UNIVERSITY OF GOTHENBURG

DEPARTMENT OF BIOLOGICAL AND ENVIRONMENTAL SCIENCES

IDENTIFICATION OF CLINICALLY RELEVANT STRAINS OF THE STREPTOCOCCUS MITIS GROUP BASED ON SEQUENCES COMPARISONS OF HOUSEKEEPING GENES 16S rRNA, *GROEL*, AND *RPOB*

Rosa Márcia De Souza Eriksson

Degree project for Master of Science (120 hec) with a major in Biology

BIO 727, Physiology and Cell Biology (60 hec)

Second cycle

Semester/year: Autumn 2024- Summer 2025)

Supervisor: Edward Moore, Francisco Salvà Serra & Liselott Svensson, Department of Infectious Diseases, University of Gothenburg

Examiner: Henrik Nilsson, Department of Biological and Environmental Sciences

Table of Contents

Abstract 2
Sammanfattning 3
1. Introduction 4
1.1 The genus Streptococcus 4
1.2 The classification of genus *Streptococcus* 4
1.3 The Streptococcus mitis group 5
1.4 Taxonomy of *Streptococcus* mitis group (SMG) 5
1.5 Pathogenicity and antibiotic resistance in the *Streptococcus* mitis group 6
1.6 Clinical relevance of the *Streptococcus* mitis group (SMG) 6
1.7 The Culture Collection Gothenburg University (CCUG) 7
1.8 Methodologies for taxonomic identification 7
1.8.1 Housekeeping genes in Bacteria 7
1.8.2 Polymerase Chain Reaction (PCR) 8
1.8.3 DNA Sequencing – Sanger Sequencing 8
1.8.4 Multi-locus sequence analysis (MLSA) 9
1.8.5 BioNumerics and BLAST (Basic Local Alignment Search Tool) 9
1.9 Project aim 9
2. Materials and Methods 9
2.1. Bacterial Strains 10
2.2. Bacterial strains cultivation 12
2.3. Bacterial DNA extraction 13
2.4. DNA amplification by Polymerase Chain Reaction (PCR) 13
2.5. Gel electrophoresis 14
2.6. Preparation of DNA samples for sending to Eurofins for Sanger Sequencing 14
2.7 BioNumerics and BLASTn 15
3. Results 16
3.1 Cultivation of strains 16
3.2 DNA extraction and PCR results. 16
3.3 Sanger sequencing, BioNumerics and BLASTn 18
3.4 Comparisons of the 16S rRNA gene, *groEL* and *rpoB BLAST results* 19
3.5 Identification of the strains as SMG at species level 20
3.6 Phylogenetic tree 22
4. Discussion 25
5. Conclusion 27
6. References 28
Appendix 1- Popular Science summary 35
Appendix 2 - PCR templates to the genes 16S rRNA, *groEL and rpoB* 36
Appendix 3 – BLAST results 39
Appendix 4- Phylogenetic trees of the housekeeping genes 16S rRNA, groEL and rpoB 42

Abstract

The *Streptococcus* Mitis group (SMG) is composed of closely related species, making it often difficult to distinguish them from each other. It represents a significant challenge for both taxonomy and the routines by clinical diagnostics laboratories. Precise identification of SMG members is crucial, since it contains highly pathogenic species, such as *S. pneumoniae* and *S. mitis,* which are major causes of invasive infections worldwide. Misidentification can lead to wrong antimicrobial therapy and contribute to the development of antimicrobial resistance. While single-gene sequencing methods, such as 16S rRNA gene analysis, are widely applied, it is not recommended for identification of species of the SMG because of their limited discriminatory power.

In this study, 33 bacterial strains are isolated from clinical samples, previously classified by the Culture Collection Gothenburg University (CCUG), as *S. mitis*, *S. mitis* complex or *S. mitis g*roup, were analyzed using multi-locus sequence analysis (MLSA) with three housekeeping genes: 16S rRNA, *groEL*, and *rpoB*. Of these, 31 strains produced high-quality 16S rRNA gene sequences, while 30 yielded reliable sequences for *groEL* and *rpoB*. Comparative analysis against GenBank references revealed that 16S rRNA gene alone was insufficient for species-level resolution due to the high sequence similarity among SMG members. In contrast, *groEL* and *rpoB* provided greater discriminatory power, particularly when used in combination.

The results in this study demonstrated that the combination of multiple housekeeping genes significantly improves resolution compared to single-gene approaches. It emphasized the importance of MLSA as a robust method when identifying SMG with high accuracy.

The implementation of such strategies in diagnostic microbiology can raise the knowledge about SMG's functional structure, improve clinical decision-making in health care, strengthen epidemiological surveillance, and reduce the risks associated with antimicrobial resistance.

Keywords: *Streptococcus* Mitis group, MLSA, 16S rRNA gene, *groEL*, *rpoB*, bacterial identification, housekeeping genes

Sammanfattning

Streptococcus mitis gruppen (SMG) anses ha nära besläktade arter, vilket ofta gör det svårt att särskilja arterna från varandra. Detta innebär en stor utmaning för både taxonomin och för kliniska diagnostiska laboratorier. Noggrann identifiering av SMG-medlemmar är avgörande eftersom den gruppen innehåller potenta patogena arter, såsom *S. pneumoniae* och *S. mitis*, som orsaka flera invasiva infektioner runt om i världen. Felaktig identifiering kan leda till felaktig antimikrobiell behandling samt bidra till utvecklingen av antimikrobiell resistens. Metoder med en enda gensekvensering, såsom 16S rRNA-genanalys, används i stor utsträckning, även om det inte rekommenderas för identifiering av arter av SMG på grund av deras begränsade särskiljningsförmåga.

I denna studie analyserades 33 bakteriestammar isolerade från kliniska prover, tidigare klassificerade av ”Culture Collection Gothenburg University” (CCUG), som S. *mitis* eller *S. mitis*-komplex eller *S. mitis*-grupp, med hjälp av multi-lokus-sekvensanalys (MLSA) med tre ”housekeeping genes”: 16S rRNA, *groEL* och *rpoB*. Av dessa, producerade 31 stammar högkvalitativa 16S rRNA-gensekvenser, medan 30 stammar gav tillförlitliga sekvenser för *groEL* och *rpoB*. Den jämförande analysen av studerade stammar mot GenBank-referenssekvenserna visade att 16S rRNA-genen ensam var otillräcklig för upplösning på artnivå på grund av den höga sekvenslikheten mellan SMG-medlemmar. Däremot gav *groEL* och *rpoB* större urskiljningsförmåga, särskilt när de användes i kombination.

Resultaten i denna studie belyser vikten av MLSA som en robust metod för korrekt SMG-identifiering, vilket visar att kombinationen av flera ”housekeeping genes” avsevärt förbättrar upplösningen jämfört med metoder med en enda gen. Implementeringen av sådana strategier inom diagnostisk mikrobiologi kan öka kunskapen om SMG:s funktionella struktur, förbättra kliniskt beslutsfattande inom hälso- och sjukvården, stärka epidemiologisk övervakning och minska riskerna i samband med antimikrobiell resistens.

Nyckelord: *Streptococcus* mitis grupp, MLSA, 16S rRNA-gen, *groEL*, *rpoB*, bakteriell identifiering, ”housekeeping genes”, antimikrobiell resistens.

1. Introduction

1.1 The genus Streptococcus

Bacteria of the genus *Streptococcus* were first observed in 1868 by Theodor Billroth and, since 1884, they have been classified into the family *Streptococcaceae,* order *Lactobacillales* and phylum *Firmicutes* (Deibel and Seeley, 1974; Ludwig et al., 2009; Schleifer, 2009). The genus *Streptococcus* is comprised of facultative anaerobic organisms, which are Gram- positive and catalase-negative (Schleifer, 2009; Hossain, 2014). Morphologically, they are non-motile, do not form spores, undergo cell division along a single axis, which leads to the formation of pairs, clusters or chains (Hossain, 2014).

