Genomic Characterisation of *Streptococcus dysgalactiae* subsp. *equisimilis* Associated with Respiratory Tract Infections

Filipe Barbosa Valcovo
filipe.valcovo@tecnico.ulisboa.pt

IST, University of Lisbon, Portugal
November 2022

Abstract

Human infections caused by *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) have been rising in recent decades. Typing of the *emm* gene, which encodes an important virulence factor in SDSE, revealed two *emm* types to be almost exclusively associated with respiratory tract (RT) infections in Portugal. This motivated the characterisation of 199 RT isolates recovered in Portugal in the years 2011 to 2019 to define the main genetic lineages responsible for RT infection, to determine antimicrobial resistance, and employ high throughput sequencing to determine *emm* types, multilocus sequence typing (MLST) and core genome MLST (cgMLST) allele profiles, known virulence factors and to perform a genome-wide association study (GWAS) to explore possibly unknown virulence factors. RT infections in Portugal were more frequent among younger patients. Typing of the *emm* gene found 31 distinct *emm* types, of which *stC36* and *stC839* were almost exclusively associated with RT infections. MLST revealed 121 sequence types (ST) distributed among 37 clonal complexes (CC) and 72 singletons, with 3 distantly related lineages, CC3, CC49 and CC68, being almost exclusively associated with RT infection. While some CCs had a diverse variety of *emm* types, the CCs associated with the RT were associated with the RT *emm* types. Among known streptococcal virulence factors, *mf3*, a phage-associated extracellular nuclease, was almost exclusively associated to RT CCs. GWAS revealed that RT CCs lacked several CRISPR-associated proteins, which serve important roles in protection against phages, and possessed several phage genes, which may affect virulence.

Keywords: *Streptococcus dysgalactiae* subsp. *equisimilis*, invasive infection, respiratory tract infection, high throughput sequencing, gene association studies, typing.

1. Introduction

Human infections caused by SDSE have been rising in the past four decades. This bacterium is genetically similar to *S. pyogenes* (SP), in the spectrum of diseases it causes, and in the virulence factors it possesses. Although widely regarded as non-pathogenic in the past, increasing awareness to this pathogen in recent years is leading to a better understanding of its epidemiological dynamics, taxonomy, pathogenicity and tissue tropism (Baracco, 2019; Brandt & Spellerberg, 2009).

1.1. Typing

The gene *emm* encodes the highly variable surface M protein, an important virulence factor in both SP and SDSE. This gene has a hypervariable region which is responsible for M serospecificity (Gherardi et al., 2013). Extensively used to characterize SP isolates, *emm* typing has become the most commonly used typing method to characterise SDSE isolates. Predominant *emm* types vary according to geographic area and specific *emm* types may be associated with different infections (Pinho et al., 2006; Sunaoshi et al., 2010). Furthermore, differences in *emm* types prevalence in a given geographic area have allowed the detection of emergence of specific genetic lineages (Oppegaard et al., 2017).

MLST is a universal standardised method for characterising bacterial isolates which uses the nucleotide sequences of housekeeping genes. MLST data can be employed to better understand the epidemiology, population dynamics, pathogenicity and evolution of bacteria. Unique sequences for these genes are assigned allele numbers, and allele combinations are designated STs. The MLST scheme currently used for SDSE uses the following housekeeping genes: *gki, gtr, murI, mutS, recP, xpt*, and *atoB* (Maiden, 2006; McMillan et al., 2010; Sang-Ik Oh et al., 2020). In a population, the majority of isolates form a cluster of closely related genotypes, referred to as a clonal complex (CC), which can be defined as the group containing the predominant genotype and its single *locus* variants (SLV), double *locus* variants (DLV), and so forth (Feil et al., 2004). The eBURST and goeBURST.java implementations of the BURST algorithm divide MLST sets into nonoverlapping groups of CCs and then subgroup the descendants of each predicted founder. This method can provide valuable insight into how bacterial clones diversify and the emergence of clinically important clones (Feil et al., 2004; Francisco et al., 2009; Nascimento et al., 2017). Studies that characterised SDSE recovered from invasive and non-invasive infections in different geographical areas have found that most *emm* types can be found in different STs and that the same ST may have a variety of different *emm* types. This suggests that there is recombination involving the *emm* gene and thus it may not be a good indicator of genetic relatedness of SDSE strains (Matsue et al., 2020; McMillan et al., 2011; Wajima et al., 2016).

1.2. Whole Genome Sequencing

With WGS becoming a more common practise due to the increasing availability of next generation sequencing, MLST, cgMLST and *emm* typing can be done with WGS data. Information on virulence, drug resistance, phages, serotype, phylogeny, and epidemiology can also be extracted from WGS data (Shimomura et al., 2011; Uelze et al., 2020; Watanabe et al., 2013).

1.2.1. Core Genome MLST

In core genome MLST (cgMLST), rather than using 7 genes, gene-by-gene allelic profiling is done to the core genome (the set of genes present in at least 95% of the strains in a group of strains) of a set of isolates in the same species, allowing high-resolution typing of closely related strains that would otherwise be indistinguishable because they may possess the same MLST profile (Maiden, 2006; Ruppitsch et al., 2015). Furthermore, cgMLST is especially useful to define relationships between isolates in the same serogroup as some isolates can be more closely related to isolates belonging to other serogroups than to those in their own due to events of capsular transformation (Medini et al., 2005).

1.2.2. The Pan-Genome

The concept of the pan-genome was introduced by in 2005 and it was defined as the complete set of genes present in a group of strains of interest, which is comprised of the core genome, the set of genes present in at least 95% of all strains in a group, and the dispensable or accessory genome, present in fewer than 95% of the strains. (Medini et al., 2005). Applying this concept to a specific clade results in the exclusive accessory genome for that clade or group of strains, which may contain genes that confer selective advantages to that group, such as adaptation to different niches, antibiotic resistance or colonisation.

1.2.3. Genome Wide Association Studies

WGS produces large amounts of genomic data, much of which is poorly understood. Bacterial genome-wide association studies (GWAS) have allowed us to find associations between genetic variants and observed phenotypes. GWAS are typically used to link certain genes to clinically relevant traits, such as virulence or antibiotic resistance (Farhat et al., 2013; Holt et al., 2015). Typically, in eukaryotes, these studies link single nucleotide polymorphisms (SNP) to a specific phenotype or trait. However, due to the nature of bacterial genomes and fundamental differences in bacterial evolution and genomes, an SNP approach may be inadequate and gene-by-gene presence/absence methods are often used (Brynildsrud et al., 2016).

1.3. Colonisation

SDSE colonises the pharynx, gastrointestinal and female genitourinary tracts of humans and is often isolated from wounds (Park et al., 2016; Traverso et al., 2016). Studies on colonisation by SDSE have found different rates of asymptomatic pharyngeal carriage depending on geographic area and usually focus on adolescents and young adults: the age groups in which tonsillopharyngitis is more prevalent. Agerhäll and co-workers found an asymptomatic pharyngeal carriage rate of 7.8% in Sweden in adolescents and young adults (Agerhäll et al., 2021). A study on Indian school children found a colonisation rate of almost 10% amongst school children with an active infection and asymptomatic carriers (Bramhachari et al., 2010).

1.4. Infection

Although it is a part of the normal human microbiota, reports on the pathogenicity of SDSE are increasing. SDSE has been considered less virulent than SP, however, recent reports show high pathogenicity in some strains (Matsue et al., 2020; M. D. Pinho, 2014). The acquisition of virulence factors from SP by horizontal gene transfer may also increase virulence in some strains (McMillan et al., 2011). Different *emm* types have typically been associated with varying severity of infection (Rantala et al., 2010). The fact that *emm* types do not cycle through communities suggests that there is no acquisition of type-specific immunity (McDonald et al., 2007).

