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

Filipe Barbosa Valcovo

Thesis to obtain the Master of Science Degree in

Microbiology

Supervisors: Doutor Marcos Daniel Caetano Borges de Pinho

Professora Leonilde de Fátima Morais Moreira

Examination Committee

Chairperson: Professor Jorge Humberto Gomes Leitão

Supervisor: Doutor Marcos Daniel Caetano Borges de Pinho

Members of committee: Professor Francisco Rodrigues Pinto

November 2022

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis 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. The thesis was co-supervised at Instituto Superior Técnico by Prof. Leonilde Moreira.

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Marcos Pinho, for entrusting me with the project that led to this thesis and for all the guidance, support and shared knowledge along the way. I also want to thank Professor Mário Ramirez for all the insightful questions and for the encouragement I was offered.

To my colleagues and researchers in the Molecular Microbiology and Infection Unit, I am grateful for all the collaboration and contributions, without which this work would not have been possible. To Joana, thank you for all the advice, patience and eagerness to help at any time, inside and outside the lab. To my colleagues in the bioinformatics team, thank you for the collaboration and advice.

To Inês, thank you for your friendship, for the company in the late hours of lab work and writing, and for all the advice, in and outside the lab. To Ivan, Koen, Lucas and Tiago, thank you for your friendship, lighters and all the amazing food we shared. To all my friends, you all contributed to my success in different ways, and I am grateful to have you in my life.

To my family, especially my parents and my sister, I am deeply grateful for all the support, sacrifices and all the opportunities that have allowed me to come this far.

Abstract

Human infections caused by *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) have been rising in recent decades. This bacterium is genetically similar to *S. pyogenes* (SP), in the spectrum of diseases it causes, and in the virulence factors it possesses.

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 association motivated the genomic characterisation of 199 isolates recovered from RT infections in the years 2011 to 2019 to define the main genetic lineages responsible for RT infection in Portugal, to identify virulence factors that may be involved in respiratory tract tropism, and to determine antimicrobial resistance in these isolates, using genomic data from invasive infections for comparison purposes.

Phenotypic methods included Lancefield group determination and antimicrobial susceptibility testing, while genotypic characterisation employed 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. To date, no large-scale gene association studies performed on SDSE RT infection isolates have been published.

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. 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 with the RT-associated CCs. The GWAS revealed that RT-associated CCs lacked several CRISPR-associated proteins, which serve important roles in protection against phages, and had several phage genes, which may affect virulence.

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

Resumo

Infecções humanas por *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) têm aumentado nas últimas décadas. Esta bactéria é geneticamente próxima de *S. pyogenes* (SP), no espectro de doenças que causa e nos factores de virulência que possui. Tipagem do gene *emm*, que codifica um factor de virulência importante em SDSE, revelou dois tipos *emm* com associação quase exclusiva a infecções do tracto respiratório.

Esta descoberta motivou a caracterização genómica de 199 estirpes provenientes de infecções do tracto respiratório em Portugal durante os anos 2011 a 2019 para definir as principais linhagens genéticas responsáveis por infecção do tracto respiratório, para identificar factores de virulência que possam ser responsáveis por tropismo para o tracto respiratório e para determinar os perfis de resistência a agentes antimicrobianos.

Os métodos fenotípicos incluíram a determinação do grupo de Lancefield e testes de susceptibilidade a agentes antimicrobianos, enquanto a caracterização genotípica empregou sequenciação de alto débito para determinar os tipos *emm*, *multilocus sequence typing* (MLST), *core genome* MLST (cgMLST), factores de virulência conhecidos, e um estudo de associação genética para explorar factores de virulência possivelmente desconhecidos. Até à data, não foram publicados estudos de associação genética de grande escala em SDSE associado a infecções respiratórias.

As infecções do tracto respiratório causadas por SDSE em Portugal foram mais frequentes em pacientes mais jovens. Os pacientes de infecção respiratória tiveram uma idade média de 27.7 anos de idade, enquanto os pacientes de infecção invasiva tiveram uma idade média de 69.8 anos. A tipagem do gene *emm* revelou 31 *sequence types* (ST) distintos, dos quais *stc36* e *stC839* estavam quase exclusivamente associados a infecções respiratórias. A MLST revelou 121 STs distribuídos por 37 complexos clonais (CC) e por 72 *singletons*. Três CCs, CC3, CC49 e CC68 estavam quase exclusivamente associados a infecção respiratória. Enquanto alguns CCs possuíam vários tipos *emm,* os CCs associados ao tracto respiratório estavam também associados aos tipos *emm stC36* e *stC839.* Entre os factores de virulência conhecidos, o gene *mf3* estava quase exclusivamente associado aos CCs do tracto respiratório. O estudo de associação genética mostrou que as estirpes destes CCs não possuíam várias proteínas do sistema CRISPR/Cas, cujo papel é importante na protecção contra fagos.

Palavras-chave: *Streptococcus dysgalactiae* subsp. *equisimilis*, infecção invasiva, infecção do tracto respiratório, sequenciação de alto débito, estudo de associação genética, tipagem.

Contents

List of Figures

List of Tables

Abbreviations

ATP - Adenosine triphosphate

BLASTP - Protein basic local alignment search tool

CC - Clonal complex

CG - Core genome

cgMLST – Core genome multilocus sequence typing

CLSI - Clinical and Laboratory Standards Institute

CN - Gentamycin

DA - clindamycin

DLV - Double *locus* variant

DNA - Deoxyribonucleic acid

FBP - Fibronectin-binding protein

FCT - Fibronectin and collagen-binding protein

FOG - Fibrinogen-binding protein of group G streptococci

GAS - Group A Streptococcus

GCGS - Group C and G Streptococcus

GCS - Group C Streptococcus

GGS - Group G Streptococcus

GWAS - Genome-wide association study

HGT - Horizontal gene transfer

HIV - Human immunodeficiency virus

HTS - High throughput sequencing

LEV - Levofloxacin

MALDI-TOF MS - Matrix-assisted laser desorption/ionization-time of flight mass spectrometry

MCL - Markov clustering algorithm

MF3 - Mitogen factor 3

MIC - Minimum inhibitory concentration

MLSA - Multilocus sequence analysis

MLSB - Macrolide-lincosamide-streptogramin B

MLST - Multilocus sequence typing

MUSCLE - Multiple sequence comparison by log-expectation

NAD - Nicotinamide adenine dinucleotide

NET - Neutrophil extracellular traps

OAI - Osteoarticular infection

OR - Odds ratio

PBP - Penicillin-binding protein

PCR - Polymerase chain reaction

PSS - Poststreptococcal sequelae

rDNA - Ribosomal DNA

RNA - Ribonucleic acid

RT - Respiratory tract

SDSD - *Streptococcus dysgalactiae* subsp. *dysgalactiae*

SDSE - *Streptococcus dysgalactiae* subsp. *equisimilis*

SLO - Streptolysin O

SLS - Streptolysin S

SLV - Single *locus* variant

SNP - Single nucleotide polymorphism

SOF - Serum opacity factor

SP – *Streptococcus pyogenes*

SPE - Streptococcal pyrogenic exotoxin

SST - Skin and soft tissue

ST - Sequence type

STSS - Streptococcal toxic shock syndrome

tDNA - Transfer DNA

TE - Tetracycline

TSA - Tryptone soy agar

TSB - Tryptone soy broth

VF - Virulence factor

VFDB - Virulence factor database

VP - Voges-Proskauer reaction

WGS - Whole genome sequencing

Chapter 1. Introduction

1.1. General Description of *Streptococcus dysgalactiae* subsp. *equisimilis*

Human infections caused by *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) have been rising in the past four decades. This bacterium is genetically similar to *Streptococcus 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; Pinho, 2014).

1.1.1. Morphology, Physiology, and Habitat

SDSE is a Gram-positive facultative anaerobic bacterium. The cells are cocci or ovoid, occurring in pairs or chains. Colonies are greyish and display a wide zone of beta-haemolysis when grown on blood agar (Figure 1).

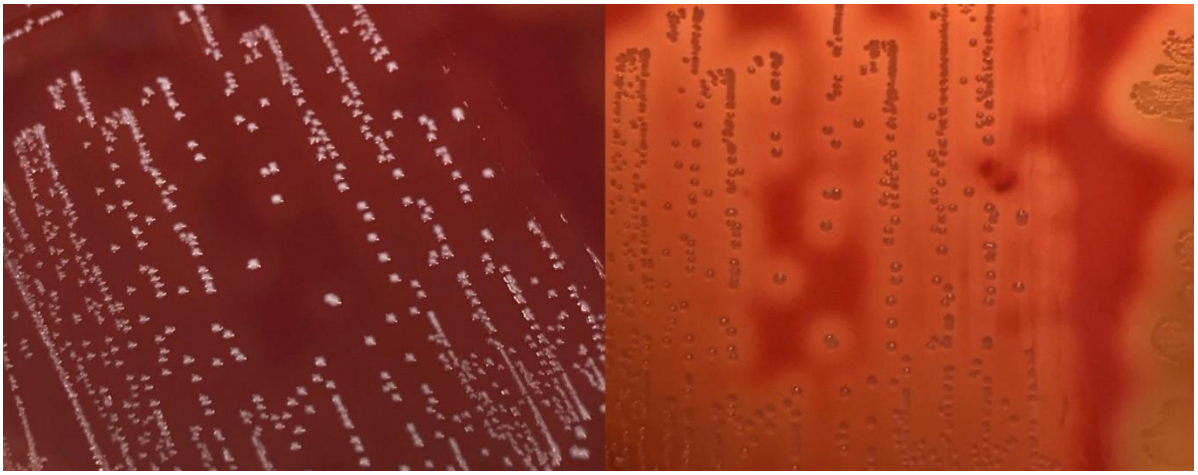

Figure 1. SDSE colonies grown for 24 h at 35 ºC in tryptone soy agar supplemented with 5% (v/v) sheep blood. (A) Typical SDSE colonies seen with above lighting; (B) Typical SDSE colonies, seen with back lighting to evidence beta-haemolysis.

The optimal temperature for growth is around 37 ºC, with growth being inhibited at 10 ºC and 45 ºC. SDSE is fastidious, growing only on complex media, usually complemented with blood or serum, and it has a fermentative metabolism (Spellberg & Brandt, 2015). Like the other members of the genus, it is catalase negative. SDSE produces acid from starch, glucose, maltose, ribose, sucrose, and trehalose, but not from arabinose, inulin, mannitol, or raffinose. Acid is also produced from glycerol and lactose by most strains when grown aerobically. It is negative for the Vogues-Proskauer test. Streptokinase activity occurs on human plasminogen (Farrow & Collins, 1984; Vandamme et al., 1996; Vieira et al., 1998). The vast majority of

isolates in this taxon react with Lancefield group G or C antisera. Consequently, SDSE is commonly aggregated with other streptococcal species that also possess these Lancefield groups, known as Lancefield group C and G streptococci (GCGS) (Baracco, 2019). However, SDSE isolates which react with group A and L antisera have also been described (Brandt et al., 1999; M. D. C. B. de Pinho, 2014). SDSE is found as a commensal or a pathogen in the human throat, skin and female genitourinary tract, as well as similar sites in other animals (Agerhäll et al., 2021). SDSE isolates can be further grouped into different ecovars with distinct phenotypes associated with colonisation of different animals (Vieira et al., 1998). The type strain of this subspecies is LMG 16026 (NCFB 1356, DSM 23147) (*Taxonomy Browser, 2022, "Streptococcus Dysgalactiae Subsp. Equisimilis"*).

1.1.2. Classification of SDSE

The classification of SDSE and other GCGS species is complex and has changed considerably in recent decades, following more in-depth analysis of taxonomic relations between isolated strains. Currently, the description of *Streptococcus dysgalactiae* subsp. *equisimilis* by Vandamme and co-workers (Vandamme et al., 1996) emended by Vieira and co-workers (Vieira et al., 1998) is widely accepted.

For much of the 20th century, streptococci were classified according to the type of haemolysis reaction, Lancefield antigens, morphology and size of the colonies, and some biochemical tests. The distinction of haemolytic properties using blood agar was one of the first tests used to differentiate strains of streptococci (Sherman, 1937). Beta-haemolytic streptococci were differentiated using Rebecca Lancefield's serotyping system, which grouped strains according to specific carbohydrate antigens in the cell wall and showed the ability to distinguish previously undifferentiated strains (Lancefield, 1933). Nowadays, it is recognised that the association between species and Lancefield group is not valid for groups other than B in human isolates (Kilpper-Bälz & Schleifer, 1984; Spellberg & Brandt, 2015). Lancefield groups A, C, G and L are expressed by various streptococcal species (Table 1) (Jensen & Kilian, 2012; Spellberg & Brandt, 2015; Vieira et al., 1998).