This genus is present in several different environments and also living in organisms. In humans it is established predominantly in the nasopharynx and oral cavity, once it is integrated into the normal microbiome (Murray, 2018; Wei, et al., 2023; Sadowy and Hryniewicz, 2020).

The species differ significantly in pathogenicity, where in some species are described as commensals, while others, such as *Streptococcus pneumoniae* are highly virulent, causing severe invasive infectious diseases in children, older people and immunosuppressed humans (Sadowy and Hryniewicz, 2020; Doern and Burnham, 2010).

Streptococcus spp. can easily adapt to new hosts, impacting and shifting their immune system by several strategies, such as expression of proteins that aid it to colonize the host cells and survive competing for nutrients (Mitchell, 2011). The genetic recombinant events contribute to increased antimicrobial resistance and spreading of diseases involving *Streptococcus* (Sadowy and Hryniewicz, 2020; Doern and Burnham, 2010).

1.2 The classification of genus *Streptococcus*

Clinically, the genus *Streptococcus* is categorized into the pyogenic and non-pyogenic, or Viridians group Streptococci (VGS) (Sadowy and Hryniewicz, 2020). The haemolytic reaction of *Streptococcus* colonies when cultivated in agar plates, allows them to be classified into three different groups: (A) α-haemolytic species, that oxide the iron in haemoglobin molecules within red blood cells, resulting in a greenish colour on Blood Agar (Murray, 2018; McDevitt et al., 2020) and presenting a yellowing colour when cultivated on Chocolate Agar, due to the production of peroxides that react with the blood cells in the chocolate agar plates (Gunn, 1984); the *Streptococcus* mitis group (SMG) are included in this group; (B) β-haemolytic species, also known as Lancefield group, which completely lyse red blood cells (Murray, 2018, Bloch et al., 2024); and (C) γ-hemolytic species, with no lysis of red cells (Murray, 2018; Bloch et al., 2024) (Figure 1).

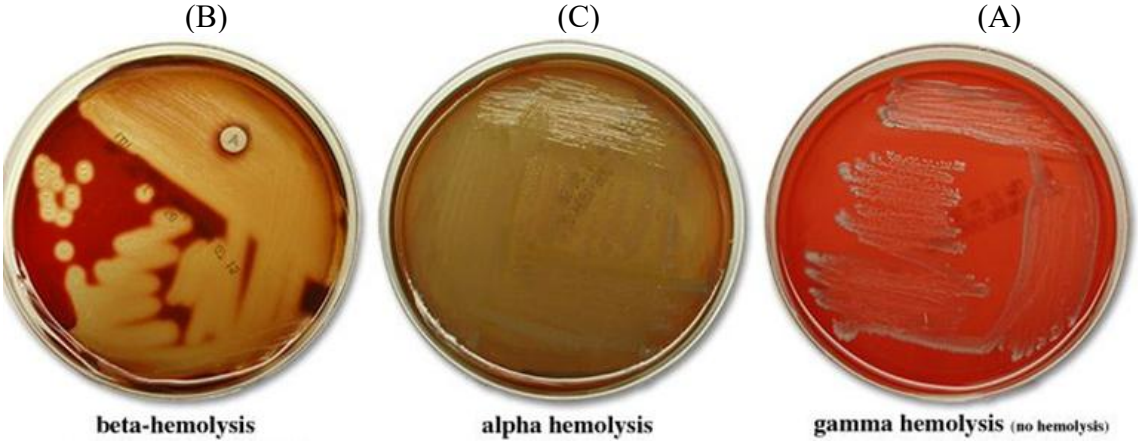

Figure 1. Blood agar plates showing different types of haemolysis (Source: Abdulla, 2015)

Group β-hemolytic *Streptococcus* are further classified, using the serological properties, also named Lancefield classification, based on serotyping of the cell-wall polysaccharides (Murray, 2018; Bloch et al., 2024). Lancefield classification is not reliable when identifying *Streptococcus* mitis group (SMG) to the species level because the Lancefield antigens in this group are missing (Doern and Burnham, 2010; Facklam, 2002; Anonymous, 2021).

Based on phylogenetic relationships between *Streptococcus*, the 16S rRNA gene sequencing technique made it possible to categorize the genus *Streptococcus* into eight groups: *mitis*, *sanguinis*, *anginosus*, *salivarius*, *downei*, *mutans*, *pyogenic*, and *bovis* groups (Bloch et al., 2024). The genetic similarity is high in these groups and advanced, and more specific molecular biology methods are required to get deeper genetic information, providing better comprehension about evolutionary relationships among streptococcal species (Sadowy et al., 2020; Kawamura et al., 1999). In this way, the sequences of housekeeping genes have successfully been used to assess clustering patterns among very closely related species, making it easier to identify the species of SMG (Sadowy et al., 2020; Kawamura et al., 1999; Bishop et al., 2009). Techniques for analysing whole genome sequences, such as digital DNA-DNA hybridization (dDDH) and average nucleotide identity (ANI), have also been used in studies to accurately identify clinical bacterial isolates contributing to the taxonomy and nomenclature of new species (Versmessen et al., 2024). Those techniques still present some disadvantages, such as being time consuming and having relatively high costs (Gao et al., 2014; Versmessen et al., 2024).

1.3 The Streptococcus mitis group

The *Streptococcus* mitis group (SMG) is part of the human normal microbiota with some species being commensal, while others are pathogenic and associated with serious clinical infections, causing millions of deaths every year around the world (Sherman, 1937; WHO, 2019; O'Brien et al., 2009).

This group is represented by at least 26 species, three subspecies and four not validly published: *S. gordonii*, *S. mitis, S. pneumoniae*, *S. sanguinis, S. oralis*, *S. parasanguinis*, *S. cristatus*, *S. infantis*, *S. peroris, S. pseudopneumoniae S. australis*, *S. panodentis*, *S. oligofermentans, S. orisratti, S. downii, S. massiliensis, S. lactarius, S. rubneri, S. panodentis, S. sinensis* and *S. oricebi* (Jensen et al., 2016; Kilian et al., 2025). The names of the following species are validly published after being identified and featured as belonging to the SMG by studies of only one isolate: *Streptococcus toyakuensis*, *Streptococcus thalassemiae*, *Streptococcus chosunensis*, *Streptococcus humanilactis, Streptococcus gwangjuensis* (Kilian et al., 2025). The subspecies of *S. oralis* are subsp. *tigurinus,* subsp *dentisani* and subsp. *oralis* (Kilian et al., 2025).

Finally, there are also other species considered to belong to SMG, but whose names are not yet validly published. Those species have their names indicated by apostrophes ("-") because its species do not yet have taxonomic standing. Those species include: *"Streptococcus vulneris*", "*Streptococcus shenyangsis"*, "*Streptococcus bouchesdurhonensis" and "Streptococcus symci"* (Kilian et al., 2025).

1.4 Taxonomy of *Streptococcus* mitis group (SMG)

The genetic diversity within the genus *Streptococcus* is still a challenge and explains the high number of taxonomic revisions done during the last decades, aiming to make it easier to distinguish species from each other (Facklam 2009; Kawamura et al., 1999; Sadowy et al., 2020). The high similarities between SMG's species also reflects the complexity of the antimicrobial resistance pattern, for which the accurate identification to the species level is crucial for deciding the most effective first-line empirical antimicrobial medicines to treat infections caused by *SMG* -species (Sadowy et al., 2020, Chun et al., 2015). The development of high-throughput DNA sequencing technologies (including next-generation sequencing and third generation sequencing) has aided to correctly identify species of SMG, although it is still a challenge, since the diversity in this group

is huge and a significant part of its diversity remains unexplored (Kawamura et al., 1999; Sadowy et al., 2020).

