.

Pigott, D. M., Deshpande, A., Letourneau, I., Morozoff, C., Reiner Jr, R. C., Kraemer, M. U., Brent, S. E., Bogoch, I. I., Khan, K., Biehl, M. H., et al. (2017). Local, national, and regional viral haemorrhagic fever pandemic potential in africa: a multistage analysis. *The Lancet*, 390(10113):2662–2672.

Pigott, D. M., Golding, N., Mylne, A., Huang, Z., Henry, A. J., Weiss, D. J., Brady, O. J., Kraemer, M. U., Smith, D. L., Moyes, C. L., et al. (2014). Mapping the zoonotic niche of ebola virus disease in africa. *Elife*, 3:e04395.

Pinho, M., Foster, G., Pomba, C., Machado, M., Baily, J., Kuiken, T., Melo-Cristino, J., Ramirez, M., Vaz, T., Giao, M., et al. (2019). Streptococcus canis are a single population infecting multiple animal hosts despite the diversity of the universally present m-like protein scm. *Frontiers in microbiology*, 10:631.

Pinho, M., Matos, S., Pomba, C., Lübke-Becker, A., Wieler, L., Preziuso, S., Melo-Cristino, J., and Ramirez, M. (2013). Multilocus sequence analysis of streptococcus canis confirms the zoonotic origin of human infections and reveals genetic exchange with streptococcus dysgalactiae subsp. equisimilis. *Journal of clinical microbiology*, 51(4):1099–1109.

Pisani, E. and AbouZahr, C. (2010). Sharing health data: good intentions are not enough. *Bulletin of the World Health Organization*, 88:462–466.

Playfair, W. (1801a). *The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure and debts of england during the whole of the eighteenth century*. T. Burton.

Playfair, W. (1801b). *The statistical breviary*. Wallis.

Podbielski, A., Woischnik, M., Leonard, B. A., and Schmidt, K.-H. (1999). Characterization of nra, a global negative regulator gene in group a streptococci. *Molecular microbiology*, 31(4):1051–1064.

Polonsky, J. A., Baidjoe, A., Kamvar, Z. N., Cori, A., Durski, K., Edmunds, W. J., Eggo, R. M., Funk, S., Kaiser, L., Keating, P., et al. (2019). Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B*, 374(1776):20180276.

Porcasi, X., Rotela, C. H., Introini, M. V., Frutos, N., Lanfri, S., Peralta, G., De Elia, E. A., Lanfri, M. A., and Scavuzzo, C. M. (2012). An operative dengue risk stratification system in argentina based on geospatial technology. *Geospatial health*, pages S31–S42.

Prescott, J., Mathews, K., Gyles, C., Matsumiya, L., Miller, C., Rinkhardt, N., Yager, J., Hylands, R., and Low, D. (1995). Canine streptococcal toxic shock syndrome in ontario: An emerging disease? *The Canadian Veterinary Journal*, 36(8):486.

Putman, M., van Veen, H. W., Degener, J. E., and Konings, W. N. (2001). The lactococcal secondary multidrug transporter lmrp confers resistance to lincosamides, macrolides, streptogramins and tetracyclines. *Microbiology*, 147(10):2873–2880.

Quainoo, S., Coolen, J. P., van Hijum, S. A., Huynen, M. A., Melchers, W. J., van Schaik, W., and Wertheim, H. F. (2017). Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical microbiology reviews*, 30(4):1015–1063.

Rabinowitz, P. M., Kock, R., Kachani, M., Kunkel, R., Thomas, J., Gilbert, J., Wallace, R., Blackmore, C., Wong, D., Karesh, W., et al. (2013). Toward proof of concept of a one health approach to disease prediction and control. *Emerging Infectious Diseases*, 19(12).

Radford, A., Tierney, A., Coyne, K., Gaskell, R., Noble, P., Dawson, S., Setzkorn, C., Jones, P., Buchan, I., Newton, J., et al. (2010). Surveillance: Developing a network for small animal disease surveillance. *Veterinary Record*, 167(13):472–474.

Ralph, A. P. and Carapetis, J. R. (2012). Group a streptococcal diseases and their global burden. In *Host-pathogen interactions in streptococcal diseases*, pages 1–27. Springer.

Rato, M. G., Bexiga, R., Florindo, C., Cavaco, L. M., Vilela, C. L., and Santos-Sanches, I. (2013). Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. *Veterinary microbiology*, 161(3-4):286–294.