Table 1. Streptococcus species belonging to Lancefield groups C and G (Facklam, 2002; Pinho, 2014; Turner et al. 2019).

Species	Host	Lancefield group
S. dysgalactiae subsp. *dysgalactiae*	Animals	C
S. dygalactiae subsp. *equisimilis*	Humans, Animals	A, C, G, L
S. equi subsp. *equi*	Animals	C
S. equi subsp. *zooepidemicus*	Animals, Humans	C
S. equi subsp. *ruminatorum*	Animals, Humans	C
S. canis	Animals, Humans	G
S. anginosus group	Humans, Animals	A, C, F, G
S. phocae	Animals	C

The *S. dysgalactiae* epithet has been present in the literature since the early 20th century, but it was not included in the Approved Lists of Bacterial Names in 1980. Following this, Garvie and co-workers (Garvie et al., 1983) published a description of the species based on a previous taxonomic study employing DNA-DNA hybridisation on group C alpha-haemolytic strains from bovine mastitis and provided a reference strain. Genetic analyses conducted by Farrow and Collins (Farrow & Collins, 1984) determined that strains from a variety of animal hosts identified at the time as "*S. equisimilis*" and large-colony-forming streptococci of groups G and L formed a homology group with *S. dysgalactiae* and they were thus included in the species description.

The division of *S. dysgalactiae* into two subspecies was originally proposed by Vandamme and co-workers (Vandamme et al., 1996), in which SDSE was comprised of beta-haemolytic *S. dysgalactiae* strains isolated from humans. The other subspecies, *S. dysgalactiae* subsp. *dysgalactiae* (SDSD), differed from SDSE by including only alpha-haemolytic animal strains. Later, Vieira and co-workers (Vieira et al., 1998) broadened this SDSE description to include strains with Lancefield group L and strains isolated from a variety of animal hosts. Although it has been challenged occasionally, this description is currently widely accepted (Jensen & Kilian, 2012; Kloos et al., 2001).

1.2. Identification of SDSE

1.2.1. Observation of Colonies

SDSE colonies are greyish, display a wide zone of beta-haemolysis when grown on blood agar, and have a diameter greater than 0,5 mm after 24 h of growth. Beta-haemolysis differentiates SDSE from SDSD, which presents alpha-haemolysis or no haemolysis. Colony size differentiates SDSE from other beta-haemolytic streptococci in the *S. anginosus* group (*S. anginosus*, *S. constellatus*, and *S. intermedius*), which produce small colonies after 24 hours growth in solid media (Nybakken et al., 2021; Spellberg & Brandt, 2015).

1.2.2. Biochemical Tests

Biochemical testing systems are routinely employed in the identification of clinically relevant bacteria according to their enzymatic profile, sugar fermentation pattern, and other biochemical phenotypes. Examples of these include API-20 Strep testing (SYSMEX bioMérieux), manual biochemical gallery systems, and automated systems like VITEK® 2 (bioMérieux) and Phoenix (BD Diagnostics). Often, these systems cannot distinguish SDSE from SDSD, but observation of colonies on blood agar plates is typically sufficient to distinguish these subspecies because they possess different haemolysis types (Fujita et al., 2017; Park et al., 2016). In case of dubious identification, additional tests can be done to differentiate between SDSE and other human beta-haemolytic streptococci (Table 2).

Table 2. Identification of human beta-haemolytic streptococci of Lancefield groups A, C and G (Facklam, 2002).

Species	Lancefield group	Bac	PYR	Cam	VP	Esc	Str	Sbl	Tre	Rib	αGal	βGal	βGlu
S. pyogenes	A	+	+	-	-	v	-	-	NA	-	NA	NA	NA
S. dysgalactiae subsp. *dysgalactiae*	C	-	-	-	-	v	-	v	+	+	NA	NA	NA
S. dysgalactiae subsp. *equisimilis*	A, C, G, L	-	-	-	-	+	-	-	+	+	-	-	+
S. equi	C	-	-	-	-	v	+	v	v	NA	NA	NA	NA
S. canis	G	-	-	+	-	+	-	-	v	NA	+	+	-
S. anginosus group	A, C, F, G	-	-	-	+	+	-	-	+	NA	NA	NA	NA

Abbreviations: Bac, bacitracin susceptibility; PYR, pyrrolidonylarylamidase; Cam, CAMP reaction; VP, Voges-Proskauer reaction; Esc, hydrolysis of esculin; Str, hydrolysis of starch; Sbl, Tre, and Rib, production of acid in sorbitol, trehalose, and ribose broth, respectively; αGal, α-galactosidase test; βGal, β-galactosidase test, βGlu, β-glucuronidase. +, positive reaction > 95% of strains; -, negative reaction >95% of strains; v, variable: 6 to 94% positive; NA, not applicable. *S. dysgalactiae* subsp. *dysgalactiae* is not beta-haemolytic but it is included due to taxonomic reasons.

1.2.3. Lancefield Antigen Immunoassays

Lancefield antigen grouping follows a simple and quick protocol. Streptococcal colonies can be assigned to a Lancefield group by commercially available latex agglutination techniques. SDSE isolates are mostly in groups C and G, although some may belong to groups A and L, but this alone is of little help considering there are several species that can be isolated from humans and belong to the aforementioned groups (Tables 1 and 2).

1.2.4. Mass Spectrometry

Matrix-assisted laser desorption/ionisation - time of flight mass spectrometry (MALDI-TOF MS) is an easy and cost-effective method for primary identification of microorganisms routinely employed in clinical settings. It is based on the mass distribution of proteins in bacterial cultures, generating fingerprint signatures that can be compared to reference spectra in a database (Nybakken et al., 2021; Spellberg & Brandt, 2015). Until recently, identification of SDSE by MALDI-TOF MS was limited, as most strains could not be identified with this system and some strains present similar spectra to SP or *S. canis*. Furthermore, differentiation between SDSE and SDSD was not accomplished (Tsuyuki et al., 2017).

A recent study on the identification of SDSE by MALDI-TOF showed that only 62% of SDSE isolates were identified with high confidence using the database provided by the manufacturer. By eliminating the spectra of some *S. canis* and SP reference strains in the spectrum database that interfere with SDSE identifications, correct identification could be increased from 62% to 94% (Nybakken et al., 2021).

In case of a low confidence identification or an identification category including more than one closely related species, combination of this method with Lancefield grouping and biochemical tests (Table 2) is sufficient for a correct identification (Nybakken et al., 2021; Spellberg & Brandt, 2015).

1.2.5. Molecular Methods

Amplification of specific sequences by polymerase chain reaction (PCR) has been used to identify SDSE, as it is rapid and cost-effective. Target sequences include the chaperonin *cpn60* gene, 16S rRNA gene, 23S rRNA gene, and a fragment of the streptokinase precursor gene. This method has the limitation of being species specific, being used only when this particular species is suspected (Kawata et al., 2004; Preziuso et al., 2010).

Analysis of streptococcal 16S rRNA sequences found this gene to have sufficient diversity between species and to be sufficiently conserved in each species, even if distinct clusters were present, for identification purposes. Unique highly conserved signature sequences are present in *S. mutans, S. dysgalactiae, S. equi, S. sobrinus, S. thermophilus, S. gallolyticus, S. agalactiae, S. pyogenes, S. bovis-equinus, S. anginosus* and *S. pneumoniae*. Multilocus sequence analysis of several housekeeping genes by amplification and sequencing can also be employed for the consistent and correct identification of a broad range of bacterial species (Lal et al., 2011).

1.3. Typing

1.3.1. Phenotypic Methods

Several phenotypic typing methods have been used for typing SDSE isolates. Among them are the aforementioned biochemical tests (Table 2), Lancefield antigen grouping, and other serotyping techniques initially developed for SP typing, such as M serotyping or T typing. M serotyping was initially developed for the detection of antigenic differences in the M protein of SP, but after the discovery of serological cross-reactions in GCGS with typing sera for group A streptococci, it was also used for M protein typing in these organisms (Fischetti, 1989). T typing was based on T protein serotypes. Most SDSE strains are T-typeable and at least seven distinct T serotypes have been found (Efstratiou, 1983).

1.3.2. *emm* Typing

The gene *emm* (Figure 2) 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. New *emm* types are assigned to sequences with <92% identity to reference *emm* types in the first 30 codons of the processed M protein (Gherardi et al., 2013; CDC, 2022, *Streptococcus Laboratory: M Protein Gene (Emm) Typing* Section).

Extensively used to characterize SP isolates, *emm* typing has become the most commonly used typing method to characterise SDSE isolates. Typing of this gene also provides a basis for the understanding of SDSE epidemiology and disease association. Predominant *emm* types vary according to geographic area and specific *emm* types may be associated with

different infections (M. D. 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).

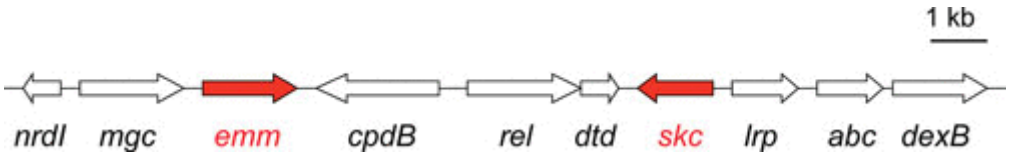

Figure 2. Representation of the streptokinase and M protein region of SDSE H46A (Malke, 2019). Legend: *nrdI*, ribonucleotide reductase; *mgc*, multigene regulator of GCS; *emm*, M protein; *cpdB*, 2′,3′-cyclic nucleotide 2′-phosphodiesterase, *rel*, bifunctional (p)ppGppase and (p)ppGpp synthetase; *dtd*, d-tyrosyl-tRNATyr deacylase; *skc*, streptokinase; *lrp*, leucine-rich protein, now recognized as a transcriptional regulator of the *PucR* family featuring a C-terminal helix-turn-helix domain; *abc*, ATP binding cassette transporter; *dexB*, α-glucosidase.

1.3.3. Multilocus Sequence Typing

Multilocus sequence typing (MLST) is a universal standardised method for characterising bacterial isolates which uses the nucleotide sequences of housekeeping genes (usually seven). 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 (a single nucleotide variation is sufficient for a sequence to be considered a new allele), and allele combinations are designated sequence types (ST). The larger the number of differences in the allelic profiles of two isolates, the farthest they are from each other, genetically. 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).

Relationships between species or individuals have been traditionally represented by dendrograms based on a matrix of pair-wise differences in the allelic profiles of the isolates. However, due to the subjective topology of this type of representation, information on the patterns of evolutionary descent cannot be obtained from them. Analyses of bacterial populations by MLST show that, in one population, the majority of isolates form a cluster of closely related genotypes, referred to as a clonal complex (CC). Clonal complexes can be defined as the group containing the predominant genotype and its single *locus* variants (SLV), double *locus* variants (DLV), and so forth, but the former is the most common definition (Feil et al., 2004).

The aforementioned topology issue is solved by using a BURST algorithm. Java implementations of this algorithm include eBURST and goeBURST. This approach is based on the biological concept of clonal complexes. It divides MLST sets into nonoverlapping groups of CCs and then subgroups 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 CCs and that the same CC 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).

1.4. Whole Genome Sequencing

The complete genome of SDSE was first sequenced in 2011. Since then, whole genome sequencing (WGS) has been increasingly used to characterise SDSE strains. The SDSE genome consists of one circular chromosome with typically 2.1 Mbp and a G+C content of 39%. Approximately 90% of the genome is shared between SDSE strains and 72% is homologous with SP (Watanabe et al., 2013). Typing may be done by amplifying specific regions of the genome with amplification and sequencing of those regions, but with WGS becoming a more common practise due to the increasing availability of next generation sequencing, MLST, core genome 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).

1.4.1. Core Genome Multilocus Sequence Typing

Core genome MLST applies the same principle as MLST to a much larger set of genes. 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).

1.4.2. The Pan-Genome

The concept of the pan-genome was introduced by Medini and colleagues in 2005 and it was defined as the complete set of genes present in a group of strains of interest. This complete set 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. The exclusive accessory genome may contain genes that confer selective advantages to that group, such as adaptation to different niches, antibiotic resistance or colonisation. Pan-genomes can be defined as open or closed and specialised species tend to have closed pan-genomes.

1.4.3. Genome Wide Association Studies

WGS produces large amounts of genomic data, much of which is poorly understood. Bacterial 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).