1.4.1. Tonsillopharyngitis

In cases in which SDSE has been isolated, patients show signs and symptoms indistinguishable from those presented by patients with SP tonsillopharyngitis (Brandt & Spellerberg, 2009). In some geographical areas, the disease burden of SDSE associated with pharyngitis is estimated to be greater than that of SP. In Australian native communities and in India, for example, the isolation rates of SP are very low compared to those of SDSE in cases of pharyngitis. The *emm* types in SDSE isolated from the pharynx are very diverse and prevalent types vary according to geographical area (Bramhachari et al., 2010; McDonald et al., 2007).

1.4.2. Skin and Soft Tissue Infections

Skin and soft tissue (SST) infections caused by SDSE may manifest as pyoderma, erysipelas, cellulitis, abscesses, pyomyositis or necrotising soft tissue infections (Brandt & Spellerberg, 2009a). These infections may also serve as the portal of entry for bacteraemia and other invasive infections. Ulcers caused by conditions such as diabetes mellitus or any kind of lymphatic or venous compromise may be complicated by these infections (Baracco, 2019; Ciszewski & Szewczyk, 2017).

1.4.3. Invasive Infections

In some geographic locations, isolation of SDSE from blood is almost as frequent as the isolation of SP (Park et al., 2016; Traverso et al., 2016). Bacteraemia, an invasive infection commonly caused by SDSE, is commonly secondary to skin and soft tissue infections. The aforementioned underlying diseases also constitute risk factors for invasive infections and are present in about 70% of cases. Community acquired bacteraemia accounts for 70% of cases and the portal of entry is the skin in most of them. A study found recurrence of bacteraemia caused by SDSE in 6 out of 84 patients, two of which with the same *emm* type recovered as in the previous infection (*stG840*), which may suggest that infection by these organisms does not confer protective immunity (Cohen-Poradosu et al., 2004).

1.5. Pathogenesis and Virulence

Approximately 72% of the SDSE genome is homologous with SP and many common virulence factors are expressed, such as the surface M-protein, streptolysins O and S, streptokinase, and C5a peptidase (Watanabe et al., 2013). Additionally, horizontal gene transfer involving virulence factors has been described between these species (McNeilly & McMillan, 2014). SDSE uses fibronectin as one of the main targets for attachment, but other host cell adherence strategies include binding to additional extracellular matrix molecules, such as fibrinogen, vitronectin, collagen and plasminogen. The M protein and the M-like fibrinogen binding protein of group G streptococci (FOG) serve important roles in

adhesion as they bind to fibrinogen. The multidomain surface of these proteins with a coiled-coil secondary structure form irregularities that are essential for fibrinogen-binding properties (Brandt & Spellerberg, 2009; Johansson et al., 2004). The ability to resist phagocytosis is a common characteristic amongst the pathogenic streptococci. The major antiphagocytic factor in streptococci is the M protein, whose fibrinogen-binding properties lead to evasion of the host's nonspecific immune response (Fischetti, 1989; Mcmillan et al., 2013; C. E. Turner et al., 2019). As well as the M protein, FOG can inhibit phagocytosis by binding to fibrinogen (Brandt & Spellerberg, 200; Johansson et al., 2004). Nucleases such as Spd1 also play a role in immune evasion, as they degrade the chromatin in neutrophil extracellular traps. Extracellular nuclease activity is typically higher in clinically relevant strains of SP, which suggests it may contribute to virulence, but the specific role of nucleases has not yet been characterised and described in detail (Korczynska et al., 2012; Sumby et al., 2005).

1.6. Treatment and Antimicrobial Resistance

Penicillin and other beta-lactams remain the drugs of choice to treat SDSE infections, since, like other beta-haemolytic streptococci, SDSE are susceptible to these antibiotics (Baracco, 2019). A single study has reported the occurrence of a SDSE penicillin resistant clone. It had mutations in multiple penicillin-binding proteins (PBPs), including some similar to those in *S. pneumoniae* and *S. agalactiae* (Fuursted et al., 2016). No further cases of penicillin resistance have been documented. High resistance rates to tetracyclines, clindamycin, macrolides and fluoroquinolones demand susceptibility testing before antimicrobial therapy. (Brandt & Spellerberg, 2009). Lincosamides and streptogramin B are structurally different from macrolides but they share the same mechanism of action and consequently, resistance to these classes of antimicrobials is related (Leclercq, 2002).

1.7. Aims of the Current Work

The discovery of significant associations between *emm* types *stC36* and *stC839* and RT infections motivated further research on this topic. This association was observed in all SDSE isolates from the years 2011-2018 (unpublished results), suggesting that strains exhibiting these *emm* types exhibit tropism to the respiratory tract or are unable to cause invasive infection. This work aims to genomically characterise SDSE isolates recovered from respiratory tract infections in Portugal; to identify the main genetic lineages present in RT infections; to identify differences in known virulence factors among the detected clonal lineages, particularly virulence factors which could contribute to differences in virulence or tissue tropism; to compare RT infection isolates with isolates recovered from invasive infection in the same period; to explore and identify possible unknown virulence factors responsible for differences in virulence among the detected clonal lineages; and to determine the antimicrobial susceptibility of SDSE involved in RT infections.

2. Materials and Methods

2.1. Bacterial Strains

SDSE isolates (n=626) collected from clinical respiratory tract (RT) (n=199) or invasive (n=427) specimens in Portugal during the years 2011 to 2019 were selected for study. These isolates were recovered in Portuguese hospitals and identified in hospital laboratories. The non-invasive respiratory tract strains were recovered from pharyngeal exudate (n=140), sputum (n=55), nasal exudate (n=2) and nasopharyngeal exudate (n=2) samples. Isolates from invasive infections were collected from blood (n=390), synovial fluid (n=26), ascitic fluid (n=7), pleural fluid (n=2), bone biopsy (n=1), and cerebrospinal fluid (n=1). SDSE invasive isolates collected up to 2017 were previously characterised (Castro, 2020) and data was included for comparison purposes, while invasive isolates recovered in 2018 and 2019 (n=110) were newly characterised in this study. Five additional strains of *emm* types *stC36* (n=2) and *stC839* (n=3) isolated from skin and soft tissue infections (n=4) or urine (one *stC36* isolate) collected in the same time range were included in the pangenome and GWAS analyses due to the association between these *emm* types and RT infections.

2.2. Culture conditions

All isolates were stored at -80 ºC in tryptone soy broth (TSB) (Oxoid, Basingstoke, UK) supplemented with 15% (v/v) glycerol (Sigma-Aldrich, St. Louis, Missouri, US) until processing. Strains were cultured in tryptone soy agar (TSA) (Oxoid, Basingstoke, UK) supplemented with 5% (v/v) defibrinated sheep blood (Probiológica, Lisboa, Portugal) at 35 ºC in ambient atmosphere. Prior to genome extraction, strains were cultured in Todd Hewitt broth (BD, Sparks, MD, USA) at 35 ºC with no shaking for 16 hours. All growth media were prepared according to the manufacturer's recommendations.

2.3. Strain Identification

Identification of isolates to the species level was done by the submitting laboratory. Haemolysis and colony size (visual confirmation of size >0.5 mm) were confirmed by culture in TSA supplemented with sheep blood after incubation in ambient atmosphere at 35 ºC for 24 hours. Lancefield groups were determined with a commercially available latex agglutination test (Oxoid, Basingstoke, UK).

2.4. Antimicrobial Susceptibility Testing

Antimicrobial susceptibility testing was performed by disc diffusion using the Kirby-Bauer method according to CLSI guidelines (CLSI, 2020). Incubation was done at 35 ºC in an atmosphere enriched with 5% CO_2 on Mueller-Hinton agar (Oxoid, Basingstoke, UK) supplemented with 5% defibrinated sheep blood. The disks (Oxoid, Basingstone, UK) used were penicillin (10u), chloramphenicol (30 µg), vancomycin (30 µg) tetracycline (30 µg), levofloxacin (5 µg), linezolid (30 µg), gentamycin (120 µg), streptomycin (300 µg), clindamycin (2 µg), and erythromycin (15 µg). *Streptococcus pneumoniae* ATCC 49619 and *Enterococcus faecalis* ATCC 29212 (for aminoglycosides) were used for quality control testing according to the same guidelines.