Reinhardt, M., Elias, J., Albert, J., Frosch, M., Harmsen, D., and Vogel, U. (2008). Episcangis: an online geographic surveillance system for meningococcal disease. *International journal of health geographics*, 7(1):1–7.

Revell, L. J. (2012). phytools: an r package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, (2):217–223.

Revez, J., Espinosa, L., Albiger, B., Leitmeyer, K. C., Struelens, M. J., Points, E. N. M. F., and Group, E. (2017). Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of european national capacities, 2015–2016. *Frontiers in public health*, 5:347.

Ribardo, D. A. and McIver, K. S. (2006). Defining the mga regulon: comparative transcriptome analysis reveals both direct and indirect regulation by mga in the group a streptococcus. *Molecular microbiology*, 62(2):491–508.

Richards, V. P., Zadoks, R. N., Bitar, P. D. P., Lefébure, T., Lang, P., Werner, B., Tikofsky, L., Moroni, P., and Stanhope, M. J. (2012). Genome characterization and population genetic structure of the zoonotic pathogen, streptococcus canis. *BMC microbiology*, 12(1):1–16.

Richter, S. S., Diekema, D. J., Heilmann, K. P., Almer, L. S., Shortridge, V. D., Zeitler, R., Flamm, R. K., and Doern, G. V. (2003). Fluoroquinolone resistance in streptococcus pyogenes. *Clinical infectious diseases*, 36(3):380–383.

Rivera-Hernandez, T., Carnathan, D. G., Jones, S., Cork, A. J., Davies, M. R., Moyle, P. M., Toth, I., Batzloff, M. R., McCarthy, J., Nizet, V., et al. (2019). An experimental group a streptococcus vaccine that reduces pharyngitis and tonsillitis in a nonhuman primate model. *MBio*, 10(2):e00693–19.

Rivera-Hernandez, T., Pandey, M., Henningham, A., Cole, J., Choudhury, B., Cork, A. J., Gillen, C. M., Ghaffar, K. A., West, N. P., Silvestri, G., et al. (2016). Differing efficacies

of lead group a streptococcal vaccine candidates and full-length m protein in cutaneous and invasive disease models. *MBio*, 7(3):e00618–16.

Robbins, N. B. (2012). *Creating more effective graphs*. Wiley.

Roberts, M. C. (2005). Update on acquired tetracycline resistance genes. *FEMS microbiology letters*, 245(2):195–203.

Robinson, T. P., Bu, D., Carrique-Mas, J., Fèvre, E. M., Gilbert, M., Grace, D., Hay, S. I., Jiwakanon, J., Kakkar, M., Kariuki, S., et al. (2016). Antibiotic resistance is the quintessential one health issue. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 110(7):377–380.

Rosenberg, E. S., Grey, J. A., Sanchez, T. H., and Sullivan, P. S. (2016). Rates of prevalent hiv infection, prevalent diagnoses, and new diagnoses among men who have sex with men in us states, metropolitan statistical areas, and counties, 2012-2013. *JMIR public health and surveillance*, 2(1):e22.

Röttig, A. and Steinbüchel, A. (2013). Acyltransferases in bacteria. *Microbiol. Mol. Biol. Rev.*, 77(2):277–321.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Ruben, F. L., NORDEN, C. W., HEISLER, B., and KORICA, Y. (1984). An outbreak of streptococcus pyogenes infections in a nursing home. *Annals of internal medicine*, 101(4):494–496.

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., and Perry, R. (2016). Data visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, 2(12):11–16.

Salipante, S. J., SenGupta, D. J., Cummings, L. A., Land, T. A., Hoogestraat, D. R., and Cookson, B. T. (2015). Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *Journal of clinical microbiology*, 53(4):1072–1079.

Sanyahumbi, A. S., Colquhoun, S., Wyber, R., and Carapetis, J. R. (2016). Global disease burden of group a streptococcus. In *Streptococcus pyogenes: basic biology to clinical manifestations [Internet]*. University of Oklahoma Health Sciences Center.

Sarikaya, A. and Gleicher, M. (2017). Scatterplots: Tasks, data, and designs. *IEEE transactions on visualization and computer graphics*, 24(1):402–412.

Schar, D., Klein, E. Y., Laxminarayan, R., Gilbert, M., and Van Boeckel, T. P. (2020). Global trends in antimicrobial use in aquaculture. *Scientific reports*, 10(1):1–9.