1.5. Colonisation and Infection

S. pyogenes, the most closely related species to *S. dysgalactiae*, is estimated to cause more than 500 000 deaths per year by invasive infections or related diseases. SDSE has for a long time been recognised as an animal pathogen. Currently, it is known to also colonise humans and a diverse range of animal hosts, including wild and domesticated animals. It causes infections in animals and humans alike, with a pathogenicity profile similar to SP in humans (Jensen & Kilian, 2012; Traverso et al., 2016). Sites of colonisation and active infections serve as the principal reservoirs for the transmission of SDSE, which generally occurs from person to person via respiratory droplets and skin contact (Sunagawa et al., 2022; C. E. Turner et al., 2019).

1.5.1. 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% among

school children with an active infection and asymptomatic carriers (Bramhachari et al., 2010). Jaalama and colleagues reported a rate of vaginal colonisation of 2.9% in pregnant women in Finland in a study related to postpartum infections by SDSE (Jaalama et al., 2018).

1.5.2. Infection

Although it is a part of the normal human microbiota, reports on the pathogenicity of SDSE are increasing. Pharyngitis is a common presentation in adolescents and adults, but invasive infections have been increasingly incident in elderly people and people with underlying conditions such as diabetes mellitus, malignancy, cardiovascular diseases, immunosuppression, substance abuse, and chronic wounds. 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. A study in Finland observed that common *emm* types often caused skin and soft tissue infections and rarer *emm* types were associated with severe infection, namely bacteraemia, and with higher mortality (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).

Tonsillopharyngitis

Many studies on tonsillopharyngitis caused by streptococci of Lancefield groups other than A do not identify the streptococci to the species level, identifying them only as group C *Streptococcus* (GCS), Group G *Streptococcus* (GGS) or GCGS. SDSE is the most probable aetiological agent in these infections, as it is the most frequently isolated GCGS species in human infections. *S. equi* and *S. anginosus* are also known to cause these infections (Baracco, 2019; M. D. Pinho, 2014).

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 a study comparing exudative pharyngitis and rhinovirus infection, SDSE was significantly more frequently isolated from exudative pharyngitis than from the control groups, which supports its role in these infections (J. C. Turner et al., 1997). 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).

In pharyngitis caused by SP, poststreptococcal sequelae (PSS), such as glomerulonephritis and rheumatic fever, may develop weeks after the initial infection in some patients. It has been suggested that SDSE may also cause PSS based on reports of poststreptococcal glomerulonephritis and acute rheumatic fever following isolation of GCGS from the RT, however, these studies do not identify the aetiological agents to the species level. Although SDSE is the most common GCGS species in human infections, studies on PSS caused by GCGS which provide species level identification showed a high incidence of acute glomerulonephritis after pharyngitis due to *S. equi* subsp. *zooepidemicus.* Such a link has not been confirmed for SDSE (Balter et al., 2000; Bramhachari et al., 2010; Brandt & Spellerberg, 2009; Duca et al., 1969; Pinto et al., 2001).

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, 2009). 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; Bruun et al., 2013). Necrotising fasciitis, Fournier's gangrene, and necrotising myositis are among the more severe SST infections caused by SDSE (Anantha et al., 2013; Bruun et al., 2013). Bruun and co-workers found an in-hospital case fatality rate of 33% for SDSE necrotising soft tissue infections in Norway. Age greater than 65, heart disease, and previous presentation of streptococcal toxic shock syndrome were found to be predictors of mortality. Other factors, such as injectable-drug use and burns, increase the risk of skin abscesses and skin infections (Bruun et al., 2013).

Invasive Infections

Since the early 2000s, there has been a significant increase in reports of invasive and non-invasive infections caused by SDSE. 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 bacteraemia cases and the portal of entry is the skin in most of them. Hospital acquired bacteraemia often occurs after surgery or other transcutaneous procedures (Baracco, 2019). The relapse rate for SDSE bacteraemia is also high. In Israel, a study found recurrence of bacteraemia caused by group G SDSE in 6 out of 84 patients, two of which with the same *emm* type recovered as in the previous infection (*stG840*). This may suggest that infection by these organisms does not confer protective immunity (Cohen-

Poradosu et al., 2004). Meningitis, peritoneal abscesses, pericarditis and pneumonia may develop after SDSE bacteraemia in some patients (Baracco, 2019).

Endocarditis by SDSE has rapid and severe clinical manifestations. A study by Oppegaard and co-workers in Norway found endocarditis constituted 4% of all invasive infections due to group G SDSE between 1999 and 2013. Patients with these infections were older (median age of 63 years) than patients with SP endocarditis (median age of 51 years), predominantly male and in general presented underlying conditions such as malignancy and diabetes mellitus. The median time of onset of disease to admission was one day and the mortality at 30 days was 22% (Oppegaard et al., 2017).

Osteoarticular infections (OAI), such as infective arthritis and osteomyelitis, are most commonly caused by staphylococci, followed immediately by streptococci. OAIs due to SDSE are not common. Seng and co-workers found that 12% of streptococcal OAIs were caused by SDSE in a group of French hospitals. However, they are associated with unfavourable clinical outcomes when compared to OAIs caused by other streptococci such as *S. agalactiae* (most commonly isolated in OAIs), *S. anginosus, S. constellatus,* and *S. pneumoniae* (Baracco, 2019; Oppegaard et al., 2016, 2018; Seng et al., 2016).

1.6. Pathogenesis and Virulence

Since the first complete genome of SDSE was sequenced in 2011, knowledge on its virulence factors has increased considerably. Around 72% of its 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). As such, much of what is currently known about SDSE virulence factors derives from previous research in SP. Important virulence factors in SP that are not present in SDSE include the hyaluronate synthase operon (*hasA* and *hasB,* which code for capsule production in the former), the streptococcal cysteine protease SpeB or streptococcal superantigen genes (streptococcal pyrogenic exotoxin, *spe*) with the exception of *speG* (Ishihara et al., 2020; Matsue et al., 2020).

1.6.1. Adherence

Microorganisms involved in infection have evolved specific mechanisms which allow them to attach and adhere to the extracellular components of the host‘s cells, as adhesion is critical in the development of infection. SDSE uses fibronectin, a high molecular weight glycoprotein, as one of the main targets for attachment. This molecule has multifunction roles in eukaryotic cells, interacting specifically with other extracellular matrix components. It is present in most tissues, including plasma and other body fluids. In the human RT, it is present in

its soluble form, which covers the epithelial cells. By binding to fibronectin binding proteins (FBPs) in the bacterial surface, fibronectin enables the bacterium to subsequently attach and colonise the site of infection (Schwarz-Linek et al., 2004). Two FBPs have been identified in SDSE: FnbA and FnbB, which bind fibronectin by the C-terminal repeat regions. GfbA, another FBP, binds to fibronectin by an upstream region containing the amino acid motif LAGESGET in its secondary binding domain (Kline et al., 1996; Lindgren et al., 1993; C. E. Turner et al., 2019).

Some FBPs and collagen binding proteins are genomically located in fibronectin and collagen binding protein and T antigen encoding *loci* (FCT), of which nine have been identified in SP. In SDSE, two FCT regions sharing high identity with SP FCT regions have been identified, suggesting that HGT occurs between the two species (Oppegaard, Mylvaganam, Skrede, Jordal, et al., 2017).

Other host cell adherence strategies in SDSE include binding to additional extracellular matrix molecules, such as fibrinogen, vitronectin, collagen and plasminogen. The M protein and 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 forms irregularities that are essential for fibrinogen-binding properties (Brandt & Spellerberg, 2009; Johansson et al., 2004).

1.6.2. Antiphagocytic Factors

The ability to resist phagocytosis is a common characteristic among 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. The M protein expressed by SDSE is critical for cell survival in human blood due to its antiphagocytic action (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, 2009b; Johansson et al., 2004).

1.6.3. Toxins and Enzymes

The dissemination of streptococci in the host tissues is a critical stage in invasive infection. Streptokinase plays a major role in the process. The conversion of plasminogen into plasmin is catalysed by the formation of a streptokinase/plasminogen complex that exposes the plasminogen active site. Plasmin lyses tissue barriers, which in turn allows for the dissemination of the bacterial cells. Streptokinase also interacts with M-proteins by binding of fibrinogen or plasminogen. The streptokinase gene, *ska*, is highly variable and alleles of this gene are associated with different tissue tropisms. Streptokinases isolated from strains colonising different hosts show specificity for the plasminogen in the hosts. Invasive SDSE

strains have been found to express higher levels of streptokinase (Brandt & Spellerberg, 2009; Malke, 2019; C. E. Turner et al., 2019).

Streptolysin O (SLO), a thiol-activated cytolysin, disrupts the membrane of several cell types, including erythrocytes, leukocytes, macrophages, platelets, and epithelial cells. In addition, it can translocate NADase into the host cells, depleting them of energy and leading toward cytotoxicity. These two molecules also limit neutrophil response inside several types of cells, protecting the bacteria from immune response (Matsue et al., 2020; O'Seaghdha & Wessels, 2013; C. E. Turner et al., 2019). Streptolysin S (SLS), a small cytolysin expressed by several species of streptococci, belongs to a group of haemolytic toxins. SLS damages the extracellular and organelle membranes in a variety of cell types. In SDSE, its expression is mediated by the *covRS* and *fasCAX* two component regulatory systems, which has been shown to regulate several virulence factors in SP. It is thought to be a relevant factor in necrotising fasciitis (Matsue et al., 2020; C. E. Turner et al., 2019).

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 roles of DNases has not been characterised and described in detail (Korczynska et al., 2012; Sumby et al., 2005).

Streptococcal pyrogenic toxins (SPEs) function as superantigens, which can lead to the exacerbated proliferation of T cells and release of inflammatory cytokines. *speG* is the only SPE expressed by SDSE. In SP, it is associated with scarlet fever and disease severity, but the presence of this gene in SDSE strains is not associated with disease severity and it does not confer mitogenic activity towards mononuclear cells, so its role in disease and infection remains unclear (Malke, 2019; Sachse et al., 2002; C. E. Turner et al., 2019).

1.7. 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 beta-lactam antibiotics (Baracco, 2019). However, there has been an increase in minimum inhibitory concentration (MIC) to penicillin in Europe and North America (Biedenbach et al., 2006). This has prompted the addition of an aminoglycoside to beta-lactam therapies in serious infections to prevent a delayed response to treatment. Clindamycin is also administered alongside beta-lactams in severe infections, especially those with a higher risk of STSS, to inhibit toxin production (Brandt & Spellerberg, 2009). Streptococci are not known to be capable of acquiring exogenous beta-lactam resistance genes and thus depend on progressive mutation of PBPs for decreased susceptibility or resistance to these antimicrobial class (Haenni et al., 2018).

A single study has reported the occurrence of a SDSE penicillin resistant clone, which was isolated in Denmark in four separate occasions between 2010 and 2012. It had mutations in multiple penicillin-binding proteins (PBPs), including some mutations similar to those in *S. pneumoniae* and *S. agalactiae* (Fuursted et al., 2016). Since the publication of this report in 2016, no further cases of penicillin resistance have been documented. Glycopeptides, daptomycin, and linezolid also show consistent activity in vitro against SDSE and may be treatment options in SDSE invasive infections.

High resistance rates to tetracyclines, clindamycin, macrolides, and fluoroquinolones make susceptibility testing a requirement 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).

1.8. Aims of the Current Work

Although widely regarded as non-pathogenic in the past, increasing awareness to SDSE in recent years is leading to a better understanding of its epidemiological dynamics, taxonomy, pathogenicity and tissue tropism (Baracco, 2019; Brandt & Spellerberg, 2009).

The discovery of significant associations between *emm* types *stC36* and *stC839* and RT infections by the strains that possess these *emm* types motivated further research on this topic. This association was observed in all SDSE isolates from the years 2011-2018 (unpublished results), suggesting that strains with these *emm* types exhibit tropism to the respiratory tract or are unable to cause invasive infection, but whether this is due to differences in the expression or presence of certain virulence factors, or directly related to the M protein is not known. The use of high throughput sequencing technology to characterise SDSE isolates from RT infection will allow a more insightful analysis of these strains.

This work aims to:

- Genomically characterise SDSE isolates recovered from respiratory tract infections in Portugal;
- Identify the main genetic lineages present in RT infections;
- Identify differences in known virulence factors among the detected clonal lineages, particularly virulence factors which could contribute to differences in virulence or tissue tropism;
- Compare RT infection isolates with isolates recovered from invasive infection in the same period;
- Explore and identify possible unknown virulence factors responsible for differences in virulence among the detected clonal lineages;
- Determine the antimicrobial susceptibility of SDSE involved in RT infections.