1.5 Pathogenicity and antibiotic resistance in the *Streptococcus* mitis group

The pathogenicity of *Streptococcus* mitis group (SMG) is regulated by many genes occurring in different species, showing the diversity of regulatory events involved, for example, the capsule production and the importation of enzymes for the metabolisms of carbohydrate and proteins (Kilian and Tettelin, 2019, Gonzales-Siles et al., 2019).

The evolutionary hypotheses, considering the bacterial competence for genetic transformation, explains common ancestry, where in the horizontal gene transfer (HGT) (Fig. 2) among SMG's species plays a central role (WHO, 2019, Gonzales-Siles et al., 2019, Straume et al., 2015). Through it, some species of SMG have acquired virulence-promoting genes conferring pathogenicity (Doern and Burnham, 2010; Straume et al., 2015). According to Killian & Tettelin (2019), some commensal species of the *SMG* lost virulence-associated genes and developed a harmonious relationship with the host, getting genetic stability and advantages when living in biofilms (Kilian and Tettelin, 2019). Conjugative transposons also have been found carrying antibiotics resistance genes in SMG (Straume et al., 2015).

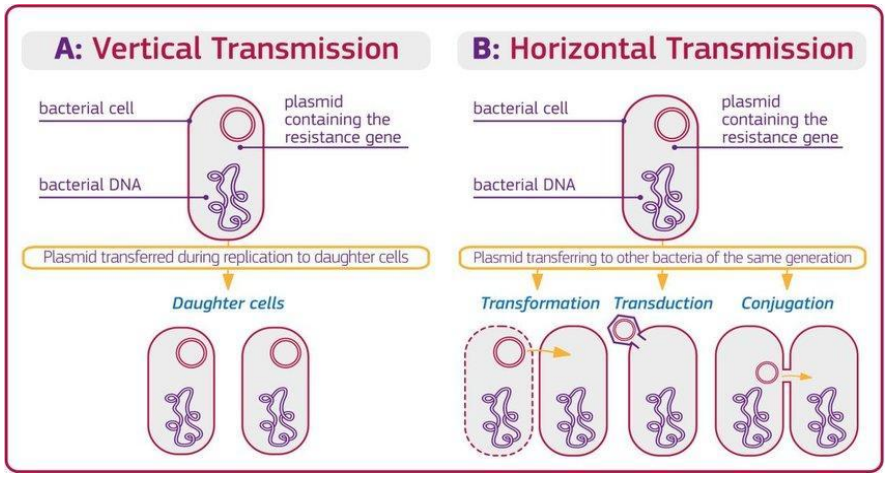

Figure 2. Vertical and horizontal gene transfer in bacteria (Sanseverino *et al*., 2018).

The identification to the species level is significant for differentiating pathogenic from non-pathogenic species, considering these bacterial groups exhibit contrasting drug susceptibility patterns during the treatment of infectious diseases (Belman et al., 2022; Sanseverino et al., 2018). Especially the species *S. mitis* that is considered to be both a reservoir of antibiotic resistance genes and the species that works as source for capsule polysaccharide diversity for the species *S. pneumoniae,* considered highly pathogenic, contributing to vaccine escape (Salvadori et al., 2019).

1.6 Clinical relevance of the *Streptococcus* mitis group (SMG)

Some species of SMG cause life-threatening infectious diseases, such as meningitis, pneumonia, infective endocarditis and bacteremia, when they reach the bloodstream, particularly in immunosuppressed patients, elderly and children, especially under five years old (Sadowy and Hryniewicz, 2020; Doern and Burnham, 2010; Kilian and Tettelin, 2019).

According to the World Health Organization (WHO), only the species *S. pneumonia*e devastates the lives of one million children every year (WHO, 2019; O'Brien et al., 2009).

A report written by Global Burden of Diseases, Injuries, and Risk Factors Study (GBD, 2021), showed that *S. pneumoniae* was one of the five pathogens, responsible to more than 500 000 deaths in 2019, where the most representative group were children younger than 5 years and living in the sub-Saharan Africa (GBD, 2022). In another report also published in 2024 by GBD, it was shown

that the *S. pneumoniae* was globally responsible for causing lower respiratory infections in circa 97,9 million and 505 000 deaths in 2021 (GDB, 2024). WHO considers *S. pneumoniae* to be a major public health problem worldwide, which explains the epidemiological and clinical relevance of SMG (WHO, 2019). The present situation, in some cases, is getting worse, since the identification of bacterial pathogens is far from contributing to an effective treatment, which decreases risk of wrong prognosis and decreases antimicrobial resistance in new generations (Rentschler *et* al., 2021).

1.7 The Culture Collection Gothenburg University (CCUG)

The Culture Collection Gothenburg University (CCUG), the largest public collection of clinically relevant bacteria in Europe, offers typing services utilizing phenotypic, chemotypic, genotypic and whole- genome sequencing analyses achieving accurate, detailed classification and identification, that comprises more than 78000 bacterial strains, included more than 4000 type strains (www.ccug.se). The convention on biological Diversity (CBD) and the Nagoya Protocol on Access and Benefit Sharing (ABS) are some regulations that CCUG follow when deciding which type of microorganisms received from environments source, clinical and industry can be added to the collection (www.ccug.se). In the Department of Infectious Diseases at Gothenburg University, it is located at the Institute of Biomedicine, which CCUG is associated with (www.ccug.se). The CCUG is included as part of the Bacterial Molecular Diagnostics section at the Department of Clinical Microbiology belonging to the Academy Hospital Sahlgrenska (www.ccug.se). The clinically relevant strains of the *S. mitis* group studied in this project were supplied by CCUG.

1.8 Methodologies for taxonomic identification

1.8.1 Housekeeping genes in Bacteria

Housekeeping genes encode proteins which manage the main functions in living organisms, such as basic metabolism, transports into the cell and the cell cycle (Wei and Ma, 2018; Joshi et al., 2022). In this project, the housekeeping genes 16S rRNA, *groEL* and *rpoB* genes were used. They are impressively conserved with a high level of genetic, evolutionary information and often used as genetic or taxonomic markers (Wei and Ma, 2018; Ogier et al., 2019).

The 16S rRNA is found in all bacterial strains, often as components of a multi-copy multi-gene family or operons, with functional structures well conserved. This enables designing common primers for PCR amplification assays for all bacteria (Janda and Abbott, 2007; Větrovský and Baldrian, 2013). The 16S rRNA gene also contains hypervariable regions named as V1–V9 (Figure 3). These regions thus have high similarity in the SMG, making it difficult to distinguish the species from each other (Větrovský and Baldrian, 2013).

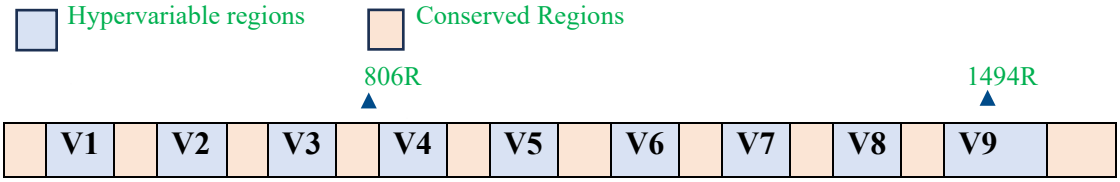

Figure 3. 16S rRNA gene and its conserved and hypervariable regions

The hypervariable regions of 16S rRNA gene have been studied both combined or alone, and have shown significant differences in taxonomic identification, once each region provides varied sensitivity and specificity to different bacterial genus, where the amplicon and the protocol used, such as the chosen regions and analytical methods can influence the bacterial identification accuracy levels (Sperling et al., 2017; López-Aladid et al., 2023). The 16S rRNA gene sequence is highly used when classifying bacterial strains, although it sometimes contains a low power of resolution when identifying at species level, showing very low discriminatory power for some genera (Janda and Abbott, 2007), but it can then be complemented with other housekeeping genes, also used as taxonomic markers (Ogier et al., 2019; Glazunova et al., 2009).