Scheiner, C. (1630). *Rosa ursina, sive Sol ex admirando facularum et macularum suarum phoenomeno varius, necnon circa centrum suum et axem fixum ab occasu in ortum annua, circaq. alium axem mobilem ab ortu in occasum conversine quasi menstrua, super polos proprios, libris IV mobilis ostensus*. A. Phaeus.

Schleifer, K. and Kilpper-Bälz, R. (1987). Molecular and chemotaxonomic approaches to the classification of streptococci, enterococci and lactococci: a review. *Systematic and Applied Microbiology*, 10(1):1–19.

Schmidt, K.-H., Mann, K., Cooney, J., and Köhler, W. (1993). Multiple binding of type 3 streptococcal m protein to human fibrinogen, albumin and fibronectin. *FEMS Immunology & Medical Microbiology*, 7(2):135–143.

Schürch, A., Arredondo-Alonso, S., Willems, R., and Goering, R. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clinical Microbiology and Infection*, 24(4):350–354.

Schwalbe, N., Wahl, B., Song, J., and Lehtimaki, S. (2020). Data sharing and global public health: defining what we mean by data. *Frontiers in Digital Health*, 2:612339.

Schwarze, K., Buchanan, J., Taylor, J. C., and Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature. *Genetics in Medicine*, 20(10):1122–1130.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Seemann, T. (2015). snippy: fast bacterial variant calling from ngs reads. `https://github.com/tseemann/snippy`.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Applied and environmental microbiology*, 51(5):873.

Severiano, A., Pinto, F. R., Ramirez, M., and Carriço, J. A. (2011). Adjusted wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, 49(11):3997–4000.

Sherwood, E., Vergnano, S., Kakuchi, I., Bruce, M. G., Chaurasia, S., David, S., Dramowski, A., Georges, S., Guy, R., Lamagni, T., et al. (2022). Invasive group a streptococcal disease in pregnant women and young children: a systematic review and meta-analysis. *The Lancet Infectious Diseases*.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic biology*, 51(3):492–508.

Shin, B. and Park, W. (2018). Zoonotic diseases and phytochemical medicines for microbial infections in veterinary science: Current state and future perspective. *Frontiers in veterinary science*, 5:166.

Siemens, N., Patenge, N., Otto, J., Fiedler, T., and Kreikemeyer, B. (2011). Streptococcus pyogenes m49 plasminogen/plasmin binding facilitates keratinocyte invasion via integrin-integrin-linked kinase (ilk) pathways and protects from macrophage killing. *Journal of Biological Chemistry*, 286(24):21612–21622.

Siirtola, H. (2019). The cost of pie charts. In *2019 23rd International Conference Information Visualisation (IV)*, pages 151–156. IEEE.

Silley, P., Simjee, S., and Schwarz, S. (2012). Surveillance and monitoring of antimicrobial resistance and antibiotic consumption in humans and animals. *Revue scientifique et technique (International Office of Epizootics)*, 31(1):105–120.

Silva-Costa, C., Carriço, J. A., Ramirez, M., and Melo-Cristino, J. (2014). Scarlet fever is caused by a limited number of streptococcus pyogenes lineages and is associated with the exotoxin genes ssa, spea and spec. *The Pediatric infectious disease journal*, 33(3):306–310.

Silva-Costa, C., Friaes, A., Ramirez, M., and Melo-Cristino, J. (2015). Macrolide-resistant streptococcus pyogenes: prevalence and treatment strategies. *Expert review of anti-infective therapy*, 13(5):615–628.

Simpson, E. H. (1949). Measurement of diversity. *nature*, 163(4148):688–688.

Single, L. A. and Martin, D. R. (1992). Clonal differences within m-types of the group a streptococcus revealed by pulsed field gel electrophoresis. *FEMS microbiology letters*, 91(1):85–89.

Sirén, J., Välimäki, N., and Mäkinen, V. (2014). Hisat2-fast and sensitive alignment against general human population. *IEEE/ACM Trans Comput Biol Bioinforma*, 11:375–388.

Skjold, S. A. and Wannamaker, L. W. (1976). Method for phage typing group a type 49 streptococci. *Journal of clinical microbiology*, 4(3):232–238.

Skjold, S. A., Wannamaker, L. W., Johnson, D. R., and Margolis, H. S. (1983). Type 49 streptococcus pyogenes: phage subtypes as epidemiological markers in isolates from skin sepsis and acute glomerulonephritis. *Epidemiology & Infection*, 91(1):71–76.