Chapter 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 (Table 3). These isolates were recovered in Portuguese hospitals and identified in hospital laboratories (supplemental Table 1) that were asked to submit all non-duplicate SDSE isolates from human infections to Instituto de Microbiologia da Faculdade de Medicina de Lisboa.

Table 3. Isolates recovered from human infections during the years 2011 to 2019 in Portugal.

Year	RT	Invasive	SSTI	Other	Total
2011	20	28	110	8	166
2012	17	35	103	1	156
2013	10	38	114	8	170
2014	29	57	125	5	216
2015	22	55	126	4	207
2016	20	45	91	2	158
2017	25	64	100	8	197
2018	34	62	99	6	201
2019	41	65	101	12	219
Total	218	449	969	54	1690

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 °C 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 (Probiológica, Lisbon, Portugal). 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). Colonies were suspended in 0.85% NaCl to a turbidity of approximately 0.5 McFarland and inoculated on the agar plates.

Macrolide-lincosamide-streptogramin B (MLSB) phenotype was determined by placing the clindamycin and erythromycin disks in proximity also according to CLSI guidelines (CLSI, 2020). Inhibition halos were measured and interpreted according to CLSI criteria for beta-haemolytic streptococci. Given that CLSI does not provide interpretive criteria for gentamycin and streptomycin testing of these streptococci, the criteria for high-level aminoglycoside resistance determination in enterococci were used instead (CLSI, 2020). *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. A volume of 1 ml of bacterial suspension in Todd-Hewitt broth was centrifuged at 15 000 rpm for 5 minutes and the supernatant was removed. Following this, 75 U of mutanolysin (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).

High throughput libraries were prepared with the Nextera XT DNA Library Prep paired-end kit and the Index v2 kit (Illumina, San Diego, California, 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 draft genomes were assembled with SPAdes (Bankevich et al., 2012) and polished with Pilon (Walker et al., 2014). 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. Prokka uses external feature prediction tools, such as Prodigal (Hyatt et al., 2010) and RNAmmer (Lagesen et al., 2007) to identify the coordinates of genomic features in assembled genomic DNA sequences, taken in FASTA format. 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). This software builds large-scale pangenomes, identifying the core and accessory genes. This tool uses annotated assembly files in GFF3 format as input. The coding regions are then extracted and converted into protein sequences and an all-against-all comparison using BLASTP is done. Subsequently, the sequences are clustered into families with MCL (using the

Markov cluster algorithm) (Enright et al., 2002). Homologous groups containing paralogous genes are separated into groups of true orthologous genes using conserved gene neighbourhood information. 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. The core genome was defined as the set of the core and soft-core genes and the accessory genome as the set of the shell and cloud genes (Page et al., 2015).

2.8. Genome-Wide Association Study

The chosen tool to perform GWAS was Scoary (Brynildsrud et al., 2016), version 1.6.16. This software was designed to be ultra-fast and score each component of a pangenome for associations with phenotypic traits, considering population stratification and with minimal assumptions about evolutionary processes.

The input is the gene presence and absence file created by Roary and a CSV file with the name of each strain and its classification for each trait/phenotype in binary categories (yes/no). Assuming the same genes could be responsible for the differences in virulence or 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.

First, Scoary collapses genotype variants. Following this, each variant receives the null hypothesis of non-association with the trait and a Fisher's exact test is done for each variant. Because of the large number of null hypotheses being tested, a Bonferroni adjustment and a Benjamini-Hochberg adjustment are also calculated for each variant. In this work, the Benjamini-Hochberg adjustment was used to determine significance.

Fisher's test assumes that traits are independently distributed among all strains. This makes the test unsuitable for inferring causal relations in real populations. as variants can be inherited by an entire group of strains and be only incidentally associated with another variant actually responsible for an observed trait. To overcome erroneous associations, Scoary calculates a phylogenetic tree from the Hamming distances in the genotype matrix and implements a pairwise comparison algorithm. The principle behind pairwise comparisons is to find the largest number of non-intersecting pairs of strains that differ both in the genotype and in the phenotype. This shifts the focus from single strains to evolutionary transitions as the unit of interest. The greatest number of differing pairs counts the lowest number of possible co-emergences of gene-trait combinations in the population. This is especially effective in reducing bias arising from clonal sampling.

The output is a CSV file with a list of significant genes for each tested trait. By default, the list is comprised only of genes with a p-value < 0.05, but for this work, cut-off value filters were removed and Scoary calculated the strength of association for each gene in the pangenome and produced a report which included the number of trait-positive (strains which belong to the CCs associated with the RT) strains in which the gene is present or absent and the number of trait-negative (strains in all other CCs) strains in which the gene is present or absent, the sensitivity and specificity, OR, and a Benjamini-Hochberg (HP) corrected p-value.

The cut-off values for genes whose presence may be relevant to RT infection were sensitivity and specificity higher than 70% and a significant HP corrected p-value and the cut-off values for genes whose absence may be relevant for RT infection were sensitivity and specificity lower than 30% and a significant HP p-value. Some genes appeared in both categories because they were erroneously separated by Roary due to allelic differences, and they were removed manually.

After this selection, sequences of each gene were searched in the UniProt BLAST search tool (https://www.uniprot.org/blast) and the annotation and identity of the gene with the best match was included in the table.

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.

Chapter 3. Results

3.1. Patient Demographics

Isolates from RT and invasive infections were recovered from patients belonging to distinct demographics (Figure 3). 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 (Table 4).

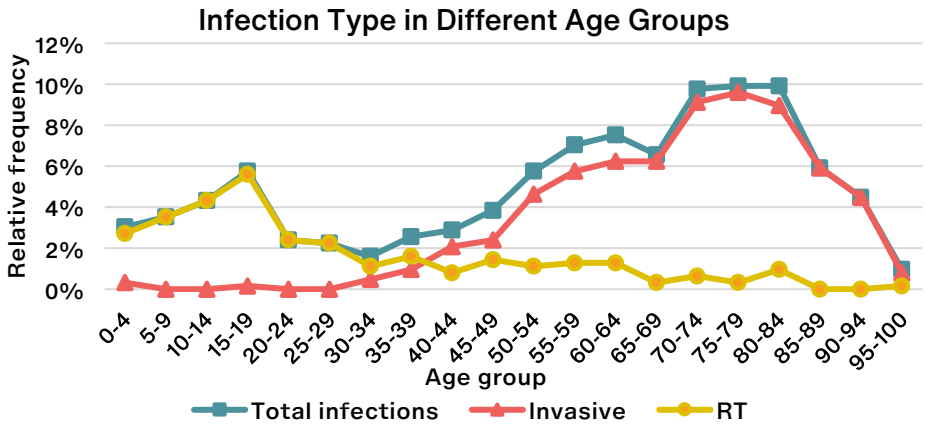

Figure 3. Relative frequency distribution of invasive and respiratory tract infections by SDSE by patient age group and gender in the years 2011 to 2019. Total infections for each age group are exclusively the sum of invasive and RT infections and not the total of infections attributed to SDSE in Portugal in that age group. One strain had no patient data and thus was excluded from demographic analyses. Frequency is calculated relative to the grand total of infections (n=625).

Table 4. Mean age of patients with invasive and RT infection.

Infection	Mean	Median
RT	27.7	19
Invasive	69.8	72

The age-adjusted distribution of SDSE invasive and RT infections (Figure 4) further emphasises the increase of invasive infection frequency with age, while the frequency of RT infections does not follow that trend.

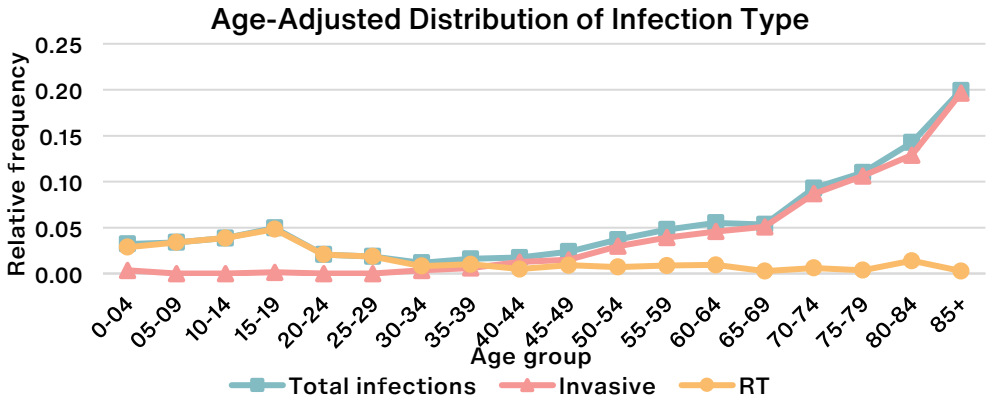

Figure 4. Age-adjusted estimated relative frequency distribution of infections by age group. Age standardisation was done with age structure data of the resident population in Portugal in the years 2011 to 2019 (https://www.pordata.pt/portugal/populacao+residente++media+anual+total+e+por+grupo+etario-10). RT infections have higher frequency in children and young adults, especially in the 15-19 age group.

3.2. Bacterial Genetic Lineages Involved in Infection

3.2.1. *emm* Typing

The association of *emm* types *stC36* and *stC839* with RT infections was observed in all SDSE isolates from the years 2011 to 2018 (Figures 5 and 6). 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. Type *stG62647* was the most frequent overall, being present in 33.3% (n=142) of invasive infections and in 19.6% (n=39) of RT infections. The second most frequent type in RT infections, *stC36*, was present in 15.6% (n=31) of these infections and it was found to be exclusively associated with them (p<0.001). Type *stC839* was present in 6.5% (n=13) of RT infections was also found to be associated with them (p<0.001).

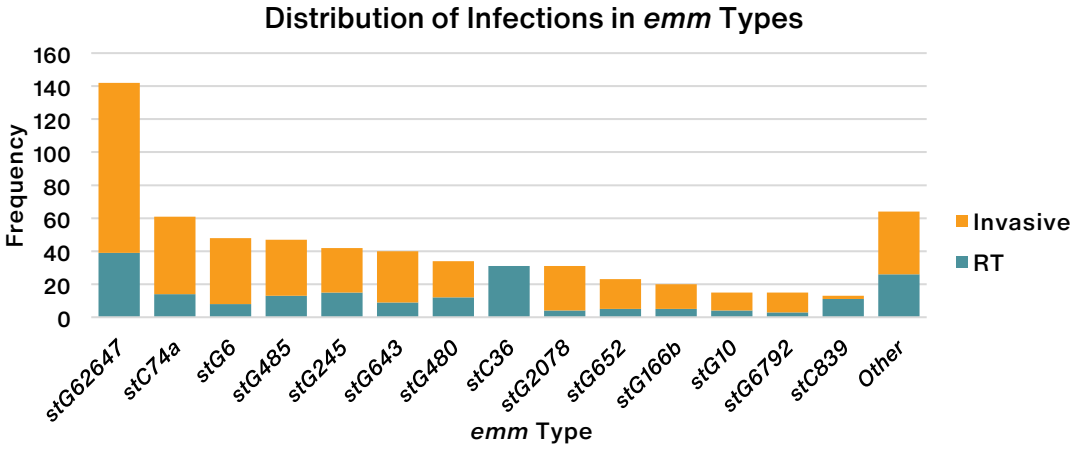

Figure 5. Distribution of RT and invasive infections for each *emm* type. Only *emm* types with frequency higher than or equal to 10 are represented.

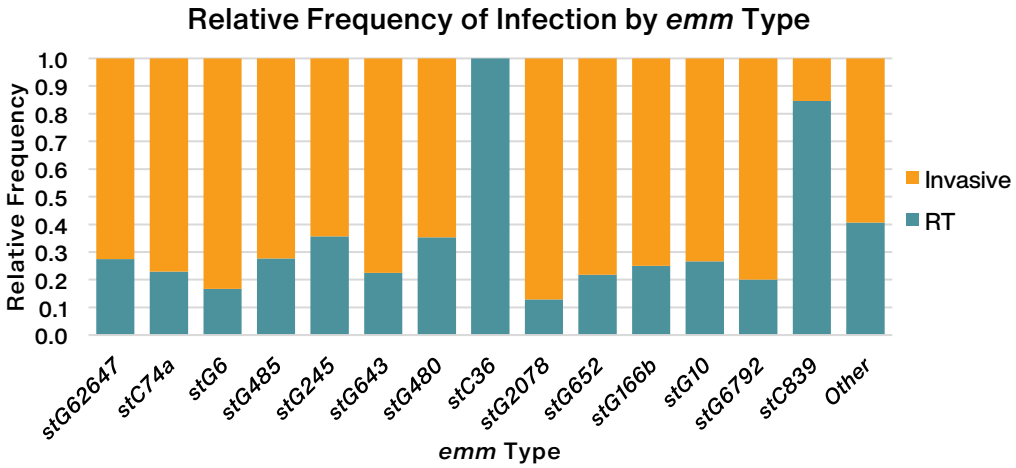

Figure 6. Distribution of RT and invasive infections for each *emm* type. Only *emm* types with 10 or more representative strains are shown. While the fraction of RT infections in most *emm* types is low, never exceeding 36%, 100% of *stC36* isolates and 84.6% of *stC839* isolates are from RT infections.