The *groEl* gene belongs to the chaperonin's protein complexes, found in all organisms and involved in the repair of misfolded polypeptides produced under stressful conditions (Glazunova et al., 2009). Previous studies have considered the *groEL* gene to be an optimal tool for distinguishing species of the SMG, as it provides insights into similarities and dissimilarities through phylogenetic analysis (Glazunova et al., 2009; Ishii, 2017).

The *rpoB* gene, encoding the β-subunit of RNA polymerase, is another housekeeping gene that is involved in essential functions, such as catalytic activity, DNA transcription, mRNA and ribonucleoside binding (Glazunova et al., 2009, Adékambi et al., 2009). The *rpoB* gene is long considered a better marker than 16S rRNA gene when identifying SMG at a species level (Glazunova et al., 2009; Adékambi et al., 2009).

1.8.2 Polymerase Chain Reaction (PCR)

Polymerase Chain Reaction (PCR) is a highly specific and sensitive molecular technique that allows amplification of a DNA region (Kralik and Ricchi,2017). To perform a PCR, it is necessary to have specific primers, DNA polymerase, nucleotides, specific ions, and a pre-selected DNA template. It consists of cycles that comprise the following steps: (1) DNA denaturation, where high temperatures (95˚C), break the hydrogen bonds and separate the double strands of DNA; (2) primer annealing or hybridization allows the hydrogen bonds to re-build because the denatured DNA is cooled at a temperature between 37-72°C depending on the length and sequence of the primers; (3) DNA extension or elongation occurring at 72°C (Adékambi et al., 2009, Kralik and Ricchi,2017). Those three PCR steps are repeated in several cycles producing many copies of the DNA sequence of interest (Kralik and Ricchi,2017; Khehra et al., 2023). In the PCR was used the enzyme DNA polymerase from the bacterium Thermus aquaticus (Taq polymerase), and it is a thermostable protein, which prevents physical and chemical modification in the DNA and RNA structures (Kralik and Ricchi,2017; Khehra et al., 2023).

1.8.3 DNA Sequencing – Sanger Sequencing

Sanger sequencing is a method developed by Fredrick Sanger to sequence DNA and is the most-used sequence method in the world during the last 48 years (Sanger et al., 1977; Crossley et al., 2020). This method is considered the gold standard, involving electrophoresis and random integration of chain-terminating dideoxynucleotides by DNA polymerase enzymes; it has been crucial to determine nucleic acid sequences (Crossley et al., 2020; Chait et al., 1988). This method allows identifying new pathogens, new genotypes of known pathogens, unravelling evolutionary changes in the pathogen's genome important to make phylogenetic analysis and epidemiologic studies (Crossley et al., 2020).

Basically, the method comprises of: (1) amplification of the DNA template (target fragment) or the complementary DNA (cDNA); (2) annealing of the DNA /cDNA to an oligonucleotide primer; (3) elongation of by the DNA polymerase. Those processes require the four deoxyribonucleotide triphosphates (dNTPD) comprising DNA, and the four dideoxyribonucleotide triphosphates (ddNTPs), marked with fluorescence labels and without 3-OH group, that work as terminators of the primers extension (Chait et al., 1988), resulting in DNA sequences of different sizes (Crossley et al., 2020).

The application of Sanger sequencing after the PCR amplification of genes, using conserved primers specifically focused on hypervariable regions on these genes, allows the identification of bacterial strains at the genus and, in many cases, at the special levels (Gu & Chiu, 2019). This method has been useful in research into microbiome, metagenomics and diagnostics of complex bacterial infections diseases (Gu & Chiu, 2019).

1.8.4 Multi-locus sequence analysis (MLSA)

Multi-locus Sequence analysis (MLSA) is a reliable approach that allows phylogenetics analysis when identifying bacterial species based on comparison of sequences between different sets of housekeeping genes (Bishop et al., 2009; Jensen et al., 2016).

Imai et al. (2020) found discrepancies in their study when comparing the MLSA methods with other methods. However, in another study, it is described that phylogenetics analysis based on MLSA and the whole-genome core sequencing are the most accurate procedure to identify species of the SMG (Jensen et al., 2016). Unfortunately, it is not always used, due to the high costs, required trained personnel and time consuming (Imai et al., 2020). By MLSA analysis sequences of taxonomically challenging groups of bacterial strains, can be analysed and compared to publicly available sequences in GenBank using for instance BLASTn and BioNumerics (Bishop et al., 2009). While effective for many common pathogens, it may struggle with rare or previously unidentified species and can sometimes be limited by the availability of sequences for all species of a genera.

1.8.5 BioNumerics and BLAST (Basic Local Alignment Search Tool)

BioNumerics is a useful software when working with multi-locus sequence analysis (MLSA), to differentiate bacterial strains based on the sequences of internal fragments of multiple housekeeping genes. BioNumerics allows analysing DNA sequences and editing the sequences by removing (trimming) primer sequences or poor-quality sequences from reads, increasing the quality of the sequences (www.aphl.org). It also allows defining the correct nucleotide in the positions where multiple nucleotides overlap in one position. Furthermore, it allows sequence alignment and phylogenetic analysis (https://oit.va.gov). It is easy to import sequence data from the National Center for Biotechnology Information (NCBI) into BioNumerics database.

The sequences can be run in the NCBI's Basic Local Alignment Search Tool (BLAST) to find the best hits to those unknown strains. BLAST is a bioinformatic tool used to find similar regions between different nucleotide or protein sequences, presenting significant statistical results that contribute to identifying organisms, at the same time, providing understanding of the functional and evolutionary relationships between the organism's sequences (https://guides.lib.berkeley.edu/; https://compss-doc.readthedocs.io/) . GenBank is a database of the National Library of Health (NLH) genetic sequence, belonging the NCBI, USA (United States of America), and is also a part of the International Nucleotide Sequence Database Collaboration, which includes the DNA DataBank of Japan (DDBJ), Europe (ENA) and The National Center for Biotechnology Information (NCBI) in USA (www.ncbi.nlm.nih.gov).

1.9 Project aim

The aim of this project is to accurately identify 33 unidentified clinical strains of clinically relevant *S.* Mitis-Group (SMG) using the DNA sequences of the housekeeping genes 16S rRNA, *groEL*, and *rpoB*. The hypothesis is that the studied clinically relevant strains will exhibit individual genetic variations in these genes, which will allow identification to the species level. The resulting phylogenetic clusters will also reveal evolutionary relationships between the studied strains of SMG.

2. Materials and Methods

The genetic materials used for this project were obtained from bacterial strains archived and provided by CCUG. All reagents, protocols, computers and material used in DNA extraction process, electrophoresis, analyses of sequencing data and PCR were also provided by the Department of Infectious Disease, Institute for Biomecine, of the University of Gothenburg and the laboratory of CCUG, where the practical work of this project was also run. The research methods used in this project were carried out following the CCUG's protocols. The cultivation

media plates used for growing the strains were produced at the Substrate Unit, Department of Clinical Microbiology, Sahlgrenska University Hospital. The 33 bacterial strains studied in this project were archived and previously identified by CCUG as *S. mitis* or *S. mitis* complex or *S. Mitis*-Group. The experimental design is illustrated in Figure 4.