Slater, J. et al. (2017). National equine health survey (nehs) 2017. *National Equine Health Survey (NEHS) 2017*.

Smeds, L. and Künstner, A. (2011). Condetri-a content dependent read trimmer for illumina data. *PloS one*, 6(10).

Smiciklas, M. (2012). *The power of infographics: Using pictures to communicate and connect with your audiences*. Que Publishing.

Smith, W. (1815). *A Delineation of the Strata of England and Wales: With Part of Scotland; Exhibiting the Collieries and Mines, the Marshes and Fen Lands... and the Varieties of Soil...* J. Carey.

Snow, J. (1855). *On the mode of communication of cholera, second edition*. John Churchill.

Somerville, G. A., Reitzer, L., Musser, J. M., et al. (2003). Rgg coordinates virulence factor synthesis and metabolism in streptococcus pyogenes. *Journal of bacteriology*, 185(20):6016–6024.

Southon, S. B., Beres, S. B., Kachroo, P., Saavedra, M. O., Erlendsdóttir, H., Haraldsson, G., Yerramilli, P., Pruitt, L., Zhu, L., Musser, J. M., et al. (2020). Population genomic molecular epidemiological study of macrolide-resistant streptococcus pyogenes in iceland, 1995 to 2016: identification of a large clonal population with a pbp2x mutation conferring reduced in vitro β-lactam susceptibility. *Journal of clinical microbiology*, 58(9).

Spellerberg, B. and Brandt, C. (2016). Laboratory diagnosis of streptococcus pyogenes (group a streptococci). In *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]*. University of Oklahoma Health Sciences Center.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *science*, 333(6048):1393–1400.

Sriskandan, S., Faulkner, L., and Hopkins, P. (2007). Streptococcus pyogenes: Insight into the function of the streptococcal superantigens. *The international journal of biochemistry & cell biology*, 39(1):12–19.

Staali, L., Bauer, S., Mörgelin, M., Björck, L., and Tapper, H. (2006). Streptococcus pyogenes bacteria modulate membrane traffic in human neutrophils and selectively inhibit azurophilic granule fusion with phagosomes. *Cellular microbiology*, 8(4):690–703.

Stafseth, H., Thompson, W., and Neu, L. (1937). Streptococcic infections in dogs. i."acid milk", arthritis, and postvaccination abscesses. *J Am Vet Med Assoc*, 90:769–781.

Stanley, J., Desai, M., Xerry, J., Tanna, A., Efstration, A., and George, R. (1996). High-resolution genotyping elucidates the epidemiology of group a streptococcus outbreaks. *Journal of Infectious Diseases*, 174(3):500–506.

Stärk, K. D., Regula, G., Hernandez, J., Knopf, L., Fuchs, K., Morris, R. S., and Davies, P. (2006). Concepts for risk-based surveillance in the field of veterinary medicine and

veterinary public health: review of current approaches. *BMC health services research*, 6(1):1–8.

Steer, A. C., Law, I., Matatolu, L., Beall, B. W., and Carapetis, J. R. (2009). Global emm type distribution of group a streptococci: systematic review and implications for vaccine development. *The Lancet infectious diseases*, 9(10):611–616.

Stevens, D. (1992). Invasive group a streptococcus infections. *Clinical Infectious Diseases*, 14(1):2–13.

Stevens, D. (2002). Streptococcal toxic shock syndrome. *Clinical Microbiology and Infection*, 8(3):133–136.

Stevens, D. L., Bisno, A. L., Chambers, H. F., Dellinger, E. P., Goldstein, E. J., Gorbach, S. L., Hirschmann, J. V., Kaplan, S. L., Montoya, J. G., and Wade, J. C. (2014). Practice guidelines for the diagnosis and management of skin and soft tissue infections: 2014 update by the infectious diseases society of america. *Clinical infectious diseases*, 59(2):e10–e52.

Stockmann, C., Ampofo, K., Hersh, A. L., Blaschke, A. J., Kendall, B. A., Korgenski, K., Daly, J., Hill, H. R., Byington, C. L., and Pavia, A. T. (2012). Evolving epidemiologic characteristics of invasive group a streptococcal disease in utah, 2002–2010. *Clinical Infectious Diseases*, 55(4):479–487.

Stoppelli, M. P. (2013). The plasminogen activation system in cell invasion. In *Madame Curie Bioscience Database [Internet]*. Landes Bioscience.