3.2.2. Multilocus Sequence Typing

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) (Figure 7). 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).

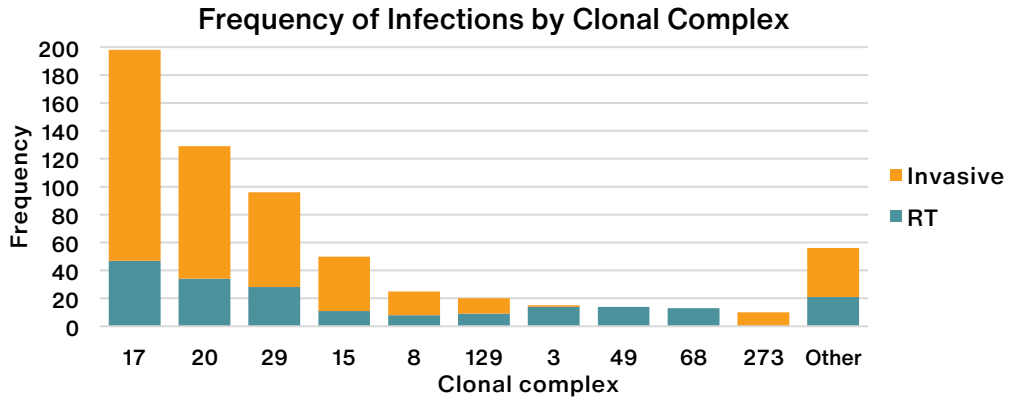

Figure 7. Frequency of RT and invasive infections for each clonal complex. Only CCs with 10 or more representative strains are shown.

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 8).

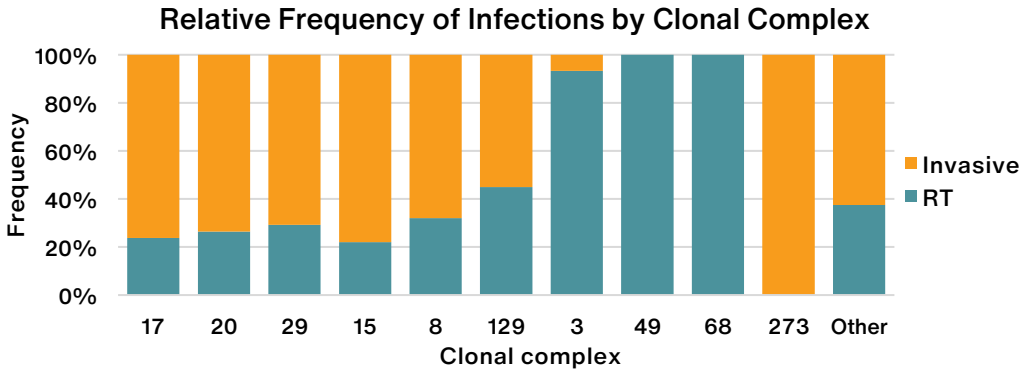

Figure 8. Distribution of RT and invasive infections for each CC. Only CCs with 10 or more representative strains are shown. While the fraction of RT infections in most CCs is low, never exceeding 36%, 100% of CC49 and CC68 isolates and 93.3% of CC3 isolates are from RT infections.

The proportion of isolates in each CC per year did not vary significantly (Figures 9 and 10) and CC17 is the most frequently isolated every year 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.

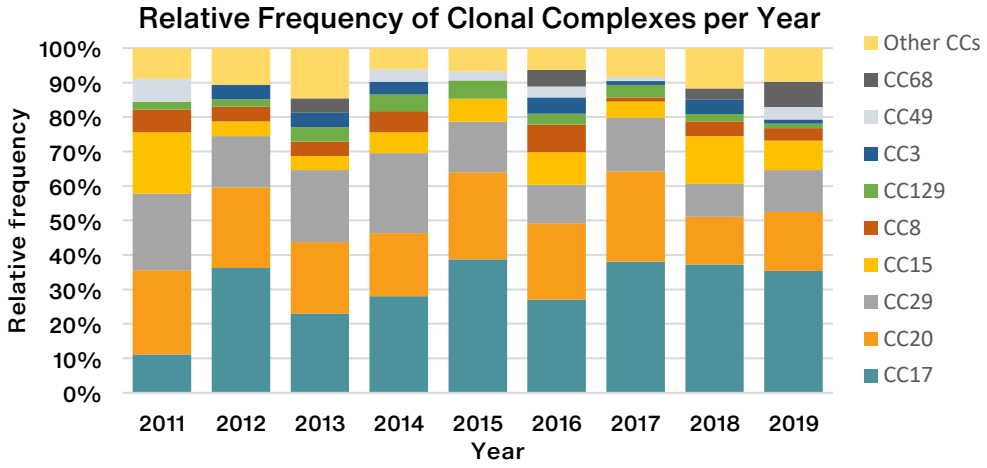

Figure 9. Relative frequency of CCs per year. Only CCs with 10 or more representative strains are shown.

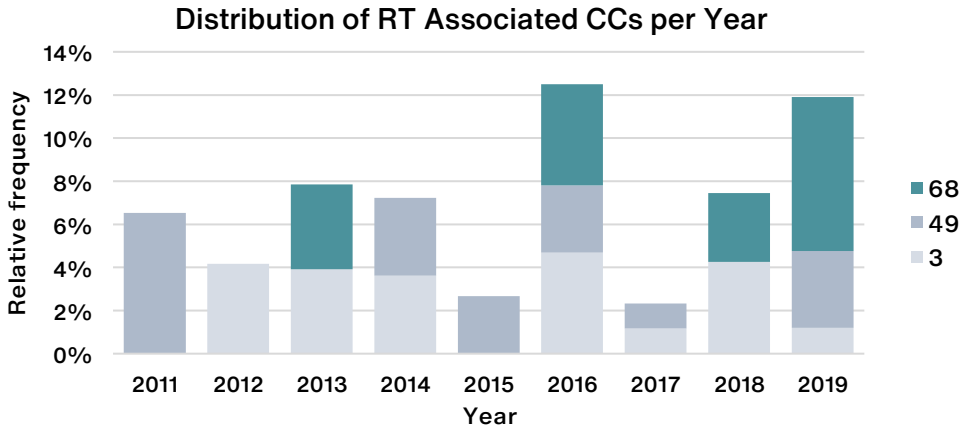

Figure 10. Relative frequency of RT-associated CCs per year.

3.2.3. Distribution of *emm* Types per Clonal Complex

While many of the most frequent CCs have a diverse variety of *emm* types, the CCs associated with RT infections are less diverse. Most strains in CC3 (n=15) are of *emm* type *stC839* (n=11) (others are *emm57* (n=3) and *stG653* (n=1)) and all strains in CC49 and CC68 possess the sequence type *stC36* (Table 5).

Table 5. Distribution of *emm* types by clonal complex. Only *emm* types and CCs with 10 or more strains are shown.

		Clonal Complex										
emm Type	*n emm*	CC17	CC20	CC29	CC15	CC8	CC129	CC3	CC49	CC68	CC273	Others
stG62647	142	18	124									
stC74a	61	41		19								1
stG6	48	3	1	35	4							5
stG485	47	15		20								12
stG245	42	27		10	3						2	
stG643	40	32			1	7						
stG480	34	21			2	9						2
stC36	31								14	13		4
stG2078	31	29										2
stG652	23			3	12	1					2	5
stG166b	20	4		4	6						4	2
stG10	15	1			14							
stG6792	15						13				1	1
stC839	13					1		11				1
Others	64	7	4	5	8	7	7	4	0	0	1	21
Total	626	198	129	96	50	25	20	15	14	13	10	56

3.2.4. Distribution of Lancefield Groups per Clonal Complex

Most CCs are exclusively or almost exclusively associated with only one Lancefield group (Table 6) (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 (Table 7).

Table 6. Distribution of Lancefield groups by clonal complex. Only CCs with 10 or more representative strains are shown.

Group		C		G		L	
CC	*n* CC	*n*	%	*n*	%	*n*	%
CC17	198	0	0.0	198	100	0	0.0
CC20	129	128	99.2	1	0.8	0	0.0
CC29	96	0	0.0	96	100	0	0.0
CC15	50	0	0.0	50	100	0	0.0
CC8	25	1	4.0	24	96.0	0	0.0
CC129	20	1	5.0	19	95.0	0	0.0
CC3	15	14	93.3	1	6.7	0	0.0
CC49	14	13	92.9	1	7.1	0	0.0
CC68	13	13	100	0	0.0	0	0.0
CC273	10	0	0.0	10	100	0	0.0
Others	56	28	50.0	27	48.2	1	1.8
Total	626	198	31.6	427	68.2	1	0.2

Table 7. Distribution of Lancefield groups in RT and invasive infections. No significant association was found between Lancefield groups and RT infections.

Lancefield Group		C		G		L	
Infection	*n*	*n*	%	*n*	%	*n*	%
RT	199	84	42.2	115	57.8	0	0.0
Invasive	427	114	26.7	312	73.1	1	0.2
Total	626	198	31.6	427	68.2	1	0.2

3.3. M Protein Analysis

A neighbour-joining tree of the M protein genes of different *emm* types (Figure 11) 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 (Table 8). 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 different patterns of hydrophobicity, G+C content, or isoelectric point 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, as far as contig length allowed sequence visualisation in that entire range.

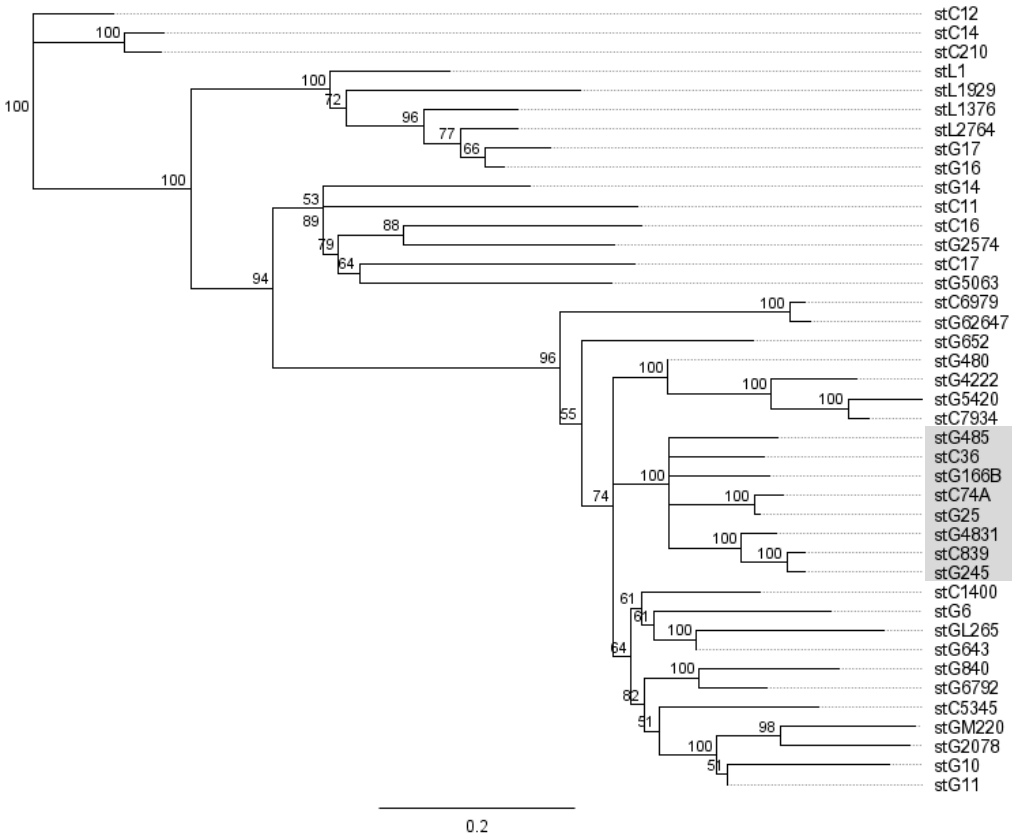

Figure 11. Neighbour-joining tree of M protein sequences of different *emm* types. The genetic distance was inferred using the Jukes-Cantor model with a bootstrap resampling method , 100 replicates, a support threshold of 50%, and sequence type *stC12* as the outgroup. M protein sequences were

translated into amino acid sequences, and a MUSCLE (multiple sequence comparison by log-expectation) alignment was performed in Geneious8. The grey triangle highlights group that includes the *emm* types that are associated with RT infection.