2.5. Genome Extraction, Quality Control and Sequencing

Genome extraction for high throughput sequencing (HTS) was done using the Invitrogen PureLink® Genomic DNA extraction kit (Thermo Fisher Scientific Inc., Waltham, MA, USA). Manufacturer instructions were used with slight modifications. Mutanolysin)75 U) (Sigma-Aldrich, St. Louis, Missouri, EUA) were added to the lysis buffer with lysozyme. After the addition of proteinase K, 400 µg of RNase (included in the kit) were added and incubated for 1 minute at room temperature. DNA purity was assessed with a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). DNA integrity was assessed by gel electrophoresis. For this purpose, a 1% (w/v) agarose gel was prepared with electrophoresis grade agarose (Bio-Rad, Hercules, California, USA) in 0.5X TBE buffer (Bio-Rad, Hercules, California, USA). The ladder 1KB plus (Invitrogen, Carlsbad, California) was used as a molecular weight marker. The nucleic acid concentration was measured by fluorometry with an Invitrogen Qubit™ dsDNA HS assay (Thermo Fisher Scientific Inc., Waltham, MA, USA) according to the manufacturer's instructions and adjusted to 10 ng/µl by dilution in Tris-HCL 10 mM pH8 (Sigma-Aldrich, St. Louis, Missouri, EUA). Genome sequencing was done at Instituto Gulbenkian de Ciência, Gene Express Unit (Oeiras, Portugal) on the Illumina NextSeq 2000 (Illumina) system using the NextSeq 500/550 Mid-Output kit (version 2) (300 cycles). Genome assembly was done by the bioinformatics team in Instituto de Microbiologia da Faculdade de Medicina de Lisboa.

2.6. Genome Annotation

The 631 assembled genomes were annotated using Prokka (Seemann, 2014) version 1.14.5, an open source software tool developed to achieve reliable annotation of bacterial genomic sequences. The output consists of several nucleotide and protein FASTA files, the sequences and annotations in Genbank and GFF v3 format, a log file and an annotation summary statistics file. Prokka was run with the following parameters: *--kingdom Bacteria --genus Streptococcus*. The presence of known virulence factors in each strain was determined with VFDB (Virulence Factor Database) (http://www.mgc.ac.cn/VFs/main.htm) and genomic sequences and context of several genes were analysed with Geneious 8.1.9 (https://www.geneious.com).

2.7. Pangenome

The tool used to analyse the pangenome of the complete set of strains was Roary (Page et al., 2015). The output incudes a CSV file with gene presence and absence in all strains and a statistics file. Besides the typical core and accessory genome, Roary separates the pangenome into the following sections: core genes, present in 99% or more of the strains, soft core genes, present in 95-99% of the strains, the shell, with genes in 15-95% of the isolates, and cloud genes, present in 15% or fewer of the isolates.

2.8. Genome-Wide Association Study

The chosen tool to perform GWAS was Scoary (Brynildsrud et al., 2016), version 1.6.16. Assuming the same genes could be responsible for the differences in tissue tropism between RT-associated CCs and other CCs, strains in CC3, CC49 or CC68 were considered trait-positive for association with the RT, and all other strains negative. The cut-off values were sensitivity and specificity higher than 70% or lower than 30% and a significant HP corrected p-value. After this selection, sequences of each gene were searched in the UniProt BLAST search tool (https://www.uniprot.org/blast).

2.9. Statistical Analysis

Population diversity was characterised with the Simpson's index of diversity (SID) and corresponding CI95%. This index represents the probability of any two strains randomly selected from a population belonging to two different groups. It can be used as a measurement of the discriminative power of typing methods. It varies from 0 to 1, with population diversity increasing as SID approaches 1 (Hunter & Gaston, 1988). The adjusted Wallace coefficients (AW) and corresponding CI95% were calculated to compare the congruence of two different typing methods. It compares the partitions obtained by two typing methods and it indicates the probability of the results of one method being predicted by another method. AW values range from 0 to 1, with higher congruency in both methods as it approaches 1 (Severiano et al., 2011). Odd ratios (OR) and false discovery rate (FDR) corrected p-values were calculated (based on Fisher's exact test) to evaluate individual associations in large groups.

3. Results

3.1. Patient Demographics

The mean age of patients from which the isolates used in this work were recovered was 56.5 years. The mean ages of RT and invasive infection patients were 27.7 and 69.8 years respectively, with approximately 50% of RT patients being 19 years of age or younger and 50% of invasive infection patients being 72 years or older.

3.2. Bacterial Genetic Lineages Involved in Infection

3.2.1. *emm* Typing

Typing of the emm gene found 31 distinct *emm* types in RT isolates and 30 distinct *emm* types and two non-typeable strains in invasive infections, yielding a total of 38 distinct *emm* types among RT and invasive infections in the years 2011 to 2019 (Figure 1).

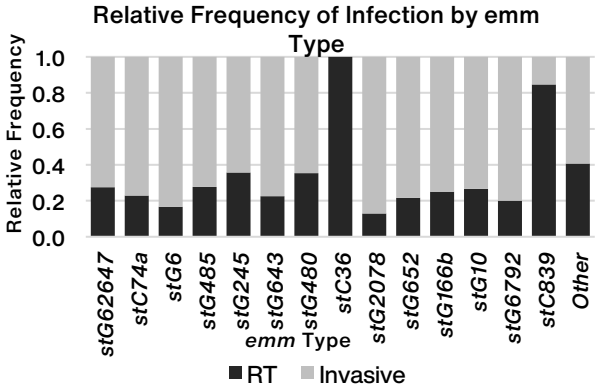

Figure 1. Distribution of RT and invasive infections for each emm type. Only emm types with 10 or more representative strains are shown. While the fraction of RT infections in most emm types is low, never exceeding 36%, 100% of *stC36* isolates and 84.6% of *stC839* isolates are from RT infections.

3.2.2. Multilocus Sequence Typing

Distribution of RT and Invasive Infections

MLST revealed 121 sequence types (ST) overall, which were distributed among 37 clonal complexes and 72 singletons. Approximately 75% of all isolates (n=626) belong to three CCs: CC17 (n=198), CC20 (n=129), CC29 (n=96), and CC15 (n=50). In the RT infection subset (n=199), approximately 55% of isolates belong to the same three CCs: CC17 (n_{RT}=47), CC20 (n_{RT}=34), CC29 (n_{RT}=28). Three CCs were found to be associated with RT infections: CC3 (n=15; p=8.17E-7) is almost exclusively associated and CC49 (n=14, p<0.001) and CC68 (n=13, p<0.001) were exclusively associated with these infections (Figure 2). The proportion of isolates in each CC per year did not vary significantly and CC17 is the most frequently isolated in all years with the exception of 2011. The CCs associated with RT infections represent a small fraction of all RT and invasive strains, varying between 2.3% in 2017 and 12.5% in 2016. While most CCs have a diverse variety of *emm* types, the CCs associated with RT infections had less diversity.

Most strains in CC3 (n=15) were of *emm* type *stC839* (n=11) (others are *emm57* (n=3) and *stG653* (n=1)) and all strains in CC49 and CC68 possessed the sequence type *stC36*. Most CCs are exclusively or almost exclusively associated with only one Lancefield group (Adjusted Wallace CC→Lancefield group ±95% CI, 0.985±0.017). All RT-associated CCs (CC3, CC49, CC68) were significantly associated with group C (p<0.01), however, group C is not associated with RT infections.