Streit, M. and Gehlenborg, N. (2014). Bar charts and box plots.

Sumby, P., Whitney, A. R., Graviss, E. A., DeLeo, F. R., and Musser, J. M. (2006). Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity. *PLoS pathogens*, 2(1):e5.

Sun, H., Ringdahl, U., Homeister, J. W., Fay, W. P., Engleberg, N. C., Yang, A. Y., Rozek, L. S., Wang, X., Sjöbring, U., and Ginsburg, D. (2004). Plasminogen is a critical host pathogenicity factor for group a streptococcal infection. *Science*, 305(5688):1283–1286.

Svensson, M. D., Scaramuzzino, D. A., Sjöbring, U., Olsén, A., Frank, C., and Bessen, D. E. (2000). Role for a secreted cysteine proteinase in the establishment of host tissue tropism by group a streptococci. *Molecular microbiology*, 38(2):242–253.

Sylva, G. L., Sturdevant, D. E., Smoot, L. M., Graham, M. R., Watson, R. O., Musser, J. M., et al. (2002). Rgg influences the expression of multiple regulatory loci to coregulate virulence factor expression in streptococcus pyogenes. *Infection and immunity*, 70(2):762–770.

Tagg, J. R. and Bannister, L. V. (1979). "fingerprinting" β-haemolytic streptococci by their production of and sensitivity to bacteriocine-like inhibitors. *Journal of Medical Microbiology*, 12(4):397–411.

Tagg, J. R. and Martin, D. R. (1984). Evaluation of a typing scheme for group a streptococci based upon bacteriocin-like inhibitor production. *Zentralblatt für Bakteriologie, Mikrobiologie und Hygiene. 1. Abt. Originale. A, Medizinische Mikrobiologie, Infektionskrankheiten und Parasitologie*, 257(1):60–67.

Takeda, N., Kikuchi, K., Asano, R., Harada, T., Totsuka, K., Sumiyoshi, T., Uchiyama, T., and Hosoda, S. (2001). Recurrent septicemia caused by streptococcus canis after a dog bite. *Scandinavian journal of infectious diseases*, 33(12):927–928.

Talbot, J., Setlur, V., and Anand, A. (2014). Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics*, 20(12):2152–2160.

Tamayo, E., Montes, M., García-Arenzana, J. M., and Pérez-Trallero, E. (2014). Streptococcus pyogenes emm-types in northern spain; population dynamics over a 7-year period. *Journal of Infection*, 68(1):50–57.

Tan, R. E. S., Yee, W. X., Cao, D. Y. H., Tan, P. L., and Koh, T. H. (2016). Zoonotic streptococcus canis infection in singapore. *Singapore medical journal*, 57(4):218.

Taniai, H., Iida, K.-i., Seki, M., Saito, M., Shiota, S., Nakayama, H., and Yoshida, S.-i. (2008). Concerted action of lactate oxidase and pyruvate oxidase in aerobic growth of streptococcus pneumoniae: role of lactate as an energy source. *Journal of bacteriology*, 190(10):3572–3579.

Taniyama, D., Abe, Y., Sakai, T., Kikuchi, T., and Takahashi, T. (2017). Human case of bacteremia caused by streptococcus canis sequence type 9 harboring the scm gene. *IDCases*, 7:48–52.

Tarabichi, M., Alvand, A., Shohat, N., Goswami, K., and Parvizi, J. (2018). Diagnosis of streptococcus canis periprosthetic joint infection: the utility of next-generation sequencing. *Arthroplasty today*, 4(1):20–23.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. (2016). Ncbi prokaryotic genome annotation pipeline. *Nucleic acids research*, 44(14):6614–6624.

Terao, Y. (2012). The virulence factors and pathogenic mechanisms of streptococcus pyogenes. *Journal of Oral Biosciences*, 54(2):96–100.

Terao, Y., Kawabata, S., Kunitomo, E., Murakami, J., Nakagawa, I., and Hamada, S. (2001). Fba, a novel fibronectin-binding protein from streptococcus pyogenes, promotes bacterial entry into epithelial cells, and the fba gene is positively transcribed under the mga regulator. *Molecular microbiology*, 42(1):75–86.

Thaker, M., Spanogiannopoulos, P., and Wright, G. D. (2010). The tetracycline resistome. *Cellular and Molecular Life Sciences*, 67(3):419–431.

Tikofsky, L. and Zadoks, R. (2005). Cross-infection between cats and cows: origin and control of streptococcus canis mastitis in a dairy herd. *Journal of dairy science*, 88(8):2707–2713.