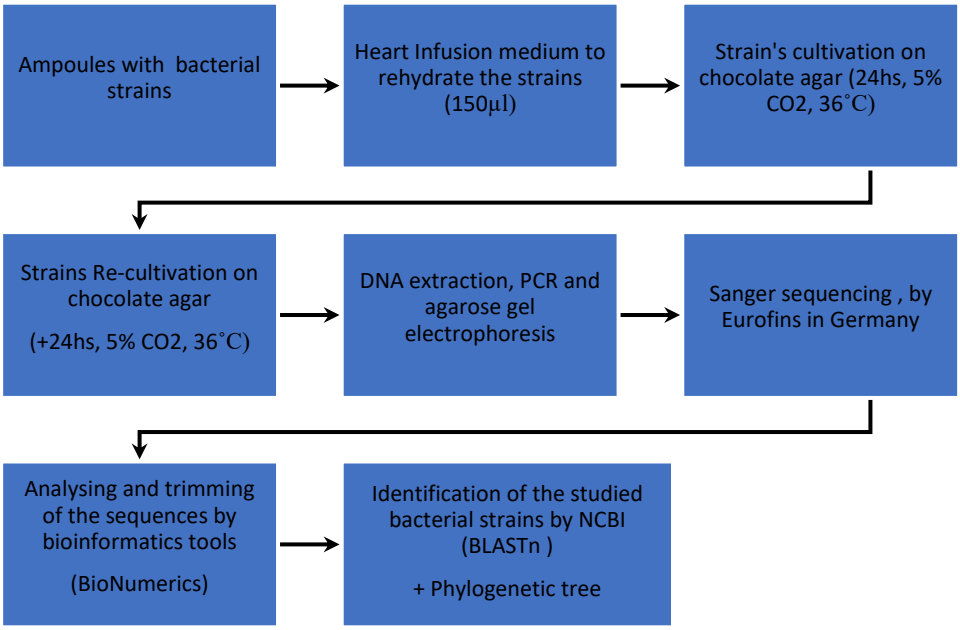

Figure 4. Experimental design for identification of SMG strains using housekeeping genes.

2.1. Bacterial Strains

A total of 33 bacterial strains were included in this project. The bacterial strains originated from clinical samples, such as blood, brain abscess, spine, oral cavity, bronchi, prosthesis, throat and sputum from patients admitted in hospitals inside and outside of Sweden. They were received and archived at CCUG between 1989 until 2007 (Table 1). Any patient information was confidential, which means that no ethical approval was required for this project. The bacterial strains were selected by the following three criteria: (1) Those 33 strains had a preliminary identification as *S. mitis* or *S. mitis* complex or *S. mitis* group at species level; (2) They should not be identified at the species level, or the species-level identification was still questionable; or (3) They should not have housekeeping genes determined, with exception to the strain CCUG 55622B, which had been analysed by 16S rRNA gene and *sodA*.

Table 1. Bacterial strains studied with the respective CCUG accession number, origin and source.

CCUG Accession Number	Archive date	Preliminary Identification (Used method)	Origin	Source
19074	1986.07.24	S. mitis (API*- biochem)	Sweden, Gothenburg	Blood
25812	1989.12.06	S. mitis (API- biochem, Tests CAT**)	Sweden, Skövde	CAPD (Continuous Ambulatory Peritoneal Dialysis)- fluid
26924	1990.07.12	S. mitis (API- biochem)	Sweden, Uddevala	Tissue

Table 1. Continued**.** Bacterial strains studied with the respective CCUG accession number, origin and source.

CCUG Accession Number	Archive date	Preliminary Identification (Used method)	Origin	Source
27740	1991.01.08	S. mitis (API- biochem, SDS-PAGE*** GelCompar)	Denmark Århus,	Dental plaque
27741	1991.01.08	S. mitis (API- biochem, SDS-PAGE-Gel Compar)	Denmark Århus,	Dental plaque
28754	1991.07.15	S. mitis (API- biochem)	Sweden, Gothenburg	Blood, newborn
31489	1993.05.06	S. mitis (Api biochem)	Sweden Gothenburg	Blood
31557	1993.05.26	S. mitis API- biochem, Test CAT)	Sweden, Gävle	Blood
32108	1993.10.06	S. mitis (API- biochem)	Sweden, Jönköping	Blood
32132	1993.10.14	S. mitis (API- biochem, TEST CAT)	Sweden, Jönköping	Blood
32331	1993.11.30	S. mitis (API- biochem)	Sweden, Uppsala	Blood
32466	1993.12.27	S. mitis API- biochem	Sweden, Gävle	Bronchi
32601	1994.02.08	S. mitis API- biochem	Sweden, Kalmar	Blood, brain abscess
33514	1994.11.03	S. mitis (API- biochem, Test CAT)	Sweden, Gothenburg	Blood
33690	1994.12.09	(S. mitis API- biochem, Test CAT)	Sweden, Gothenburg	Blood, premature
35276	1996.01.12	S. mitis (API- biochem, test CAT)	Sweden, Uppsala	Throat, relapse after treatment
35278	1996.01.12	S. mitis (API- biochem, test CAT)	Sweden, Uppsala	Throat, relapse after treatment
35580	1996.03.26	S. mitis (API- biochem, test CAT)	Sweden, Östersund	Prosthesis infection
36639	1996.08.27	S. mitis (API- biochem, test CAT)	Sweden, Borås	Blood
36753	1996.09.25	S. oralis (API- biochem, test CAT)	Sweden, Uddevalla	Blood

Table 1. Continued. Bacterial strains studied with the respective CCUG accession number, origin and source.

CCUG Accession Number	Archive date	Preliminary Identification (Used method)	Origin	Source
36755	1996.09.26	S. oralis (API- biochem, test CAT)	Sweden, Boden	Blood
37308	1997.01.03	S. mitis (API- biochem, test CAT)	Sweden, Sundsvall	Blood, neutropenic
39096B	1998.03.11	S. mitis (API- biochem, test CAT)	Sweden, Linköping	Urine
39210	1998.03.31	S. mitis (API- biochem, test CAT)	Sweden, Linköping	Blood
41450	1998.11.20	S. mitis (API- biochem, test CAT)	Sweden, Boden	Blood
42636	1999.09.02	S. mitis (API- biochem, test CAT)	Sweden, Karlskrona	Eye
42984	1999.11.19	*S. mitis* (API- biochem, test CAT)	Sweden, Karlstad	Blood
44763	2000.12.19	*S. mitis* (API- biochem)	Sweden, Falun	Blood
44969	2001.02.27	S. mitis (API- biochem, test CAT)	Sweden, Göteborg	Blood
49591	2004.07.30	S. mitis (API- biochem, test CAT, SDS-PAGE-GelCompar)	Sweden, Göteborg	Sputum
55622B	2007.12.11	S. mitis (API- biochem, test CAT, 16S rRNA and SodA genes)	Sweden, Göteborg	Blood

* API(Analytical Profile Index)-tests measure bacteria's capacity to ferment carbohydrates, to use amino acids and other biochemistry processes. (*Source*: https://www.biomerieux.com).

**CAT(Catalase Test) test - Detects the presence of the catalase enzyme in bacterial species (Khatoon et al., 2022)

***SDS-PAGE (Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis) - Separate and analyzed proteins based primarily on their molecular weight (Kielkopf et al., 2021)

2.2. Bacterial strains cultivation

The bacterial strains had been preserved freeze-dried (lyophilized), in sealed glass ampoules, at 4˚C. The ampoules were opened under aseptic conditions. The rehydration and resuspension of the dried bacterial biomass was done by adding 150 µl of BHI (Brain-Heart Infusion) Broth medium into the ampoules and mixing it until the suspension exhibited a homogenic consistency. A total of 50 µl of the bacterial strain solution was inoculated onto fresh Chocolate Agar plates, following the four quadrant streaking techniques. All strains were first incubated for 24 hours at 36˚˚ degrees and 5% CO_2 atmosphere (Thermo Scientific CO_2 incubator). The strains that grew well in the first cultivation on Chocolate Agar plates, were re-streaked on new Chocolate Agar plates and submitted for a new cultivation for 24 hours in the same conditions. The first eleven strains were also

cultivated on Blood Agar plates, on which they did not grow. The rest of twenty-two strains were cultivated only on chocolate agar. A loop-full of bacterial culture biomass from the second generation of colonies was used for DNA extraction, while the remaining part of the biomaterial was mixed in 1,000 µL glycerol, frozen and saved at -20 ˚C. For each of the three PCRs, the strains were cultured in the same way, and new batches of DNA were prepared.