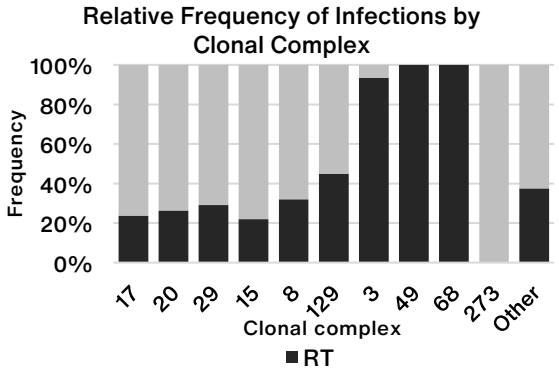

Figure 2. Distribution of RT and invasive infections for each CC. Only CCs with 10 or more representative strains are shown. While the fraction of RT infections in most CCs is low, never exceeding 36%, 100% of CC49 and CC68 isolates and 93.3% of CC3 isolates are from RT infections.

3.3. M Protein Analysis

A neighbour-joining tree of the M protein genes of different *emm* types (Figure 3) revealed that the *emm* types that are associated with RT infections (*stC36* and *stC839*) have similar sequences to types with no strong association to invasive or RT infections. M protein sequences of *emm* types *stC36, stC839, stC74A, stG245, stG485, stG4831* and *stG25* were aligned with MUSCLE and manually analysed for differences between the RT-associated *emm* types (*stC36* and *stC839*) and the others. No differences in hydrophobicity, G+C content, or isoelectric point patterns were found between the RT-associated *emm* sequences and the others. No differences were found in the genomic context of the emm gene (10 000 bp around the gene) in RT infection isolates of different CCs or emm types, in which contig length allowed sequence visualisation in that entire range.

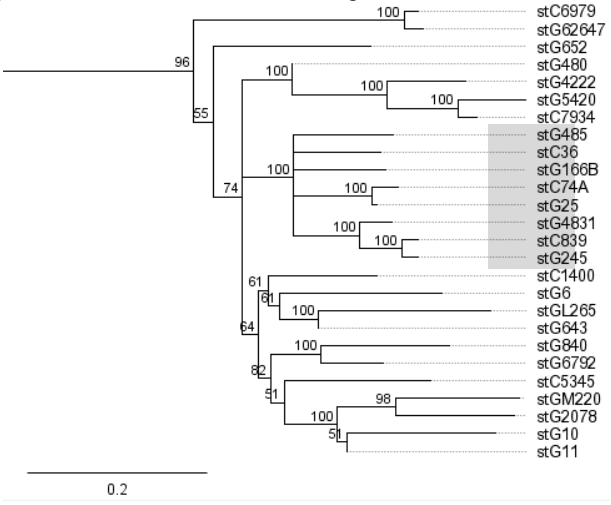

Figure 3. Neighbour-joining tree of M protein sequences of different *emm* types. The genetic distance was inferred using the Jukes-Cantor model and sequence type *stC12* as the outgroup. M protein sequences were translated into amino acid sequences, and a MUSCLE (multiple sequence comparison by log-expectation) alignment was performed in Geneious8.

3.4. Other Virulence Factors

WGS allowed the detection of known streptococcal virulence factors (VF) in all RT and invasive isolates. The genes *fbp54*, *hasC*, and *sagA* were present in all the RT and invasive infection isolates. The only gene with a significant association to RT infection was *mf3* (mitogen factor 3), an extracellular nuclease, being present in 22.6% (n=45) of RT infection isolates and only 3% (n=13) of invasive infection isolates.

When the presence of VF genes was analysed for each CC, *mf3* and *speG* had higher ORs for RT-associated CCs. The gene *mf3* was absent in CC17, CC20 and CC273 and it was found in 45% (n=9, p<0.001) of CC129 strains, in 86.7% (n=13, p<0.001) of CC3 strains, in 78.6% (n=11, p<0.001) of CC49 strains, and in 92.3% (n=12, p<0.001) of CC68 isolates. This indicates a strong association of *mf3* to the RT-associated CCs.

3.5. Pangenome and Gene Association

3.5.1. The Pangenome

Five additional strains of *emm* types *stC36* (n=2) and *stC839* (n=3) isolated from skin and soft tissue infections (n=4) or urine (one *stC36* isolate) collected in the same time range as the RT and invasive infection strains were included in the pangenome and GWAS analyses. The dataset of RT and invasive infection strains and the five extra strains (n=631) generated a pangenome with 10 684 genes. The core genome, with 1409 genes, is comprised of 1255 core genes, present in at least 99% of the isolates and 154 soft-core genes, present in at least 95% to 99% of the isolates. The accessory genome , with 9275 genes, is comprised of 1155 shell genes, present in 15% to 95% of the isolates, and 8120 cloud genes, present in fewer than 15% of the isolates.

3.5.2. Gene Association Study

A GWAS was performed to find genes that may be responsible for the differences in virulence between strains of

different CCs. Strains were classified as trait-positive if they belonged to CC3, CC49 or CC68 and trait-negative if they belonged to other CCs and the strength of association of each gene to trait-positive (RT-associated) strains was calculated. After selection, a total of 17 genes whose presence may be relevant and 24 genes whose absence may be relevant for RT infection were analysed. Analysis of a *group_5532* sequence revealed that is the *mf3* gene. That was the only nuclease whose presence in RT CCs was significant. Three genes are involved in the synthesis of rhamnose or rhamnose polysaccharides in the cell wall. In the set of genes whose absence may be relevant, CRISPR-associated proteins are among the genes with the lowest specificity to RT CCs and several proteins involved in cell wall polysaccharide synthesis are also present.

3.6. Antimicrobial Resistance

All isolates in the present work were susceptible to penicillin, vancomycin, and linezolid. Among the strains recovered from RT infections, the highest resistance rate was 28.1% (n=56), to erythromycin and no resistant strains to streptomycin or chloramphenicol were found. No significant differences in antimicrobial resistance were found between RT and invasive infection isolates. Antimicrobial resistance rates varied considerably in different CCs. However, they did not differ significantly between the CCs associated with RT infection (CC3, CC49, and CC68).

4. Discussion

This work characterised SDSE strains from human respiratory tract infections collected in Portugal in the years 2011 to 2019. To date, no large-scale gene association studies performed on SDSE RT infection isolates have been published. RT infections by SDSE in the study period were more common in younger patients. The mean age of RT infection patients is 27.7 years with a median of 19 years, which coincides with the ages in which colonisation by SDSE is more prevalent in Europe and North America. Pharyngitis is a classical presentation of SDSE infection in younger patients. (Agerhäll et al., 2021; J. C. Turner et al., 1997). In SP RT infections, repeated exposure to this pathogen confers protective immunity with antibodies against the hypervariable N-terminal portion of the M protein and that is hypothesised to reduce the frequency of RT infections with age, but no such mechanism is elucidated in SDSE (Brandt & Spellerberg, 2009; Cannon et al., 2021). In addition, SDSE emm types do not cycle through communities, suggesting there is no acquired immunity to the M protein of SDSE (McDonald et al., 2007).

4.1. Molecular Typing

Type *stC839* and stC36 were associated to the RT, which is in line with the observations made in the years 2011 to 2018 regarding these *emm* types, suggesting that strains exhibiting these *emm* types have strong tropism to the respiratory tract. These differences could be caused by the M protein itself, which is an important virulence factor responsible for adherence and antiphagocytosis (Mcmillan et al., 2013), differences in the virulome of these strains in general, or differences in expression of virulence factors.