Table 8. Distribution of RT and invasive infections for the most frequent *emm* types.

		n		%		OR
emm type	*n*	RT	Invasive	RT	Invasive	RT
stG62647	142	39	103	27.5	72.5	0.38
stC74A	61	14	47	23.0	77.0	0.30
stG6	48	8	40	16.7	83.3	0.20
stG485	47	13	34	27.7	72.3	0.38
stG245	42	15	27	35.7	64.3	0.56
stG643	40	9	31	22.5	77.5	0.29
stG480	34	12	22	35.3	64.7	0.55
stC36	31	31	0	100.0	0.0	∞
stG2078	31	4	27	12.9	87.1	0.15
stG652	23	5	18	21.7	78.3	0.28
stG166B	20	5	15	25.0	75.0	0.33
stG10	15	4	11	26.7	73.3	0.36
stG6792	15	3	12	20.0	80.0	0.25
stC839	13	11	2	84.6	15.4	5.50

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 (Table 9). 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.

Table 9. Distribution of virulence factors in RT and invasive infections. The gene *mf3* is the only gene significantly associated with RT infections.

Gene	Total	RT	Invasive	Total	RT	Invasive	OR RT
GBS_RS03565	5	2	3	0.8	1.0	0.7	1.43
GBS_RS03570	552	163	389	88.2	81.9	91.1	0.90
GBS_RS03585	545	159	386	87.1	79.9	90.4	0.88
fbp54	626	199	427	100	100	100	1.00
hasC	626	199	427	100	100	100	1.00
lmb	624	199	425	99.7	100	99.5	1.00
mf3	58	45	13	9.3	22.6	3.0	7.43
sagA	626	199	427	100	100	100	1.00
scpA/scpB	624	199	425	99.7	100	99.5	1.00
ska	623	199	424	99.5	100	99.3	1.01
slo	624	199	425	99.7	100	99.5	1.00
speG	339	117	222	54.2	58.8	52.0	1.13
srtC1	551	163	388	88.0	81.9	90.9	0.90
srtC2	551	163	388	88.0	81.9	90.9	0.90

When the presence of VF genes was analysed for each CC, *mf3* and *speG* had higher ORs for RT-associated CCs (Table 10). 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.

Table 10. Distribution of the genes *mf3* and *speG* in the most frequent CCs.

CC		mf3			
	n	n	%	OR	p
CC17	198	0	0.0	0.000	0.000
CC20	129	0	0.0	0.000	0.000
CC29	96	2	2.1	0.197	0.003
CC15	50	5	10.0	1.087	NS
CC8	25	1	4.0	0.422	NS
CC129	20	9	45.0	5.565	0.000
CC3	15	13	86.7	11.767	0.000
CC49	14	11	78.6	10.231	0.000
CC68	13	12	92.3	12.301	0.000
CC273	10	0	0.0	0.000	NS
Other	56	5	8.9	0.960	NS
Total	626	58	9.3	-	-

3.5. Core-Genome Multilocus Sequence Typing

The relationship between strains of different CCs was analysed using cgMLST. The generated minimum spanning tree showed that the CCs associated with RT infection are genetically distant from each other (Figure 12).

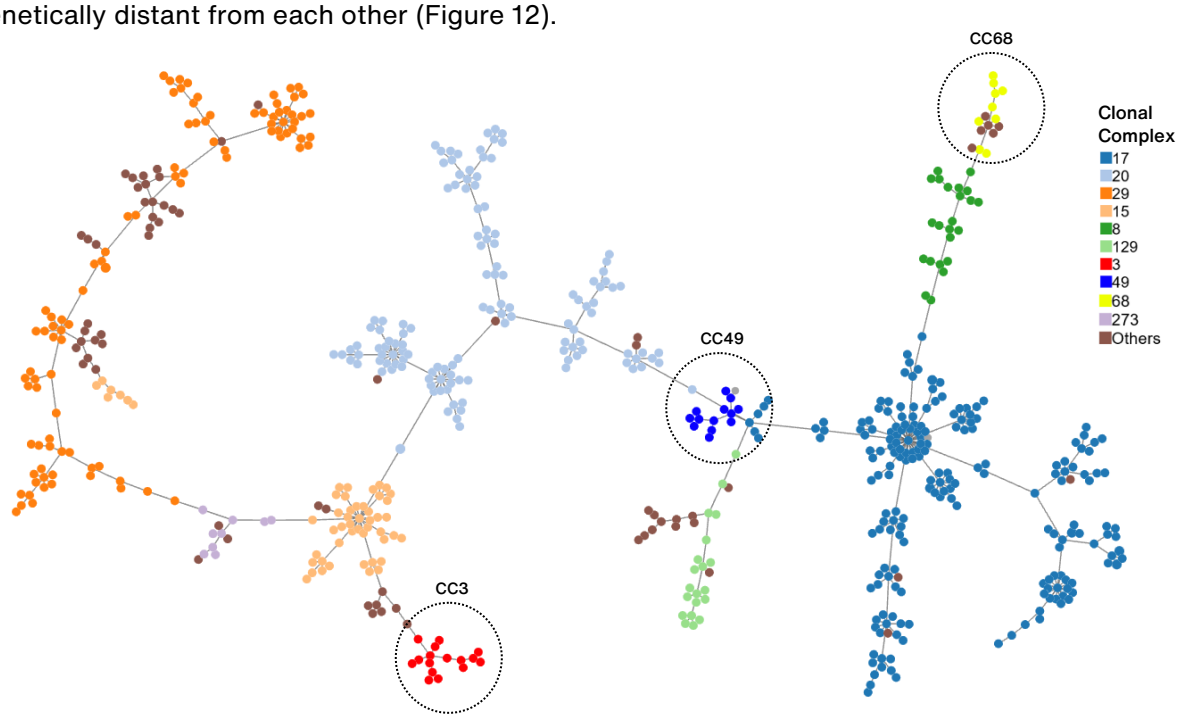

Figure 12. Minimum spanning tree based on the cgMLST scheme for the 631 isolates from RT, invasive and five adititonal isolates collected in Portugal in the years 2011 to 2019. It represents the distribution of CCs defined by MLST profiling (approximately 1600 *loci*). Isolates are represented by small circles and the distances between them are proportional to the number of allelic differences. CCs asssociated with RT infection are grouped inside the dotted circles. Within CC3, the largest distance was 318 allele differences and the distance to the closest CC was 711; for CC49, the largest distance was 147 and the distance to the closest CC was 675; and for CC68, the largest distance was 235 and the distance to the closest CC was 655. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

3.6. Pangenome Analysis and Gene Association

3.6.1. 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 (Figure 13). 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.

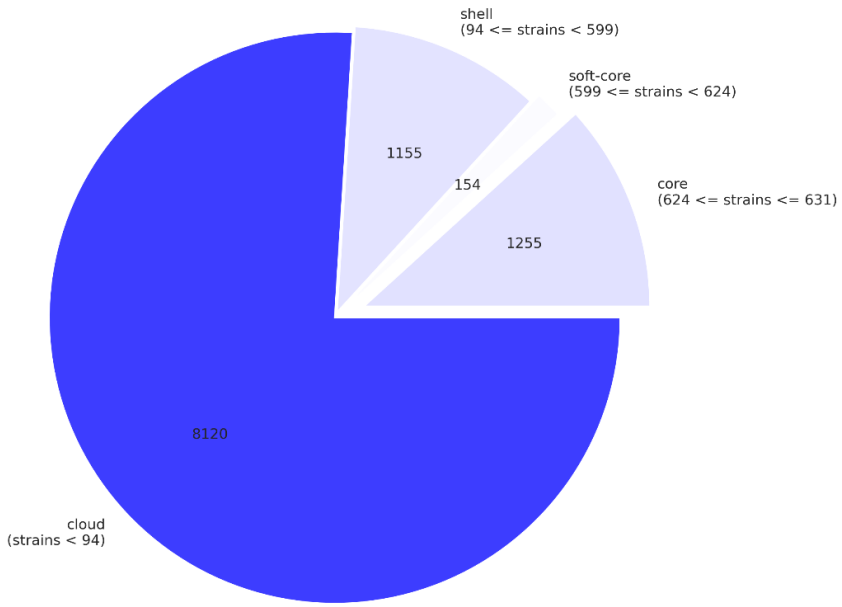

Figure 13. Pangenome breakdown for SDSE RT and invasive infection strains and 5 extra strains belonging to *emm* types *stC36* (n=2) and *stC839* (n=3) isolated from skin and soft tissue infections (n=4) or urine (one *stC36* isolate) collected in the same time range. Pie chart breakdown of the core, soft sore, shell and cloud genes and the number of isolates in which they are present.

3.6.2. 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 (Table 11) and 24 genes whose absence may be relevant for RT infection (Table 12) 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.

Table 11. Genes whose presence in RT-associated CCs is significant. Genes are sorted by decreasing specificity. Sensitivity is the sensitivity if using the presence of this gene as a diagnostic test to determine trait-positivity. Specificity is the specificity if using the non-presence of this gene as a diagnostic test to determine trait-negativity. All genes have p<0.001.

		%	*n* present		*n* not present		%		
Gene	Uniprot annotation	Identity	RT CCs	Other CCs	RT CCs	Other CCs	Sensitivity	Specificity	OR
group_5532	Extracellular nuclease 3	100	36	13	8	574	81.8	97.8	198.7
group_3382	CsbD domain-containing protein	100	36	26	8	561	81.8	95.6	97.1
group_5792	Uncharacterized protein	92.1	44	106	0	481	100	81.9	inf
group_984	Acetyltransferase	60.7	43	107	1	480	97.7	81.8	192.9
group_3532	DNA-binding helix-turn-helix protein	39.4	44	108	0	479	100	81.6	inf
group_1175	Putative transcriptional activator regulator protein	77.5	44	117	0	470	100	80.1	inf
group_2986	ABC transporter	76.5	43	149	1	438	97.7	74.6	126.4
group_4326	Uncharacterized protein	46.9	44	157	0	430	100	73.3	inf
group_1661	CAAX amino terminal protease family protein	95.8	44	158	0	429	100	73.1	inf
tagG	Transport permease protein	92.9	44	159	0	428	100	72.9	inf
group_5040	Rhamnose-glucose cell wall localization/formation ABC-transporter	85.5	44	159	0	428	100	72.9	inf
group_5042	Rhamnan synthesis protein F	71.3	44	159	0	428	100	72.9	inf
group_5041	Rhamnose-glucose cell wall localization/formation glycosyltransferase	68.7	44	159	0	428	100	72.9	inf
mgtA	Glycosyl transferase	51.1	44	159	0	428	100	72.9	inf
group_5044	DUF2304 domain-containing protein	49.1	44	159	0	428	100	72.9	inf
group_5045	Membrane protein	43.9	44	159	0	428	100	72.9	inf
group_5047	Phosphoglycerol transferase	41.3	44	159	0	428	100	72.9	inf

Legend: *n* present, number of strains in which the gene is present; *n* not present, number of strains in which the gene is not present, RT CCs, CCs associated with RT infection(CC3, CC49 and CC68).

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.

Table 12. Genes for which absence in RT CCs associated CCs is significant. All genes have p<0.001.