Timoney, J., Velineni, S., Ulrich, B., and Blanchard, P. (2017). Biotypes and scm types of isolates of streptococcus canis from diseased and healthy cats. *Veterinary Record*, 180(14):358–358.

Tiseo, K., Huber, L., Gilbert, M., Robinson, T. P., and Van Boeckel, T. P. (2020). Global trends in antimicrobial use in food animals from 2017 to 2030. *Antibiotics*, 9(12):918.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., et al. (2020). Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, 21(1):1–21.

Tory, M., Sprague, D., Wu, F., So, W. Y., and Munzner, T. (2007). Spatialization design: Comparing points and landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1269.

Trell, K., Jörgensen, J., Rasmussen, M., and Senneby, E. (2020). Management of an outbreak of postpartum streptococcus pyogenes emm75 infections. *Journal of Hospital Infection*, 105(4):752–756.

Tsai, W.-C., Shen, C.-F., Lin, Y.-L., Shen, F.-C., Tsai, P.-J., Wang, S.-Y., Lin, Y.-S., Wu, J.-J., Chi, C.-Y., and Liu, C.-C. (2021). Emergence of macrolide-resistant streptococcus pyogenes emm12 in southern taiwan from 2000 to 2019. *Journal of Microbiology, Immunology and Infection*, 54(6):1086–1093.

Tsang, A. K., Lee, H. H., Yiu, S.-M., Lau, S. K., and Woo, P. C. (2017). Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Scientific reports*, 7(1):1–12.

Tufte, E. R. (1985). The visual display of quantitative information. *The Journal for Healthcare Quality (JHQ)*, 7(3):15.

Tufte, E. R. and Robins, D. (1997). *Visual explanations*. Graphics Cheshire, CT.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67.

Turner, C. E., Abbott, J., Lamagni, T., Holden, M. T., David, S., Jones, M. D., Game, L., Efstratiou, A., and Sriskandan, S. (2015). Emergence of a new highly successful acapsular group a streptococcus clade of genotype emm89 in the united kingdom. *MBio*, 6(4).

Turner, C. E., Bedford, L., Brown, N. M., Judge, K., Török, M. E., Parkhill, J., and Peacock, S. J. (2017). Community outbreaks of group a streptococcus revealed by genome sequencing. *Scientific reports*, 7(1):1–9.

Turner, C. E., Holden, M. T., Blane, B., Horner, C., Peacock, S. J., and Sriskandan, S. (2019). The emergence of successful streptococcus pyogenes lineages through convergent pathways of capsule loss and recombination directing high toxin expression. *MBio*, 10(6):e02521–19.

Tyrrell, G. J., Lovgren, M., Forwick, B., Hoe, N. P., Musser, J. M., and Talbot, J. A. (2002). M types of group a streptococcal isolates submitted to the national centre for streptococcus (canada) from 1993 to 1999. *Journal of clinical microbiology*, 40(12):4466–4471.

Tyrrell, G. J., Lovgren, M., St. Jean, T., Hoang, L., Patrick, D. M., Horsman, G., Caeseele, P. V., Sieswerda, L. E., McGeer, A., Laurence, R. A., et al. (2010). Epidemic of group a streptococcus m/emm 59 causing invasive disease in canada. *Clinical infectious diseases*, 51(11):1290–1297.

Ubukata, K., Wajima, T., Morozumi, M., Sakuma, M., Tajima, T., Matsubara, K., Itahashi, K., and Iwata, S. (2020). Changes in epidemiologic characteristics and antimicrobial resistance of streptococcus pyogenes isolated over 10 years from japanese children with pharyngotonsillitis. *Journal of Medical Microbiology*, 69(3):443–450.

Uchiyama, S., Andreoni, F., Schuepbach, R. A., Nizet, V., and Zinkernagel, A. S. (2012). Dnase sda1 allows invasive m1t1 group a streptococcus to prevent tlr9-dependent recognition. *PLoS pathogens*, 8(6):e1002736.

Unnikrishnan, M., Altmann, D. M., Proft, T., Wahid, F., Cohen, J., Fraser, J. D., and Sriskandan, S. (2002). The bacterial superantigen streptococcal mitogenic exotoxin z is the major immunoactive agent of streptococcus pyogenes. *The Journal of Immunology*, 169(5):2561–2569.