Reports of recombination events in the *emm locus* and the fact that the M protein is a surface protein and thus subject to selective and diversifying pressure mean that *emm* typing does not reflect evolutive relations between strains (McMillan et al., 2011; McNeilly & McMillan, 2014; van Belkum et al., 2007). MLST and cgMLST were used to define the main clonal lineages present in RT infections. MLST uses housekeeping genes under low selective pressure, and it offers insight into the genetic relations between strains. Higher discriminative power is achieved by cgMLST, which uses all the genes in a defined core genome. Overall, the 4 most frequent CCs, CC17, CC20, CC29, and CC15, included approximately 75% of all RT and invasive isolates whilst in the RT infection subset, the same CCs included only 55% of these infections. Furthermore, strains of CC3, CC49 and CC68 were recovered almost exclusively from RT infections, which suggests that the strains in these CCs have strong tropism to the respiratory tract or are unable to cause invasive infection. The distribution of *emm* types varied greatly in different CCs: The RT-associated CCs presented far less diversity than most others: CC49 and CC68 are exclusively associated with *stC36*, and CC3 is mainly associated with *stC839*, but also includes *emm57* and *stG653*. Although the expected rates of spontaneous mutations in SDSE are not well defined, the greater diversity of *emm* types observed in CC17, CC29 and CC15 suggests that these CCs have undergone considerable diversification since their origin (Oppegaard et al., 2017), unlike CC20, CC49 and CC68. Typing by cgMLST confirmed these differences in diversity and that CC3, CC49 and CC68 are relatively distant from each other and less diverse than other CCs, which indicates that these lineages originated independently from each other.

4.2. Virulence Factors

The genomic sequences of different *emm* types were translated into amino acid sequences, aligned with MUSCLE, and a neighbour-joining tree was constructed with these sequences. STs *stC36* and *stC839* formed a distinct group that included *stC74A*, which was associated with invasive infections, and *stG485*, *stG166B* and *stC245*, which were not significantly associated with either type of infection. Furthermore, no other *emm* type besides *stC36* and *stC839* showed such a strong association with either type of infection. This indicates that M protein sequence identity does not predict potential to cause invasive or RT infections. When the distribution of VFs per CC was analysed, *mf3* was present in 86.7% (n=13) of CC3 isolates, 78% (n=11) of CC49 isolates and 92.3% (n=12) of CC68 isolates. This gene encodes the protein Mitogen factor 3 (MF3), an extracellular DNase. The genomic context of *mf3* in RT isolates is varied and it is usually surrounded by small phage genes (data not shown) and a UniProt BLAST search of the gene results in several 100% identity matches with several prophage extracellular DNases. Furthermore, a study on SP with a primate pharyngitis model found this gene and other DNases to increase fitness in for the primate oropharynx (Zhu et al., 2020). Some authors hypothesise that *mf3* contributes to increased virulence by two main mechanisms: immune evasion by degradation of neutrophil extracellular traps, and increased fitness by use of DNA as a nutrient source. However, these hypotheses are tentative, as the roles of specific DNases have not yet been fully elucidated in SP or SDSE (Wen et al., 2011), and there is no strong evidence to support direct involvement of this gene in

differences in virulence between CCs associated with RT infection and other CCs.

4.3. Pan-Genome and Gene Association

The set of 631 isolates recovered from invasive infections, RT infections and five additional *stC36* or *stC839* isolates generated a pan-genome comprised of 10 684 genes. The accessory genome, with 9275 genes, is comprised of 1155 shell genes, present in 15% to 95% of the isolates, and 8120 cloud genes, present in fewer than 15% of the isolates. The accessory genome contains the genes responsible for adaptation to selective pressures such as antimicrobial resistance, and adaptation to environmental niches, like the colonisation of a new host or tissue tropism (Daubin & Ochman, 2004; Medini et al., 2005). In total, 17 genes whose presence may be relevant for differences in virulence and 24 genes whose absence may be relevant were found when RT-associated CC isolates were compared to all others. No genes with both very high sensitivity and specificity for the RT-associated CCs were found. The gene with the highest specificity and sensitivity for these CCs was *group_5532*. Sequence analysis revealed that this is the same gene as *mf3*. In the set whose presence may be relevant, several genes involved in the synthesis of rhamnose-containing cell wall polysaccharides were present. These polysaccharides are critical for virulence in streptococci and play a significant role as bacteriophage receptors and interaction with hosts (Guérin et al., 2022; Mistou et al., 2016). The biosynthesis of Lancefield group A, B, C and G antigens, for example, are initiated by a rhamnosyltransferase (Zorzoli et al., 2019). In the set of genes whose absence in RT CCs is significant, CRISPR-associated endonuclease genes and a pre-crRNA processing endonuclease are among the genes with the lowest specificity. CRISPR/Cas systems play a vital role in bacterial protection against invading phages and plasmids, and the absence of these genes can impair the anti-phage immunity of these strains (Deltcheva et al., 2011; Hochstrasser & Doudna, 2015; Marraffini, 2016). The absence of these essential genes against phage defence, together with the fact that *mf3* may have a phage related origin, suggests possible phage involvement in differences in virulence in the RT CCs.

4.4. Antimicrobial Resistance

All the strains that were characterised in this work were susceptible to penicillin, vancomycin, and linezolid. To date, only one case of penicillin resistance has been documented in SDSE (Fuursted et al., 2016) and this antimicrobial remains the first line of therapy against SDSE infections (Barros, 2021). No cases of resistance to vancomycin or linezolid have been reported (Barros, 2021; Broyles et al., 2009; Loubinoux et al., 2013; Lu et al., 2016). The rate of resistance to erythromycin remains the highest among the tested antibiotics. No significant differences in antimicrobial susceptibility were observed between RT and invasive infections or between different CCs.

5. Conclusions

Typing of the *emm* gene, which encodes an important virulence factor in SDSE, revealed two *emm* types to be almost exclusively associated with respiratory tract (RT) infections in Portugal. RT infections in Portugal were more frequent among younger patients. RT infection patients had a mean age of 27.7 years, compared to 69.8 years in invasive infection patients. Typing of the *emm* gene found 31 distinct *emm* types, of which *stC36* and *stC839* were almost exclusively associated with RT infections. MLST revealed 121 sequence types (ST) distributed among 37 clonal complexes (CC) and 72 singletons, with 3 distantly related CCs, CC3, CC49 and CC68, being almost exclusively associated with RT infection, and cgMLST supported the grouping of strains with those allelic profiles and showed that they are genetically distant from each other. While some CCs had a diverse variety of *emm* types, the CCs associated with the RT were associated with the RT *emm* types. Among known streptococcal virulence factors, *mf3* was almost exclusively associated to RT CCs. GWAS revealed that RT CCs lacked several CRISPR-associated proteins, which serve important roles in protection against phages, and had several phage genes, which may affect virulence.

5.1. Future Work

This work contributes towards the knowledge of human respiratory tract by SDSE in Portugal. Future work may employ high throughput sequencing in the study of colonisation strains to evaluate the main clonal lineages responsible for colonisation and identify genomic differences between colonisation and infection strains that may influence pathogenic potential in different clonal lineages.

Without information on gene expression, no strong conclusions can be drawn on the effects of specific virulence factor genes can from the presence of a gene in strains of different clonal lineages. Genomic expression studies would elucidate whether these genes are actually involved in pathogenic potential and invasion studies on with human cell lines can provide more in-depth knowledge on varying degrees of pathogenic potential. Analysis of individual VF loci and allelic profiling of VF genes and regulators can provide knowledge on how different variants impact pathogenic potential.

The finding of a phage-associated nuclease in the RT CCs, along with the absence of CRISPR-associated genes, warrants the identification and analysis of genetic mobile elements in RT strains and their association with virulence factors, and how the differences in these CRISPR loci may affect phage defence mechanisms.

A GWAS based on presence and absence of genes offers no insight on how gene variants may affect virulence. A k-mer approach may find gene variants that influence pathogenic potential in the clonal lineages related to the respiratory tract.