2.3. Bacterial DNA extraction

The extraction of genomic DNA was done through bead-beating technique. Circa 1-2 µl of biomass was transferred to 1.5 ml Eppendorf tubes, containing 150 µl TE-buffer (Tris HCl 10mM + EDTA 0.1 mM) produced by Substrate Unit, Clinical Microbiology, Sahlgrenska University Hospital. The next step was transferring this mixture to an Eppendorf tube of 0.2 ml, containing a recommended amount of glass beads (acid-washed, 150-212 µm, Sigma Aldrich) for "bead-beating" to lyse the bacterial cells. The bead-beating machine, the TissueLyserII (Qiagen, Hilden, Germany), was used at 25 frequency 1/s for 5 minutes, to disrupt the cells in the samples and maximize DNA yields. The suspension of the broken cells was spun down, and the supernatant was transferred to new 0.5 ml Eppendorf tubes, before new centrifugation at 13000 rpm for 10 minutes. The supernatant with the extracted DNA, was transferred to new Eppendorf tubes, diluted and used as template for PCR. When the PCR could not be done on the same day, the supernatant containing the extracted DNA was saved at 4°C for a maximum 24 hrs.

2.4. DNA amplification by Polymerase Chain Reaction (PCR)

The extracted DNA was used to perform the PCR amplification of the housekeeping genes 16SrRNA*, groEl and rpoB.* The forward and reverse specific primers for each gene were produced by Eurofins Genomics (Germany) (Table 2).

Table 2**.** Information about the gene and its specific primers sequences, size and use.

Gene	Primer	Sequence	Reference	Product Size (bp)	Use
16S rRNA	16F28	5'- AAG AGT TTG ATC MTG GCT CAG-3'	Weisburg et al. (1991)	~ 1466	PCR
16S rRNA	16R1494	5'- NTA CGG YTA CCT TGT TAC GAC-3'	Weisburg et al.(1991)		PCR
16S rRNA	16R806 reverse	5'- GGA CTA CCA GGG TAT CTA AT-3'	Fredricks and Relman (1998)	~ 786	Sequencing
groEL	*groEL*-575F	5'- GAH GTN GTN GAA GGN ATG CA-3'	Glazunova et al. (2009)	~ 798	PCR - Sequencing
groEL	*groE*L-1333R	5'- ATT TGR CGN AYW GGY TCT TC-3'	Glazunova et al. (2009)		PCR - Sequencing
rpoB	*rpoB*-F	5'- NNN AAR YTN GGM CCT GAA GAA AT-3'	Ogier et al. (2019)	~ 742	PCR - Sequencing
rpoB	rpoB-R	5'- GNA RTT TRT CAT CAA CCA TGT G-3'	Ogier et al. (2019)		PCR - Sequencing

The low-TE buffer (Sigma 7-9 T3187, EDTA Triplex II, TritonX100) was used to re-suspend the primers, building a primer stock with a concentration of 100 mol/µl. The primer stocks were

vortexed and diluted 1:10 with milliQ H_2O, reaching a concentration of 10 µM, which was used in the PCR amplification.

The PCR amplification mixture contained: 2.5 µl of milliQ H_2O; 2.5 µl of each primer (forward + reverse) at 10µM original concentration; 12.5 µl of GoTaq Green Master Mix ($MgCl_2$, dNTPs, bacterial Taq DNA Polymerase, reaction buffers, loading dyes produced by Promega); 5µl of template DNA, that totalling 25 µl of PCR solution.

The PCR solutions were prepared in Eppendorf tube 0.2ml and then loaded in the GeneTouch thermal cycler (Bioer Technology). The PCR amplification began with an initial denaturation at 95°C for 2 minutes, followed by 30 cycles consisting of: denaturation at 95°C for 30 seconds, annealing at 55°C for 60 seconds, and extension at 72°C for 46 seconds. The PCR was completed with a final extension at 72°C for 10 minutes and the reaction mixtures were maintained at 4 °C until further analysis. The PCR amplification of the gene *groEL* required temperature adjustment in the annealing step from 55°C, used to amplify the 16S rRNA and *rpoB* genes, to 53°C, for 2 min, the extension at 72°C for 48 seconds. This adjustment can be explained by the short *groEl* primers compared with the primers for 16S rRNA and *rpoB* genes.

According to Obradovic et al. (2013) the primer optimal annealing temperature is dependent on the primer sequence length and, among other things, the content of C+G in the primer sequence. The positive control was 5 µl of the extracted DNA of the type of strain of *S. pneumoniae* CCUG 28588^T, diluted 1:10 for all three PCRs. The negative control was done with 5 µl milliQ H_2O.

2.5. Gel electrophoresis

The dilution of 1.0 g agarose (SeaKem LE Agarose, Cambrex Bio Science) in 100 ml E buffer (Tris Sigma 7-9 T3178, Natrium EDTA Triplex III, Distilled H_2O) (1X, 1000/1L, pH 7.9), produced by Substrate Unit, Department of Clinical Microbiology, Sahlgrenska University Hospital.

30 µl of GelRed Nucleic Acid Gel Stain (EMD Millipore) was added. The solution was heated in the microwave at full power for 1 minute and was shaken and stirred regularly until the agarose had been diluted and the solution became transparent. The solution rested around 10 minutes to cool and then poured into an electrophoresis gel casting tray. The gel combs were placed in the gel and after solidification, the combs were removed. The gel was transferred to the electrophoresis bath with the wells closest to the negative charge and the gel was completely covered by E-buffer (Tris Sigma 7-9 T3178, Natrium EDTA Triplex III, Distilled H_2O) (1X, 1000/1L, pH 7.9) produced by Substrate Unit, Department of Clinical Microbiology, Sahlgrenska University Hospital. PCR products (5 µl) were pipetted into the wells. The gel electrophoresis was run at 70 volts for circa 25-35 minutes. Once the electrophoresis was finished, the gels were visualized under UV-light (Genes flash – Syngene Bio imaging) and a picture was printed (video geographic printer UP 895MD- Sony). The analysis of the results of the PCR products run in gel electrophoresis was based on absence or presence of PCR products visualized as bands in the agarose gel and if the sizes of the bands were the same as for the control sample. If the bands were too faint or weak, a new PCR and new electrophoresis was performed.

2.6. Preparation of DNA samples for sending to Eurofins for Sanger Sequencing

The Sanger sequencing and purification of the PCR products was done by Eurofins Genomic Europe Sequence GmbG, in Germany. The PCR products were transferred into a new 1.5 ml Eppendorf tube and labelled with the CCUG code, and a Eurofins barcode etiquette were attached to each tube. The sequencing was done at the Eurofins Genomics after indicating which sequencing primer/s should be used. For the 16S rRNA gene sequencing was done only with a single reverse primer (16R806) (Table 2). The tubes containing the PCR products were placed in envelopes and sent to Eurofins Genomics after the digital order was done. After Sanger sequencing, the sequence

results were downloaded, and the quality of the sequence data was analysed. When the quality was low the whole procedure was repeated, including culturing the bacteria, DNA extraction, PCR, and sequencing.

2.7 BioNumerics and BLASTn

The raw sequences created by Sanger sequencing were imported into BioNumerics software (version 7.1; Applied Maths, Belgium). After visualization, the quality was checked, and editing was done manually. Including, trimming of the sequence reads, checking for the low-quality sequence reads and sequence artifacts. For the *rpoB* and *groE*L genes having the reverse and forward primers, both sequence reads were assembled, and any ambiguous nucleotides were replaced by IUPAC- codes (Table 3).

Table 3. IUPAC codes to nucleotides (source: https://www.gendx.com).