Unnikrishnan, M., Cohen, J., and Sriskandan, S. (1999). Growth-phase-dependent expression of virulence factors in an m1t1 clinical isolate of streptococcus pyogenes. *Infection and immunity*, 67(10):5495–5499.

Valdiserri, R. O. and Sullivan, P. S. (2018). Data visualization promotes sound public health practice: the aidsvu example. *AIDS Education and Prevention*, 30(1):26–34.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., and Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC public health*, 14(1):1–9.

Van Sorge, N. M., Cole, J. N., Kuipers, K., Henningham, A., Aziz, R. K., Kasirer-Friede, A., Lin, L., Berends, E. T., Davies, M. R., Dougan, G., et al. (2014). The classical lancefield

antigen of group a streptococcus is a virulence determinant with implications for vaccine design. *Cell host & microbe*, 15(6):729–740.

Vannice, K. S., Ricaldi, J., Nanduri, S., Fang, F. C., Lynch, J. B., Bryson-Cahn, C., Wright, T., Duchin, J., Kay, M., Chochua, S., et al. (2019). Streptococcus pyogenes pbp2x mutation confers reduced susceptibility to β-lactam antibiotics. *Clinical Infectious Diseases*.

Villalón, P., Sáez-Nieto, J. A., Rubio-López, V., Medina-Pascual, M. J., Garrido, N., Carrasco, G., Pino-Rosa, S., and Valdezate, S. (2021). Invasive streptococcus pyogenes disease in spain: a microbiological and epidemiological study covering the period 2007–2019. *European Journal of Clinical Microbiology & Infectious Diseases*, 40(11):2295–2303.

Vos, P., Garrity, G., Jones, D., Krieg, N. R., Ludwig, W., Rainey, F. A., Schleifer, K.-H., and Whitman, W. B. (2011). *Bergey's manual of systematic bacteriology: Volume 3: The Firmicutes*, volume 3. Springer Science & Business Media.

Voyich, J. M., Sturdevant, D. E., Braughton, K. R., Kobayashi, S. D., Lei, B., Virtaneva, K., Dorward, D. W., Musser, J. M., and DeLeo, F. R. (2003). Genome-wide protective response used by group a streptococcus to evade destruction by human polymorphonuclear leukocytes. *Proceedings of the National Academy of Sciences*, 100(4):1996–2001.

Walker, M. J., Barnett, T. C., McArthur, J. D., Cole, J. N., Gillen, C. M., Henningham, A., Sriprakash, K., Sanderson-Smith, M. L., and Nizet, V. (2014). Disease manifestations and pathogenic mechanisms of group a streptococcus. *Clinical microbiology reviews*, 27(2):264–301.

Wang, H., Lottenberg, R., and Boyle, M. D. (1995). Analysis of the interaction of group a streptococci with fibrinogen, streptokinase and plasminogen. *Microbial pathogenesis*, 18(3):153–166.

Wang, X., Zhang, X., and Zong, Z. (2016). Genome sequence and virulence factors of a group g streptococcus dysgalactiae subsp. equisimilis strain with a new element carrying erm (b). *Scientific reports*, 6:20389.

Wasserzug, O., Valinsky, L., Klement, E., Bar-Zeev, Y., Davidovitch, N., Orr, N., Korenman, Z., Kayouf, R., Sela, T., Ambar, R., et al. (2009). A cluster of ecthyma outbreaks caused by a single clone of invasive and highly infective streptococcus pyogenes. *Clinical Infectious Diseases*, 48(9):1213–1219.

Webb, R. H., Grant, C., and Harnden, A. (2015). Acute rheumatic fever. *Bmj*, 351:h3443.

Weber, S., Pfaller, M. A., and Herwaldt, L. A. (1997). Role of molecular epidemiology in infection control. *Infectious disease clinics of North America*, 11(2):257–278.

Weeks, C. R. and Ferretti, J. J. (1984). The gene for type a streptococcal exotoxin (erythrogenic toxin) is located in bacteriophage t12. *Infection and immunity*, 46(2):531–536.

Welsh, J. and McClelland, M. (1990). Fingerprinting genomes using pcr with arbitrary primers. *Nucleic acids research*, 18(24):7213–7218.

Wessels, M. R. (2011). Streptococcal pharyngitis. *New England Journal of Medicine*, 364(7):648–655.

Wessels, M. R. (2016). Pharyngitis and scarlet fever. In *Streptococcus pyogenes: Basic Biology to Clinical Manifestations [Internet]*. University of Oklahoma Health Sciences Center.