		%	*n* present		*n* not present		%		
Gene	Uniprot annotation	Identity	RT CCs	Other CCs	RT CCs	Other CCs	Sensitivity	Specificity	OR
group_2020	pre-crRNA processing endonuclease	100	0	560	44	27	0.0	4.6	0
cas1_1	CRISPR-associated endonuclease Cas1	98.5	0	554	44	33	0.0	5.6	0
group_2704	Uncharacterized protein	96.8	0	554	44	33	0	5.6	0
group_3222	CRISPR-associated exonuclease Cas4	95.1	0	554	44	33	0.0	5.6	0
cas2_1	CRISPR-associated endoribonuclease Cas2	96.9	0	553	44	34	0	5.8	0
group_960	Uncharacterized protein	96	0	540	44	47	0	8.0	0
group_445	HD Cas3-type domain-containing protein	95.9	0	536	44	51	0.0	8.7	0
group_3359	Uncharacterized protein	100	0	500	44	87	0	14.8	0
group_2398	Uncharacterized protein	93.5	5	495	39	92	11	15.7	2.4E-02
hpaIIM	Cytosine-specific methyltransferase	100	0	486	44	101	0	17.2	0
tipA	Putative transcriptional activator regulator	98	0	476	44	111	0	18.9	0
group_612	Sigma-70 family RNA polymerase sigma factor	95.6	0	471	44	116	0	19.8	0
group_2335	FtsK/SpoIIIE-like DNA segregation ATPase	95.9	0	464	44	123	0	21.0	0
group_2307	Enoyl-(Acyl-carrier-protein) reductase II	43	0	449	44	138	0	23.5	0
btuD_1	Bacteriocin ABC-type exporter	53.3	0	436	44	151	0	25.7	0
group_5103	Transport permease protein	91.8	0	428	44	159	0	27.1	0
group_5102	ABC transporter ATP-binding protein	89.1	0	428	44	159	0	27.1	0
group_5101	Rhamnosyltransferase	71.3	0	428	44	159	0	27.1	0
group_4132	Glycosyltransferase	67.1	0	428	44	159	0	27.1	0
group_5099	Glycosyl transferase family protein	61.4	0	428	44	159	0	27.1	0
group_5100	Membrane protein	49	0	428	44	159	0	27.1	0
group_5098	Uncharacterized conserved protein	44.4	0	428	44	159	0	27.1	0
group_4131	Glycosyltransferase family 1protein	89.5	0	427	44	160	0	27.3	0
group_1780	Sigma-70 family RNA polymerase sigma factor	97.8	1	416	43	171	2	29.1	9.6E-03

3.7. 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 (Table 13). No significant differences in antimicrobial resistance were found between RT and invasive infection isolates.

Table 13. Frequency of resistant isolates and rate of antimicrobial resistance for RT and invasive infections. The first column indicates the total number of strains for each infection type and the others indicate the number of resistant strains and rate of resistance to each antimicrobial. Legend: DA, clindamycin; E, erythromycin; TE, tetracycline; LEV, levofloxacin; CN, gentamycin; S, streptomycin; C, chloramphenicol.

		DA		E		TE		LEV		CN		S		C	
Infection	*n*	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
RT	199	18	9.0	56	28.1	32	16.1	2	1.0	0	0.0	1	0.5	0	0.0
Invasive	427	33	7.7	138	32.3	74	17.3	13	3.0	3	0.7	11	2.6	2	0.5
Total	626	51	8.1	194	31.0	106	16.9	15	2.4	3	0.5	12	1.9	2	0.3

Antimicrobial resistance rates vary considerably in different CCs, however, they do not differ significantly between the CCs associated with RT infection (CC3, CC49, and CC68) (Table 14).

Table 14. Frequency and rate of antimicrobial resistance by CC. Only CCs with 10 or more representatives are shown.

		DA		E		TE		LEV		CN		S		C	
CC	*n*	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
17	198	14	7.1	107	54.0	6	3.0	10	5.1	1	0.5	4	2.0	0	0.0
20	129	1	0.8	6	4.7	1	0.8	1	0.8	0	0.0	3	2.3	0	0.0
29	96	10	10.4	27	28.1	15	15.6	0	0.0	0	0.0	1	1.0	0	0.0
15	52	6	12.0	15	30.0	42	84.0	2	4.0	2	4.0	1	2.0	0	0.0
8	25	6	24.0	12	48.0	8	32.0	1	4.0	0	0.0	1	4.0	0	0.0
129	20	2	10.0	5	25.0	1	5.0	0	0.0	0	0.0	0	0.0	0	0.0
3	15	0	0.0	5	33.3	1	6.7	1	6.7	0	0.0	0	0.0	0	0.0
49	14	2	14.3	2	14.3	3	21.4	0	0.0	0	0.0	0	0.0	0	0.0
68	14	1	7.7	1	7.7	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
273	10	1	10.0	3	30.0	3	30.0	0	0.0	0	0.0	0	0.0	0	0.0
Other	57	8	14.0	11	19.3	26	45.6	0	0.0	0	0.0	2	3.5	2	3.5

The evolution of resistance rates to different antimicrobials throughout the years 2011 to 2019 is shown in Table 15. Erythromycin resistance increased from 19.6% in 2011 to 38% in 2019 (p=0.04). All chloramphenicol resistant strains were recovered in 2018.

Table 15. Frequency and rate of antimicrobial resistance by year.

		DA		E		TE		LEV		CN		S		C	
Year	*n*	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
2011	46	6	13.0	9	19.6	17	37.0	2	4.3	1	2.2	0	0.0	0	0.0
2012	48	1	2.1	12	25.0	6	12.5	5	10.4	0	0.0	1	2.1	0	0.0
2013	48	4	8.3	17	35.4	8	16.7	2	4.2	0	0.0	2	4.2	0	0.0
2014	83	9	10.8	25	30.1	9	10.8	1	1.2	0	0.0	4	4.8	0	0.0
2015	75	4	5.3	24	32.0	11	14.7	1	1.3	0	0.0	2	2.7	0	0.0
2016	64	6	9.4	15	23.4	12	18.8	0	0.0	0	0.0	0	0.0	0	0.0
2017	86	6	7.0	20	23.3	9	10.5	2	2.3	0	0.0	1	1.2	0	0.0
2018	92	7	7.6	40	43.5	18	19.6	1	1.1	2	2.2	1	1.1	2	2.2
2019	84	8	9.5	32	38.1	16	19.0	1	1.2	0	0.0	1	1.2	0	0.0

Chapter 4. Discussion

The rise in human infections caused by SDSE, in the past four decades, notably invasive infections, has led to increased awareness about its pathogenic potential and consequently, more research on this organism is being undertaken. This work characterised SDSE strains from human respiratory tract infections collected in Portugal in the years 2011 to 2019. A total of 199 RT infection strains were characterised and compared to invasive infection strains collected in the same period. For phenotypic characterisation, haemolysis type, Lancefield group, and antimicrobial susceptibility were determined. Genomic characterisation was done by sequencing the genomes of RT infection strains and some invasive strains for comparison purposes, which allowed the analysis of the main genetic lineages responsible for RT infection, known VFs, their pangenome, and a gene association study to explore other possible VFs. To date, no large-scale gene association studies performed on SDSE RT infection isolates were found in the literature.

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). Invasive infection was much more frequent in older patients and frequency increased with age. 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, 2009b; 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

Typing of the *emm* gene has been widely used to characterise SDSE strains and it is often the only available genotypic information in some studies (Bramhachari et al., 2010; Gherardi et al., 2014; Rantala et al., 2010).

Typing of the emm gene among RT and invasive infection isolates in the years 2011 to 2019 revealed 38 distinct *emm* types and two non-typeable strains, with 31 distinct types in RT infection isolates and 30 in invasive isolates. Type *stG62647* was the most frequent overall, being present in 33.3% (n=142) of invasive infections and in 19.6% (n=39) of RT infections. The second most frequent type in RT infections, *stC36*, was present in 15.6% (n=31) of these infections and it was exclusively associated with them. Type *stC839*, present in 6.5% (n=13) of RT infections, was also associated with them with an odds ratio (OR) of 5.5. This 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 or are unable to cause invasive infection. These differences in virulence and tissue tropism 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. MLST yielded 121 STs distributed among 37 CCs and 72 singlets, with a vast majority of isolates belonging to the few most frequent CCs. 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, with only one CC3 isolate from invasive infection. This suggests that some CCs have reduced potential to cause invasive infection.

The distribution of *emm* types varied greatly in different CCs: among the most frequent CCs, CC17 included 10 of the most frequent *emm* types, CC29 included 6, and CC15 included 7. The RT-associated CCs, on the contrary, presented far less diversity: CC49 and CC68 are exclusively associated with *stC36*, and CC3 is mainly associated with *stC839*, but also includes *emm57* and *stG653*. Each CC was associated with either Lancefield group C or G, with at least 92.9% of the strains of each CC possessing one of these groups. The majority of isolates in the RT-associated CCs possessed group C, the exceptions being one CC3 and one CC49 isolates possessing group G.

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, Mylvaganam, Skrede, Lindemann, 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 association between *emm* types *stC36* and *stC839* and RT infection may be due to differences in the M protein expressed by isolates with these *emm* types or due to differences in VF presence or expression in the clonal lineages that possess these *emm* types. No differences were found in the genomic context of the *emm* gene in RT infection isolates of different CCs or *emm* types.

Genomic sequences of *emm* genes of different *emm* types were translated into amino acid sequences, aligned with MUSCLE, and a neighbour-joining tree was constructed with these sequences. All the most frequently isolated *emm* types were used and other *emm* types were included for comparison. 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.

The genes *fbp54*, *hasC*, and *sagA* were present in all the RT and invasive infection isolates and the genes *lmb*, *scpA/scpB*, *ska*, and *slo* were present in all RT isolates. This is expected due to the near ubiquity of these genes reported in SDSE (Kittang et al., 2011; Lo & Cheng, 2015; Lother et al., 2017). The absence of the genes *hasA* and *hasB* from the *hasABC* operon was observed in all isolates and it is also expected, as SDSE does not possess a hyaluronic acid capsule (Shimomura et al., 2011; Watanabe et al., 2013).

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. Apart from CC129, in which it was present in around half of the isolates, it was only sporadically found in other CCs. This gene codes the protein Mitogen factor 3, an extracellular DNase. The genomic context of *mf3* in RT isolates is varied but it is usually surrounded by small phage genes (data not shown). 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). A study on *mf3* expression in SP with a skin infection model concluded that it contributes to improved bacterial dissemination and lesion size (Wen et al., 2011). Because it is a DNase, 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 Study

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 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. In total, the core genome included only 13.2% of all genes in the pan-genome. Such a small core/pan-genome ratio suggests that SDSE has an open genome. The core genome typically contains the genes responsible for the basic biologic aspects of a species and phenotypic traits associated with all its members. In contrast, 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). Thus, a gene presence/absence association study performed for the group of strains belonging to the RT-associated CCs may find a larger number of genes which may be responsible for differences in virulence.

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*. Alignment and a neighbour joining tree of all the *mf3* sequences revealed 3 main distinct groups in which almost all sequences were included (data not shown). 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 the RT-associated 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.

Chapter 5. Conclusions

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 RT infections in Portugal. This association motivated the genomic characterisation of 199 isolates recovered from RT infections in the years 2011 to 2019 to define the main genetic lineages responsible for RT infection in Portugal, to identify virulence factors that may be involved in respiratory tract tropism, and to determine antimicrobial resistance in these isolates, using genomic data from invasive infections for comparison purposes. To date, no large-scale gene association studies performed on SDSE RT infection isolates have been published.

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.

References

RAgerhäll, 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

Anantha, R. V., Kasper, K. J., Patterson, K. G., Zeppa, J. J., Delport, J., & McCormick, J. K. (2013). Fournier's gangrene of the penis caused by Streptococcus dysgalactiae subspecies equisimilis: Case report and incidence study in a tertiary-care hospital. *BMC Infectious Diseases*, *13*(1), 1–5. https://doi.org/10.1186/1471-2334-13-381/TABLES/2

Balter, S., Benin, A., Wyton Lima Pinto, S., Teixeira, L. M., Alvim, G. G., Luna, E., Jackson, D., Laclaire, L., Elliott, J., Facklam, R., & Schuchat, A. (2000). Epidemic nephritis in Nova Serrana, Brazil. *The Lancet*, *355*(9217), 1776–1780. https://doi.org/10.1016/S0140-6736(00)02265-0

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., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477. https://doi.org/10.1089/CMB.2012.0021

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

Biedenbach, D. J., Toleman, M. A., Walsh, T. R., & Jones, R. N. (2006). Characterization of fluoroquinolone-resistant beta-hemolytic Streptococcus spp. isolated in North America and Europe including the first report of fluoroquinolone-resistant Streptococcus dysgalactiae subspecies equisimilis: report from the SENTRY Antimicr. *Diagnostic Microbiology and Infectious Disease*, *55*(2), 119–127. https://doi.org/10.1016/J.DIAGMICROBIO.2005.12.006

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., Haase, G., Schnitzler, N., Zbinden, R., & Lütticken, R. (1999). Characterization of blood culture isolates of Streptococcus dysgalactiae subsp. equisimilis possessing Lancefield's group A antigen. *Journal of Clinical Microbiology*, *37*(12), 4194–4197. https://doi.org/10.1128/jcm.37.12.4194-4197.1999

Brandt, C. M., & Spellerberg, B. (2009a). *Human Infections Due to Streptococcus dysgalactiae*

Subspecies equisimilis. *49*. 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

Brandt, C. M., & Spellerberg, B. (2009c). 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

Bruun, T., Kittang, B. R., de Hoog, B. J., Aardal, S., Flaatten, H. K., Langeland, N., Mylvaganam, H., Vindenes, H. A., & Skrede, S. (2013). Necrotizing soft tissue infections caused by Streptococcus pyogenes and Streptococcus dysgalactiae subsp. equisimilis of groups C and G in western Norway. *Clinical Microbiology and Infection*, *19*(12), E545–E550. https://doi.org/10.1111/1469-0691.12276

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*.