Symbol	Meaning	Origin of designation
G	G	**G**uanine
A	A	**A**denine
T	T	**T**hymine
C	C	Cytosine
R	G or A	Purine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**eto
S	G or C	**S**trong interaction (3 H bonds)
W	A or T	**W**eak interaction (2 H bonds)
H	A or C or T	Not-G, **H** follows G in the alphabet
B	G or T or C	Not-A, **B** follows A
V	G or C or A	Not-T (not-U), **V** follows U
D	G or A or T	Not-C, **D** follows C
N	G or A or T or C	A**n**y

The sequence identifications and also sequence similarities for all three genes were done by comparison with known sequences deposited in the GenBank database. The studied sequences were uploaded to BLASTn, and the three best hits were sorted by highest percent identity, where the highest one, was considered conclusive to identify the strain.
According to earlier studies, the threshold higher than 97.0% was used for *groEL* and *rpoB* genes, while the threshold higher than 99.0% was used for 16S rRNA, when deciding if it could be a new species (Adékambi et al., 2009; Glazunova et al., 2009; Janda and Abbott, 2007). Moreover, the sequence should be different from the sequences of all other species. The bioinformatic tool BioNumerics was used to compare clinical and reference strains, and phylogenetic trees were created (Appendix 4).

3. Results

3.1 Cultivation of strains

The cultivation of bacterial strains started with 33 strains, but only 31 showed satisfactory growth (Figure 5), and were used in this project. Two strains (CCUG 11772 and CCUG 33349) did not grow well and were excluded. The reactions of Gram-positive strains on Chocolate Agar have been correlated to the strain identity, where yellow colour of the bacterial colonies has been associated with alpha-haemolytic species of genus *Streptococcus* (Gunn, 1984).

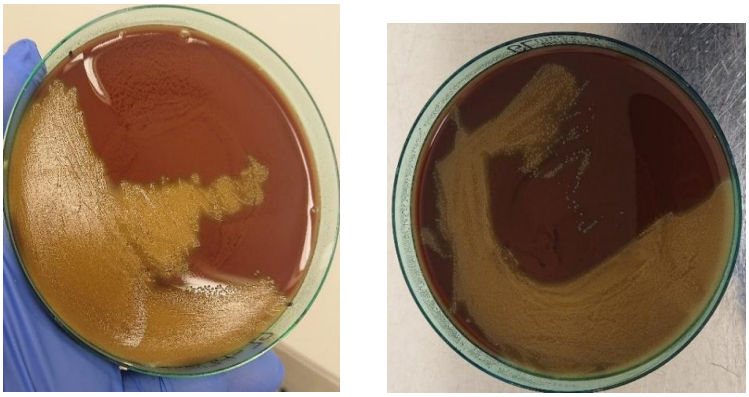

Figure 5. Strains of SMG cultivated on chocolate agar plates.

3.2 DNA extraction and PCR results.

The performed PCRs with positive and negative results for the three housekeeping genes 16S rRNA, *groEL* and *rpoB*. The PCR for 16S rRNA gene started with 21 strains and none of those samples showed PCR reaction yields to detect amplicons when running in gel electrophoresis. The mother solution containing the extracted DNA was diluted at 1:10 according to the protocol, but there was a suspicion about too high amount of DNA in the samples. A NanoDrop instrument was used to evaluate the quality of the DNA. A Qubit Broad Range was used to quantify the DNA. The analysis showed DNA concentrations between 274.6-1210.3 ng/µl, which was around 5-55 times higher than the recommended 20 ng/µl.

To troubleshoot this problem and optimize the PCR, four samples with different concentrations of DNA were performed in the new PCR for 16S rRNA (Figure 6). The adjustments were made by changing the size of the inoculation loops, lowering the amount of biomass. Continued control of the amount of DNA in the samples by using Nanodrop, allowed dilution of the samples and to take the necessary amount of DNA to perform the PCR.

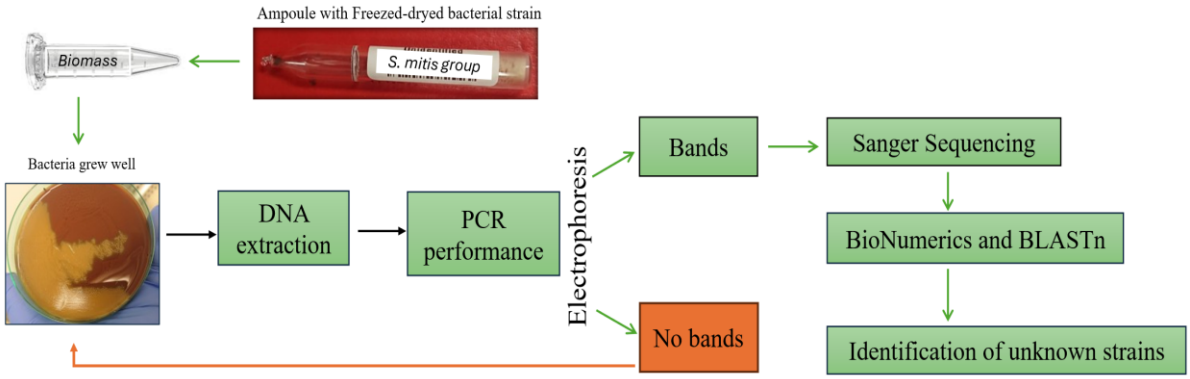

Figure 6. Workflow from the strain's cultivation until repeated PCR templating or sending to Sanger sequencing.

The concentrations of the DNA after measuring it on the Nanodrop were between 6.0 -101.4 ng/µl and it resulted in satisfactory bands on the gel electrophoresis (Figure 7).

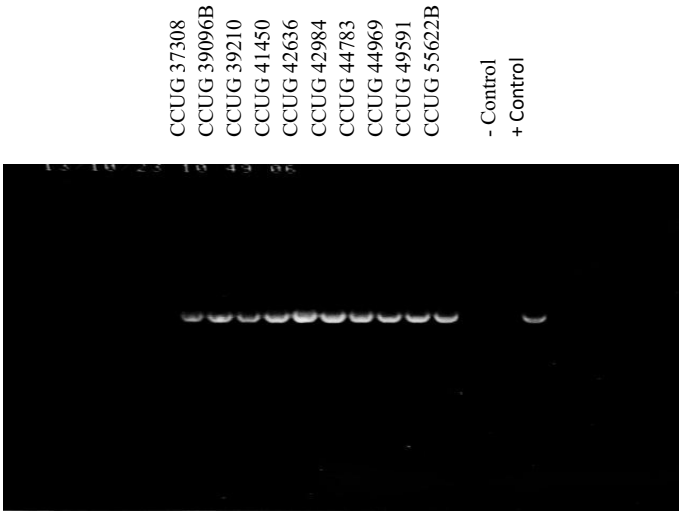

Figure 7. Positive PCR with positive control to 16S rRNA gene.

Even though DNA concentrations were measured with Nanodrop machines, the PCR performance was challenging with different DNA concentrations, and several PCRs were required . The appendix 2, tables 1,2 and 3 shows the performed PCR with the concentration's adjustment when trying to get bands. Some samples required dilutions while others did not. The DNA concentrations in the samples for the gene *groEL* were between 54.8 – 290.8 ng/µl. PCR annealing temperature conditions were also adjusted, and it was decreased from 55˚ Celsius degrees to 53˚ Celsius degrees (Figure 8 and Figure 9). After temperature adjustment, it provided better PCR amplicons of *groEL* genes producing satisfactory sequences in Sanger sequencing.

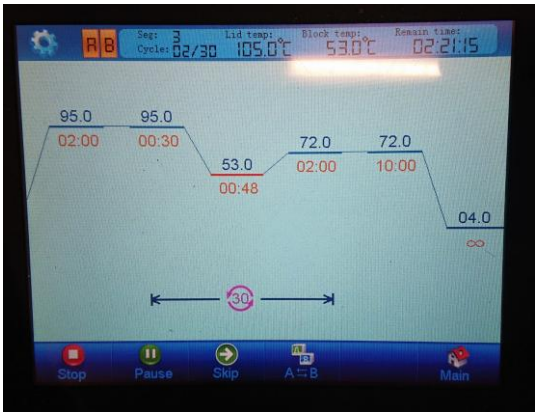

Figure 8. PCR thermal program to the *groEL* gene.