Finally, while no atypical patterns of antimicrobial resistance were found, continued epidemiological surveillance is important to monitor antimicrobial resistance and to detect the emergence of new clonal lineages.

6. Acknowledgements

This work was performed at the Molecular Microbiology and Infection lab, Instituto de Medicina Molecular João Lobo Antunes, University of Lisbon (Lisbon, Portugal), during the period of September 2021 to November 2022, under the supervision of Dr. Marcos Pinho and co-supervised at Instituto Superior Técnico by Prof. Leonilde Moreira.

7. References

Agerhäll, M., Henrikson, M., Johansson Söderberg, J., Sellin, M., Tano, K., Gylfe, Å., & Berggren, D. (2021). High prevalence of pharyngeal bacterial pathogens among healthy adolescents and young adults. *Apmis*, *129*(12), 711–716. https://doi.org/10.1111/apm.13179

Baracco, G. J. (University of M. M. S. of M. and M. V. A. (2019). Infections Caused by Group C and G Streptococcus (Streptococcus dysgalactiae subsp. equisimilis and Others): Epidemiological and Clinical Aspects. *Microbiology Spectrum*, *163*(HS1), 1–11. https://doi.org/10.1016/S0035-3787(07)92157-8

Barros, R. R. (2021). Antimicrobial Resistance among Beta-Hemolytic Streptococcus in Brazil: An Overview. *Antibiotics*, *10*(8). https://doi.org/10.3390/ANTIBIOTICS10080973

Bramhachari, P. V., Kaul, S. Y., McMillan, D. J., Shaila, M. S., Karmarkar, M. G., & Sriprakash, K. S. (2010). Disease burden due to Streptococcus dysgalactiae subsp. equisimilis (group G and C streptococcus) is higher than that due to Streptococcus pyogenes among Mumbai school children. *Journal of Medical Microbiology*, *59*(2), 220–223. https://doi.org/10.1099/jmm.0.015644-0

Brandt, C. M., & Spellerberg, B. (2009a). Human infections due to streptococcus dysgalactiae subspedes equisimilis. *Clinical Infectious Diseases*, *49*(5), 766–772. https://doi.org/10.1086/605085

Brandt, C. M., & Spellerberg, B. (2009b). Human infections due to streptococcus dysgalactiae subspedes equisimilis. *Clinical Infectious Diseases*, *49*(5), 766–772. https://doi.org/10.1086/605085

Broyles, L. N., Van Beneden, C., Beall, B., Facklam, R., Lynn Shewmaker, P., Malpiedi, P., Daily, P., Reingold, A., & Farley, M. M. (2009). Population-based study of invasive disease due to β-Hemolytic streptococci of groups other than A and B. *Clinical Infectious Diseases*, *48*(6), 706–712. https://doi.org/10.1086/597035

Brynildsrud, O., Bohlin, J., Scheffer, L., & Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, *17*(1), 1–9. https://doi.org/10.1186/S13059-016-1108-8/FIGURES/4

Cannon, J. W., Zhung, J., Bennett, J., Moreland, N. J., Baker, M. G., Geelhoed, E., Fraser, J., Carapetis, J. R., & Jack, S. (2021). The economic and health burdens of diseases caused by group A Streptococcus in New Zealand. *International Journal of Infectious Diseases*, *103*, 176–181. https://doi.org/10.1016/J.IJID.2020.11.193

Castro, A. C. (2020). *Caraterização genómica de Streptococcus dysgalactiae subsp . equisimilis responsável por infeção invasiva no Homem*.

Ciszewski, M., & Szewczyk, E. M. (2017). Potential Factors Enabling Human Body Colonization by Animal Streptococcus dysgalactiae subsp. equisimilis Strains. *Current Microbiology*, *74*(5), 650–654. https://doi.org/10.1007/s00284-017-1232-z

CLSI. (2020). CLSI M100-ED29: 2021 Performance Standards for Antimicrobial Susceptibility Testing, 30th Edition. In *Clsi* (Vol. 40, Issue 1).

Cohen-Poradosu, R., Jaffe, J., Lavi, D., Grisariu-Greenzaid, S., Nir-Paz, R., Valinsky, L., Dan-Goor, M., Block, C., Beall, B., & Moses, A. E. (2004). Group G Streptococcal Bacteremia in Jerusalem - Volume 10, Number 8—August 2004 - Emerging Infectious Diseases journal - CDC. *Emerging Infectious Diseases*, *10*(8), 1455–1460. https://doi.org/10.3201/EID1008.030840

Daubin, V., & Ochman, H. (2004). Bacterial genomes as new gene homes: The genealogy of ORFans in E. coli. *Genome Research*, *14*(6), 1036–1042. https://doi.org/10.1101/gr.2231904

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602. https://doi.org/10.1038/NATURE09886

Farhat, M. R., Shapiro, B. J., Kieser, K. J., Sultana, R., Jacobson, K. R., Victor, T. C., Warren, R. M., Streicher, E. M., Calver, A., Sloutsky, A., Kaur, D., Posey, J. E., Plikaytis, B., Oggioni, M. R., Gardy, J. L., Johnston, J. C., Rodrigues, M., Tang, P. K. C., Kato-Maeda, M., … Murray, M. (2013). Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nature Genetics*, *45*(10), 1183–1189. https://doi.org/10.1038/NG.2747

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. *Journal of Bacteriology*, *186*(5), 1518–1530. https://doi.org/10.1128/JB.186.5.1518-1530.2004

Fischetti, V. A. (1989). Streptococcal M protein: Molecular design and biological behavior. *Clinical Microbiology Reviews*, *2*(3), 285–314. https://doi.org/10.1128/CMR.2.3.285

Francisco, A. P., Bugalho, M., Ramirez, M., & Carriço, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, *10*(1), 1–15. https://doi.org/10.1186/1471-2105-10-152/FIGURES/5

Fuursted, K., Stegger, M., Hoffmann, S., Lambertsen, L., Andersen, P. S., Deleuran, M., & Thomsen, M. K. (2016a). Description and characterization of a penicillin-resistant Streptococcus dysgalactiae subsp. equisimilis clone isolated from blood in three epidemiologically linked patients. *Journal of Antimicrobial Chemotherapy*, *71*(12), 3376–3380. https://doi.org/10.1093/JAC/DKW320

Fuursted, K., Stegger, M., Hoffmann, S., Lambertsen, L., Andersen, P. S., Deleuran, M., & Thomsen, M. K. (2016b). Description and characterization of a penicillin-resistant Streptococcus dysgalactiae subsp. equisimilis clone isolated from blood in three epidemiologically linked patients. *Journal of Antimicrobial Chemotherapy*, *71*(12), 3376–3380. https://doi.org/10.1093/JAC/DKW320

Gherardi, G., Imperi, M., Palmieri, C., Magi, G., Facinelli, B., Baldassarri, L., Pataracchia, M., & Creti, R. (2013). Genetic diversity and virulence properties of Streptococcus dysgalactiae subsp. Equisimilis from different sources. *Journal of Medical Microbiology*, *63*(PART 1), 90–98. https://doi.org/10.1099/jmm.0.062109-0

Guérin, H., Kulakauskas, S., & Chapot-Chartier, M. P. (2022). Structural variations and roles of rhamnose-rich cell wall polysaccharides in Gram-positive bacteria. *Journal of Biological Chemistry*, *298*(10), 102488. https://doi.org/10.1016/j.jbc.2022.102488

Haenni, M., Lupo, A., & Madec, J.-Y. (2018). Antimicrobial Resistance in Streptococcus spp . *Microbiology Spectrum*, *6*(2). https://doi.org/10.1128/MICROBIOLSPEC.ARBA-0008-2017/ASSET/44046F2D-90DC-4275-A357-7A743035CC26/ASSETS/GRAPHIC/ARBA-0008-2017-FIG1.GIF