Whatmore, A. M., Engler, K. H., Gudmundsdottir, G., and Efstratiou, A. (2001). Identification of isolates of streptococcus canisinfecting humans. *Journal of clinical microbiology*, 39(11):4196–4199.

WHO (2005). The current evidence for the burden of group a streptococcal diseases. Technical report, World Health Organization.

WHO (2020). Glass whole-genome sequencing for surveillance of antimicrobial resistance.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6):e1005595.

Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome biology*, 20(1):129.

Wilkening, R. V. and Federle, M. J. (2017). Evolutionary constraints shaping streptococcus pyogenes–host interactions. *Trends in microbiology*, 25(7):562–572.

Wilkinson, L. (2012). The grammar of graphics. In *Handbook of computational statistics*, pages 375–414. Springer.

Williamson, D. A., Morgan, J., Hope, V., Fraser, J. D., Moreland, N. J., Proft, T., Mackereth, G., Lennon, D., Baker, M. G., and Carter, P. E. (2015). Increasing incidence of invasive group a streptococcus disease in new zealand, 2002–2012: a national population-based study. *Journal of Infection*, 70(2):127–134.

Wilson, D. J., Gonzalez, R. N., and Das, H. H. (1997). Bovine mastitis pathogens in new york and pennsylvania: prevalence and effects on somatic cell count and milk production. *Journal of dairy science*, 80(10):2592–2598.

Xiong, C., Setlur, V., Bach, B., Koh, E., Lin, K., and Franconeri, S. (2021). Visual arrangements of bar charts influence comparisons in viewer takeaways. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):955–965.

Xu, K., Rooney, C., Passmore, P., Ham, D.-H., and Nguyen, P. H. (2012). A user study on curved edges in graph visualization. *IEEE transactions on visualization and computer graphics*, 18(12):2449–2456.

Yang, W., Li, Z., Lan, Y., Wang, J., Ma, J., Jin, L., Sun, Q., Lv, W., Lai, S., Liao, Y., et al. (2011). A nationwide web-based automated system for outbreak early detection and rapid response in china. *Western Pacific surveillance and response journal: WPSAR*, 2(1):10.

Yoshida, H., Goto, M., Fukushima, Y., Maeda, T., Tsuyuki, Y., and Takahashi, T. (2021). Intracellular invasion ability and the associated microbiological characteristics of streptococcus canis in isolates from japan. *Japanese Journal of Infectious Diseases*, 74(2):129–136.

You, Y., Davies, M. R., Protani, M., McIntyre, L., Walker, M. J., and Zhang, J. (2018). Scarlet fever epidemic in china caused by streptococcus pyogenes serotype m12: epidemiologic and molecular analysis. *EBioMedicine*, 28:128–135.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Yu, L., Petros, A. M., Schnuchel, A., Zhong, P., Severin, J. M., Walter, K., Holzman, T. F., and Fesik, S. W. (1997). Solution structure of an rrna methyltransferase (ermam) that confers macrolide-lincosamide-streptogramin antibiotic resistance. *Nature structural biology*, 4(6):483–489.

Yuzenkova, Y., Gamba, P., Herber, M., Attaiech, L., Shafeeq, S., Kuipers, O. P., Klumpp, S., Zenkin, N., and Veening, J.-W. (2014). Control of transcription elongation by grea determines rate of gene expression in streptococcus pneumoniae. *Nucleic acids research*, 42(17):10987–10999.

Zachariadou, L., Stathi, A., Tassios, P., Pangalis, A., Legakis, N., Papaparaskevas, J., Group, H. S.-E. S., et al. (2014). Differences in the epidemiology between paediatric and adult invasive streptococcus pyogenes infections. *Epidemiology & Infection*, 142(3):512–519.

Zaidi, S. M. and Eranki, A. (2019). Streptococcus canis bacteremia in a renal transplant recipient. *Journal of investigative medicine high impact case reports*, 7:2324709619834592.

Zinsstag, J., Schelling, E., Bonfoh, B., Fooks, A. R., Kasymbekov, J., Waltner-Toews, D., Tanner, M., et al. (2009). Towards a 'one health'research and application tool box. *Veterinaria italiana*, 45(1):121–133.

Zinsstag, J., Schelling, E., Waltner-Toews, D., and Tanner, M. (2011). From "one medicine" to "one health" and systemic approaches to health and well-being. *Preventive veterinary medicine*, 101(3-4):148–156.