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

Duca, E., Teodorovici, G., Radu, C., VÎTĂ, A., Talasman-Niculescu, P., Bernescu, E., Feldi, C., & ROŞCA, V. (1969). A new nephritogenic streptococcus. *Epidemiology & Infection*, *67*(4), 691–698. https://doi.org/10.1017/S0022172400042145

Efstratiou, A. (1983). The serotyping of hospital strains of streptococci belonging to Lancefield group C and group G. *Epidemiology & Infection*, *90*(1), 71–80. https://doi.org/10.1017/S0022172400063865

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. https://doi.org/10.1093/NAR/30.7.1575

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

Farrow, J. A. E., & Collins, M. D. (1984). Taxonomic Studies on Streptococci of Serological Groups C, G and L and Possibly Related Taxa. *Systematic and Applied Microbiology*, *5*(4), 483–493. https://doi.org/10.1016/S0723-2020(84)80005-3

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

Fujita, T., Horiuchi, A., Ogawa, M., & Yoshida, H. (2017). *Genetic Diversity in Streptococcus dysgalactiae subsp . equisimilis Isolates from Patients with Invasive and Noninvasive Infections in a Japanese University*. *2015*, 100–104. https://doi.org/10.7883/yoken.JJID.2015.602

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

Garvie, E. I., Farrow, J. A. E., & Bramley, A. J. (1983). Streptococcus dysgalactiae. *Definitions*, 404–405. https://doi.org/10.32388/ljnwnh

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

Gherardi, G., Imperi, M., Palmieri, C., Magi, G., Facinelli, B., Baldassarri, L., Pataracchia, M., & Creti, R. (2014). Genetic diversity and virulence properties of Streptococcus dysgalactiae subsp. equisimilis from different sources. *Journal of Medical Microbiology*, *63*(Pt 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

Ishihara, H., Ogura, K., Miyoshi-Akiyama, T., Nakamura, M., Kaya, H., & Okamoto, S. (2020). Prevalence and genomic characterization of Group A Streptococcus dysgalactiae subsp. equisimilis isolated from patients with invasive infections in Toyama prefecture, Japan. *Microbiology and Immunology*, *64*, 113. https://doi.org/10.1111/1348-0421.12760

Jaalama, M., Palomäki, O., Vuento, R., Jokinen, A., & Uotila, J. (2018). Prevalence and Clinical Significance of Streptococcus dysgalactiae subspecies equisimilis (Groups C or G Streptococci) Colonization in Pregnant Women: A Retrospective Cohort Study. *Infectious Diseases in Obstetrics and Gynecology*, *2018*. https://doi.org/10.1155/2018/2321046

Jensen, A., & Kilian, M. (2012). Delineation of Streptococcus dysgalactiae, its subspecies, and its clinical

and phylogenetic relationship to Streptococcus pyogenes. *Journal of Clinical Microbiology*, *50*(1), 113–126. https://doi.org/10.1128/JCM.05900-11

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

Kawata, K., Anzai, T., Senna, K., Kikuchi, N., Ezawa, A., & Takahashi, T. (2004). Simple and rapid PCR method for identification of streptococcal species relevant to animal infections based on 23S rDNA sequence. *FEMS Microbiology Letters*, *237*(1), 57–64. https://doi.org/10.1016/j.femsle.2004.06.015

Kilpper-Bälz, R., & Schleifer, K. H. (1984). *Nucleic acid hybridization and cell wall composition studies of pyogenic streptococci*. *24*, 355–364.

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

Kloos, W. E., Hardie, J. M., & Whiley, R. A. (2001). International Committee on Systematic Bacteriology Subcommittee on the taxonomy of staphylococci and streptococci. *International Journal of Systematic and Evolutionary Microbiology*, *51*(2), 717–718. https://doi.org/10.1099/00207713-51-2-717

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

Lal, D., Verma, M., & Lal, R. (2011). Exploring internal features of 16S rRNA gene for identification of clinically relevant species of the genus Streptococcus. *Annals of Clinical Microbiology and Antimicrobials*, *10*(1), 28. https://doi.org/10.1186/1476-0711-10-28

Lancefield, B. R. C. (1933). A SEROLOGICAL DIFFERENTIATION OF HUMAN AND (From the Hospital of The Rockefeller Institute for Medical Research). *The Journal of Experimental Medicine*, *1919*(1), 571–595.

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

Malke, H. (2019). Genetics and Pathogenicity Factors of Group C and G Streptococci. *Microbiology Spectrum*, *7*(2). https://doi.org/10.1128/MICROBIOLSPEC.GPP3-0002-2017/ASSET/C508E25E-7EA2-4CA9-8C30-62217665CC20/ASSETS/GRAPHIC/GPP3-0002-2018-FIG7.GIF

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

Nybakken, E. J., Oppegaard, O., Gilhuus, M., Jensen, C. S., & Mylvaganam, H. (2021). Identification of Streptococcus dysgalactiae using matrix-assisted laser desorption/ionization-time of flight mass spectrometry; refining the database for improved identification. *Diagnostic Microbiology and Infectious Disease*, *99*(1), 115207. https://doi.org/10.1016/j.diagmicrobio.2020.115207

O'Seaghdha, M., & Wessels, M. R. (2013). Streptolysin O and its Co-Toxin NAD-glycohydrolase Protect Group A Streptococcus from Xenophagic Killing. *PLOS Pathogens*, *9*(6), e1003394. https://doi.org/10.1371/JOURNAL.PPAT.1003394

Oppegaard, O., Mylvaganam, H., Skrede, S., Jordal, S., Glambek, M., & Kittang, B. R. (2017). Clinical and molecular characteristics of infective β-hemolytic streptococcal endocarditis. *Diagnostic Microbiology and Infectious Disease*, *89*(2), 135–142. https://doi.org/10.1016/J.DIAGMICROBIO.2017.06.015

Oppegaard, O., Mylvaganam, H., Skrede, S., & Kittang, B. R. (2018). Exploring the arthritogenicity of Streptococcus dysgalactiae subspecies equisimilis. *BMC Microbiology*, *18*(1), 1–10. https://doi.org/10.1186/S12866-018-1160-5/FIGURES/4

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

Oppegaard, O., Skrede, S., Mylvaganam, H., & Kittang, B. R. (2016). Temporal trends of β-haemolytic streptococcal osteoarticular infections in western Norway. *BMC Infectious Diseases*, *16*(1). https://doi.org/10.1186/S12879-016-1874-7

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

Pinto, S. W. L., Sesso, R., Vasconcelos, E., Watanabe, Y. J., & Pansute, A. M. (2001). Follow-up of patients with epidemic poststreptococcal glomerulonephritis. *American Journal of Kidney Diseases*, *38*(2), 249–255. https://doi.org/10.1053/AJKD.2001.26083

Preziuso, S., Laus, F., Tejeda, A. R., Valente, C., & Cuteri, V. (2010). Detection of Streptococcus dysgalactiae subsp. equisimilis in equine nasopharyngeal swabs by PCR. *Journal of Veterinary Science*, *11*(1), 67. https://doi.org/10.4142/JVS.2010.11.1.67

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

Sachse, S., Seidel, P., Gerlach, D., GÃ¼nther, E., RÃ¶del, J., Straube, E., & Schmidt, K.-H. (2002). Superantigen-like gene(s) in human pathogenic Streptococcus dysgalactiae, subsp equisimilis: genomic localisation of the gene encoding streptococcal pyrogenic exotoxin G (speG(dys)). *FEMS Immunology and Medical Microbiology*, *34*(2), 159–167. https://doi.org/10.1111/J.1574-695X.2002.TB00618.X

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

Schwarz-Linek, U., Höök, M., & Potts, J. R. (2004). The molecular basis of fibronectin-mediated bacterial adherence to host cells. *Molecular Microbiology*, *52*(3), 631–641. https://doi.org/10.1111/J.1365-2958.2004.04027.X

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

Seng, P., Vernier, M., Gay, A., Pinelli, P. O., Legré, R., & Stein, A. (2016). Clinical features and outcome of bone and joint infections with streptococcal involvement: 5-year experience of interregional reference centres in the south of France. *New Microbes and New Infections*, *12*, 8–17. https://doi.org/10.1016/J.NMNI.2016.03.009

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

Sherman, J. M. (1937). the Streptococci. *Bacteriological Reviews*, *1*(1), 3–97. https://doi.org/10.1128/mmbr.1.1.3-97.1937

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

Spellberg, B., & Brandt, C. M. (2015). Manual of Clinical Microbiology. In J. H. Jorgensen, M. A. Pfaller, K. C. Carroll, M. L. Landry, G. Funke, S. S. Richter, & D. W. Warnock (Eds.), *Manual of Clinical Microbiology* (11th ed., pp. 383–402). ASM Press. https://doi.org/10.1128/9781555817381.ch22

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

Sunagawa, K., Shirafuji, T., Sun, G., Arai, R., Azuma, H., Miyoshi-Akiyama, T., & Katano, H. (2022). Intra-familial transmission of Streptococcus dysgalactiae subsp. equisimilis (SDSE): A first case report and review of the literature. *Journal of Infection and Chemotherapy : Official Journal of the Japan Society of Chemotherapy*, *28*(6), 819–822. https://doi.org/10.1016/J.JIAC.2022.01.017

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

Taxonomy browser (Streptococcus dysgalactiae subsp. equisimilis). (n.d.). Retrieved January 3, 2022, from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=119602

Taxonomy browser (Streptococcus pyogenes). (n.d.). Retrieved January 6, 2022, from https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=Streptococcus+hemolyticus

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

Tsuyuki, Y., Kurita, G., Murata, Y., Goto, M., & Takahashi, T. (2017). Identification of group G streptococcal isolates from companion animals in Japan and their antimicrobial resistance patterns. *Japanese Journal of Infectious Diseases*, *70*(4), 394–398. https://doi.org/10.7883/yoken.JJID.2016.375

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

Vandamme, P., Pot, B., Falsen, E., Kersters, K., & Devriese, L. A. (1996). Taxonomic study of lancefield streptococcal groups C, G, and L (Streptococcus dysgalactiae) and proposal of S. dysgalactiae subsp. equisimilis subsp. nov. *International Journal of Systematic Bacteriology*, *46*(3), 774–781. https://doi.org/10.1099/00207713-46-3-774

Vieira, V. V., Teixeira, L. M., Zahner, V., Momen, H., Facklam, R. R., Steigerwalt, A. G., Brenner, D. J., & Castro, A. C. D. (1998). Genetic relationships among the different phenotypes of Streptococcus dysgalactiae strains. *International Journal of Systematic Bacteriology*, *48*(4), 1231–1243. https://doi.org/10.1099/00207713-48-4-1231

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

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/JOURNAL.PONE.0112963

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

Supplemental Data

Supplemental Table 1. List of hospital laboratories.

Centro Hospitalar de Cascais
Centro Hospitalar de Leiria - Hospital de Santo André
Centro Hospitalar de Lisboa Ocidental
Centro Hospitalar de Vila Nova de Gaia e Espinho
Centro Hospitalar do Baixo Vouga - Hospital Infante D. Pedro
Centro Hospitalar do Funchal - Hospital Dr Nélio Mendonça
Centro Hospitalar do Médio Ave
Centro Hospitalar e Universitário de Coimbra
Centro Hospitalar Entre Douro e Vouga - Hospital São Sebastião
Centro Hospitalar Póvoa de Varzim - Vila do Conde
Centro Hospitalar Tondela Viseu
Centro Hospitalar Universitário de São João
Centro Hospitalar Universitário do Algarve - Hospital de Portimão
Centro Hospitalar Universitário do Porto - Hospital de Santo António
Centro Hospitalar Universitário Lisboa Central
Centro Hospitalar Universitário Lisboa Norte
Clínica Lusíadas
Clínica Parque dos Poetas
General Lab Portugal
Hospital Beatriz Ângelo - Loures
Hospital da Luz Lisboa
Hospital da Senhora da Oliveira Guimarães
Hospital de Braga
Hospital Espírito Santo - Évora
Hospital Garcia de Orta - Almada
Hospital Santa Luzia – Elvas
SAMS Lisboa
Synlab
Unidade Local de Saúde de Matosinhos - Hospital Pedro Hispano
Unidade Local de Saúde do Alto Minho - Hospital Santa Luzia de Viana do Castelo