Hochstrasser, M. L., & Doudna, J. A. (2015). Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends in Biochemical Sciences*, *40*(1), 58–66. https://doi.org/10.1016/j.tibs.2014.10.007

Holt, K. E., Wertheim, H., Zadoks, R. N., Baker, S., Whitehouse, C. A., Dance, D., Jenney, A., Connor, T. R., Hsu, L. Y., Severin, J., Brisse, S., Cao, H., Wilksch, J., Gorrie, C., Schultz, M. B., Edwards, D. J., Van Nguyen, K., Nguyen, T. V., Dao, T. T., … Thomson, N. R. (2015). Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(27), E3574–E3581. https://doi.org/10.1073/PNAS.1501049112

Hunter, P. R., & Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *Journal of Clinical Microbiology*, *26*(11), 2465–2466.

https://doi.org/10.1128/JCM.26.11.2465-2466.1988
Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-119
Johansson, H. M., Mo, M., & Frick, I. (2004). *Protein FOG – a streptococcal inhibitor of neutrophil function*. https://doi.org/10.1099/mic.0.27269-0
Kittang, B. R., Skrede, S., Langeland, N., Haanshuus, C. G., & Mylvaganam, H. (2011). emm gene diversity, superantigen gene profiles and presence of SlaA among clinical isolates of group A, C and G streptococci from western Norway. *European Journal of Clinical Microbiology & Infectious Diseases : Official Publication of the European Society of Clinical Microbiology*, *30*(3), 423–433. https://doi.org/10.1007/S10096-010-1105-X
Kline, J. B., Xu, S., Bisno, A. L., & Collins, C. M. (1996). Identification of a fibronectin-binding protein (GfbA) in pathogenic group G streptococci. *Infection and Immunity*, *64*(6), 2122–2129. https://doi.org/10.1128/IAI.64.6.2122-2129.1996
Korczynska, J. E., Turkenburg, J. P., & Taylor, E. J. (2012). The structural characterization of a prophage-encoded extracellular DNase from Streptococcus pyogenes. *Nucleic Acids Research*, *40*(2), 928. https://doi.org/10.1093/NAR/GKR789
Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/NAR/GKM160
Leclercq, R. (2002). Mechanisms of resistance to macrolides and lincosamides: nature of the resistance elements and their clinical implications. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *34*(4), 482–492. https://doi.org/10.1086/324626
LINDGREN, P. -E, McGAVIN, M. J., SIGNÄS, C., GUSS, B., GURUSIDDAPPA, S., HÖÖK, M., & LINDBERG, M. (1993). Two different genes coding for fibronectin-binding proteins from Streptococcus dysgalactiae. *European Journal of Biochemistry*, *214*(3), 819–827. https://doi.org/10.1111/J.1432-1033.1993.TB17985.X
Lo, H. H., & Cheng, W. S. (2015). Distribution of virulence factors and association with emm polymorphism or isolation site among beta-hemolytic group G Streptococcus dysgalactiae subspecies equisimilis. *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, *123*(1), 45–52. https://doi.org/10.1111/APM.12305
Lother, S. A., Demczuk, W., Martin, I., Mulvey, M., Dufault, B., Lagacé-Wiens, P., & Keynan, Y. (2017). Clonal Clusters and Virulence Factors of Group C and G Streptococcus Causing Severe Infections, Manitoba, Canada, 2012-2014. *Emerging Infectious Diseases*, *23*(7), 1092–1101. https://doi.org/10.3201/EID2307.161259
Loubinoux, J., Plainvert, C., Collobert, G., Touak, G., Bouvet, A., & Poyart, C. (2013). Adult Invasive and Noninvasive Infections Due to Streptococcus dysgalactiae subsp. equisimilis in France from 2006 to 2010. *Journal of Clinical Microbiology*, *51*(8), 2724. https://doi.org/10.1128/JCM.01262-13
Lu, B., Fang, Y., Huang, L., Diao, B., Du, X., Kan, B., Cui, Y., Zhu, F., Li, D., & Wang, D. (2016). Molecular characterization and antibiotic resistance of clinical Streptococcus dysgalactiae subsp. equisimilis in Beijing, China. *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *40*, 119–125. https://doi.org/10.1016/J.MEEGID.2016.01.030
Maiden, M. C. J. (2006). Multilocus sequence typing of bacteria. *Annual Review of Microbiology*, *60*, 561–588. https://doi.org/10.1146/annurev.micro.59.030804.121325
Marraffini, L. A. (2016). The CRISPR-Cas system of Streptococcus pyogenes: function and applications. *Streptococcus Pyogenes: Basic Biology to Clinical Manifestations*. https://www.ncbi.nlm.nih.gov/books/NBK355562/
Matsue, M., Ogura, K., Sugiyama, H., Miyoshi-Akiyama, T., Takemori-Sakai, Y., Iwata, Y., Wada, T., & Okamoto, S. (2020). Pathogenicity Characterization of Prevalent-Type Streptococcus dysgalactiae subsp. equisimilis Strains. *Frontiers in Microbiology*, *11*(February), 1–13. https://doi.org/10.3389/fmicb.2020.00097
McDonald, M., Towers, R. J., Andrews, R. M., Carapetis, J. R., & Currie, B. J. (2007). Epidemiology of Streptococcus dysgalactiae subsp. equisimilis in Tropical Communities, Northern Australia. *Emerging Infectious Diseases*, *13*(11), 1694. https://doi.org/10.3201/EID1311.061258
McMillan, D. J., Bessen, D. E., Pinho, M., Ford, C., Hall, G. S., Melo-Cristino, J., & Ramirez, M. (2010). Population genetics of Streptococcus dysgalactiae subspecies equisimilisreveals widely dispersed clones and extensive recombination. *PLoS ONE*, *5*(7). https://doi.org/10.1371/journal.pone.0011741
Mcmillan, D. J., Drèze, P. A., Vu, T., Bessen, D. E., Guglielmini, J., Steer, A. C., Carapetis, J. R., Van Melderen, L., Sriprakash, K. S., Smeesters, P. R., Michael Batzloff, Towers, R., Goossens, H., Malhotra-Kumar, S., Guilherme, L., RosangelaTorres, Low, D., Mc Geer, A., Krizova, P., … Tanz, R. (2013). Updated model of group A Streptococcus M proteins based on a comprehensive worldwide study. *Clinical Microbiology and Infection*, *19*(5), E222–E229. https://doi.org/10.1111/1469-0691.12134
McMillan, D. J., Kaul, S. Y., Bramhachari, P. V., Smeesters, P. R., Vu, T., Karmarkar, M. G., Shaila, M. S., & Sriprakash, K. S. (2011). Recombination Drives Genetic Diversification of Streptococcus dysgalactiae Subspecies equisimilis in a Region of Streptococcal Endemicity. *PLOS ONE*, *6*(8), e21346. https://doi.org/10.1371/JOURNAL.PONE.0021346
McNeilly, C. L., & McMillan, D. J. (2014). Horizontal gene transfer and recombination in Streptococcus dysgalactiae subsp. equisimilis. *Frontiers in Microbiology*, *5*(DEC), 676. https://doi.org/10.3389/FMICB.2014.00676/BIBTEX
Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594. https://doi.org/10.1016/J.GDE.2005.09.006
Mistou, M. Y., Sutcliffe, I. C., & Van Sorge, N. M. (2016). Bacterial glycobiology: Rhamnose-containing cell wall polysaccharides in gram-positive bacteria. *FEMS Microbiology Reviews*, *40*(4), 464–479. https://doi.org/10.1093/femsre/fuw006
Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A., & Vaz, C. (2017). PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics*, *33*(1), 128–129. https://doi.org/10.1093/BIOINFORMATICS/BTW582
Oppegaard, O., Mylvaganam, H., Skrede, S., Lindemann, P. C., & Kittang, B. R. (2017). Emergence of a Streptococcus dysgalactiae subspecies equisimilis stG62647-lineage associated with severe clinical manifestations. *Scientific Reports 2017 7:1*, *7*(1), 1–10. https://doi.org/10.1038/s41598-017-08162-z
Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/BIOINFORMATICS/BTV421
Park, J. S., Kim, H. W., Chae, J. D., Hur, J. W., Park, Y. S., Yoo, M. S., & Ryu, H. Y. (2016). Identification of streptococcus dysgalactiae subsp. equisimilis from septic knee by 16S rRNA gene sequencing. *Kosin Medical Journal*, *31*(1), 79. https://doi.org/10.7180/kmj.2016.31.1.79
Pinho, M. D. C. B. de. (2014). Lancefield group C and G streptococci in human infection: Molecular typing, virulence and antimicrobial susceptibility. *PQDT - Global*, 238. http://ezaccess.libraries.psu.edu/login?url=https://search.proquest.com/docview/1991489049?accountid=13158%250Ahttp://sk8es4mc2l.search.serialssolutions.com?ctx_

ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF-8&rfr_id=info:sid/ProQuest+Dissertations+%2526+Theses+A%252

Pinho, M. D., Melo-Cristino, J., & Ramirez, M. (2006). Clonal relationships between invasive and noninvasive Lancefield group C and G streptococci and emm-specific differences in invasiveness. *Journal of Clinical Microbiology*, *44*(3), 841–846. https://doi.org/10.1128/JCM.44.3.841-846.2006

Rantala, S., Vähäkuopus, S., Vuopio-Varkila, J., Vuento, R., & Syrjänen, J. (2010). Streptococcus dysgalactiae subsp. equisimilis Bacteremia, Finland, 1995-2004. *Emerging Infectious Diseases*, *16*(5), 843–846. https://doi.org/10.3201/eid1605.080803

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H. L., Allerberger, F., Harmsen, D., & Mellmann, A. (2015). Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of Listeria monocytogenes. *Journal of Clinical Microbiology*, *53*(9), 2869–2876. https://doi.org/10.1128/JCM.01193-15

Sang-Ik Oh, Jong Wan Kim, Jongho Kim, Byungjae So, Bumseok Kim, & Ha-Young Kim. (2020). *Molecular subtyping and antimicrobial susceptibility of Streptococcus dysgalactiae subspecies equisimilis isolates from clinically diseased pigs*. https://doi.org/10.4142/jvs.2020.21.e57

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/BIOINFORMATICS/BTU153

Severiano, A., Pinto, F. R., Ramirez, M., & Carriço, J. A. (2011). Adjusted Wallace coefficient as a measure of congruence between typing methods. *Journal of Clinical Microbiology*, *49*(11), 3997–4000. https://doi.org/10.1128/JCM.00624-11

Shimomura, Y., Okumura, K., Murayama, S. Y., Yagi, J., Ubukata, K., Kirikae, T., & Miyoshi-Akiyama, T. (2011). Complete genome sequencing and analysis of a Lancefield group G Streptococcus dysgalactiae subsp. equisimilis strain causing streptococcal toxic shock syndrome (STSS). *BMC Genomics*, *12*, 17. https://doi.org/10.1186/1471-2164-12-17

Streptococcus Laboratory: M Protein Gene (emm) Typing | CDC. (n.d.). Retrieved February 4, 2022, from https://www.cdc.gov/streplab/groupa-strep/emm-background.html

Sumby, P., Barbian, K. D., Gardner, D. J., Whitney, A. R., Welty, D. M., Long, R. D., Bailey, J. R., Parnell, M. J., Hoe, N. P., Adams, G. G., DeLeo, F. R., & Musser, J. M. (2005). Extracellular deoxyribonuclease made by group A Streptococcus assists pathogenesis by enhancing evasion of the innate immune response. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(5), 1679–1684. https://doi.org/10.1073/PNAS.0406641102/SUPPL_FILE/06641FIG10.PDF

Sunaoshi, K., Murayama, S. Y., Adachi, K., Yagoshi, M., Okuzumi, K., Chiba, N., Morozumi, M., & Ubukata, K. (2010). Molecular emm genotyping and antibiotic susceptibility of Streptococcus dysgalactiae subsp. equisimilis isolated from invasive and non-invasive infections. *Journal of Medical Microbiology*, *59*(1), 82–88. https://doi.org/10.1099/jmm.0.013201-0

Traverso, F., Blanco, A., Villalón, P., Beratz, N., Nieto, J. A. S., & Lopardo, H. (2016). Molecular characterization of invasive Streptococcus dysgalactiae subsp. Equisimilis. Multicenter study: Argentina 2011-2012. *Revista Argentina de Microbiología*, *22*(2), 247–254. https://doi.org/10.1016

Turner, C. E., Bubba, L., & Efstratiou, A. (2019). Pathogenicity factors in group C and G streptococci. *Gram-Positive Pathogens*, *May 2019*, 264–274. https://doi.org/10.1128/9781683670131.ch16

Turner, J. C., Hayden, F. G., Lobo, M. C., Ramirez, C. E., Murren, D., Al, T. E. T., & Icrobiol, J. C. L. I. N. M. (1997). *Epidemiologic Evidence for Lancefield Group C Beta-Hemolytic Streptococci as a Cause of Exudative Pharyngitis in College Students*. *35*(1), 1–4.

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., & Malorny, B. (2020). Typing methods based on whole genome sequencing data. *One Health Outlook*, *2*(1), 1–19. https://doi.org/10.1186/s42522-020-0010-1

van Belkum, A., Tassios, P. T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N. K., Fussing, V., Green, J., Feil, E., Gerner-smidt, P., Brisse, S., & Struelens, M. (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *13 Suppl 3*(SUPPL. 3), 1–46. https://doi.org/10.1111/J.1469-0691.2007.01786.X

Wajima, T., Morozumi, M., Hanada, S., Sunaoshi, K., Chiba, N., Iwata, S., & Ubukata, K. (2016). Molecular Characterization of Invasive Streptococcus dysgalactiae subsp. equisimilis, Japan. *Emerging Infectious Diseases*, *22*(2), 247. https://doi.org/10.3201/EID2202.141732

Watanabe, S., Kirikae, T., & Miyoshi-Akiyama, T. (2013). Complete genome sequence of Streptococcus dysgalactiae subsp. equisimilis 167 carrying Lancefield group C antigen and comparative genomics of S. dysgalactiae subsp. equisimilis strains. *Genome Biology and Evolution*, *5*(9), 1644–1651. https://doi.org/10.1093/GBE/EVT117

Wen, Y. T., Tsou, C. C., Kuo, H. T., Wang, J. S., Wu, J. J., & Liao, P. C. (2011). Differential secretomics of Streptococcus pyogenes reveals a novel peroxide regulator (PerR)-regulated extracellular virulence factor mitogen factor3 (MF3). *Molecular and Cellular Proteomics*, *10*(9), 1–11. https://doi.org/10.1074/mcp.M110.007013

Zhu, L., Olsen, R. J., Beres, S. B., Saavedra, M. O., Kubiak, S. L., Cantu, C. C., Jenkins, L., Waller, A. S., Sun, Z., Palzkill, T., Porter, A. R., DeLeo, F. R., & Musser, J. M. (2020). Streptococcus pyogenes genes that promote pharyngitis in primates. *JCI Insight*, *5*(11). https://doi.org/10.1172/jci.insight.137686

Zorzoli, A., Meyer, B. H., Adair, E., Torgov, V. I., Veselovsky, V. V., Danilov, L. L., Uhrin, D., & Dorfmueller, H. C. (2019). Group A, B, C, and G Streptococcus Lancefield antigen biosynthesis is initiated by a conserved α-D-GlcNAc-β-1,4-L-rhamnosyltransferase. *Journal of Biological Chemistry*, *294*(42), 15237–15256. https://doi.org/10.1074/jbc.RA119.009894