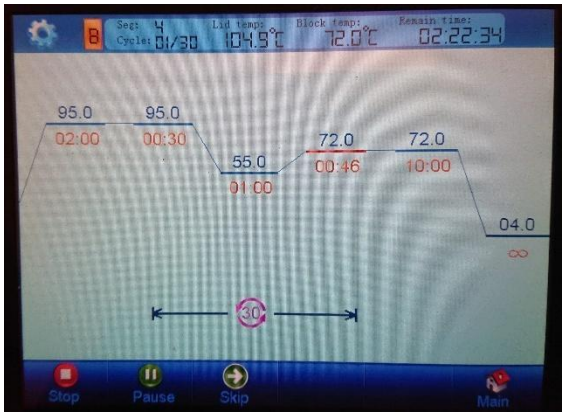

Figure 9. PCR thermal program to the *16S rRNA* and *rpoB* genes.

The gel electrophoresis showed bands with different strengths and the strains corresponding to too weak bands or absence bands, required new PCR performance. (Figure 10 and Figure 11).

The *rpoB* gene was easier to work with and the DNA concentrations measured in Nanodrop machines were between 86.7 – 290.0 ng/µl.

Figure 10. Negative PCR with positive control of 16S rRNA gene.

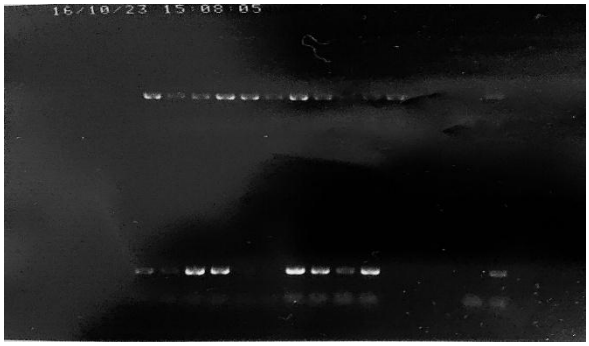

Figure 11. PCR results on gel 11+ 10 strains to *Groel* Gene.

PCR performance of *rpoB* started with twelve strains and the samples we diluted according to DNA concentrations measured with Nanodrop machines (Appendix 2, Table 3,). There was one strain performed for *rpoB*, CCUG 41450, where the PCR amplicon was positive, but it didn't get sequences with good quality by Sanger sequencing. It became impossible to make its analysis and identification. In total, sequences of two strains were disregarded, to *groEL* for CCUG 33514 and *rpoB* for CCUG 41450 (Figure 12).

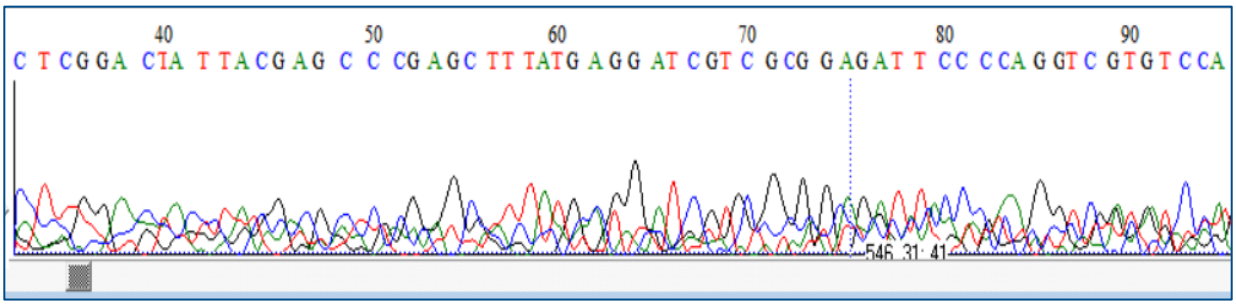

Figure 12. Chromatogram of Sanger sequencing to *rpoB* gene amplification of strain CCUG 41450 (positions 31-95).

On the other hand, the Sanger sequencing of the strain CCUG 41450 using 16S rRNA gene, got a sequence with good quality, which was possible to make the analysis and identification (Figure 13).

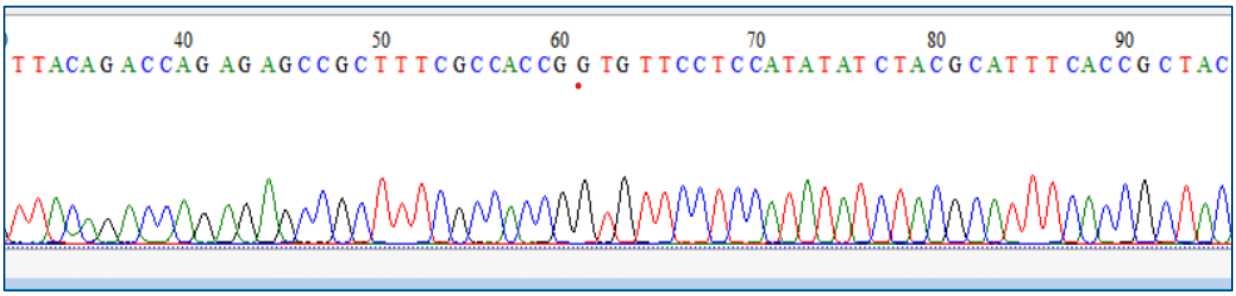

Figure 13. Chromatogram of Sanger sequencing to16S rRNA gene amplification of strain CCUG 41450 (positions 31-95).

3.3 Sanger sequencing, BioNumerics and BLASTn

The results of Sanger sequencing for the three housekeeping genes showed 31, out of 31 strains, with high quality of the 16S rRNA gene and 30, out of 31 strains, with high quality of the *groEL* and *rpoB* genes. After quality control, all sequences were assembled and the reads with low quality were trimmed, using BioNumerics.

The sequences were submitted to the BLASTn program to compare the nucleotides sequences to sequences available in GenBank. The matching type strains found in GenBank were verified to be published type strains in the online database List of Prokaryotic names with Standing in Nomenclature (LPSN) (https://lpsn.dsmz.de/text/introduction).

The values for the three best matches were chosen from BLASTn and the level of similarity by percent identity was recorded. For a strain to be identified as a certain species, the level of similarity had to be higher than 99.0% (16S rRNA) and 97.0% (*rpoB* and *groEL*). In addition, the similarities to all other species needed to be approximately 1% lower, than the best matching species.

The comparison in BLASTn of 16S rRNA gene for 29 strains showed similarities to several species higher than 99.0% (Appendix 3, table 2). Only two strains, CCUG 39210 and CCUG 42984 showed similarity lower than 99.0%.

Multi-locus sequence analysis (MLSA) of a set of housekeeping genes was used to improve the bacterial identification at the species level. MLSA has earlier been used to improve the species identification of the genus *Streptococcus* (Glazunova *et al*., 2009; Lal *et al*., 2011). In this study the genes *groEl* and *rpoB* were used.

The alignment in BLASTn results to *groEL* gene, showed nine different species as the best match (Appendix 3, table 2), but only three strains exhibited similarity ranges higher than 97.0% to the best hit. The sequence's similarities to the *groEL* gene were quite low ranging from 91.0% – 100.0%. The *groEL* gene is considered one of the best tools when identifying several bacterial species, subspecies and when performing phylogenetic analysis (Glazunova et al., 2009; Lal et al., 2011).

For BLASTn results to *rpoB* gene sequences, fourteen different species were identified (Appendix 3, table 3), A total of twenty-three strains had a best hit similarity range higher than 97.0%, while seven strains had a similarity range lower than 97.0%, but higher than 94.0%. In earlier studies, *rpoB* gene has been used together with other housekeeping genes when identifying bacterial strains at the species level (Glazunova et al., 2009; Bivand et al., 2024).

The high mutation rates occurred in *rpoB* gene through the years contribute to better discrimination one species from the other (Bivand et al., 2024).

3.4 Comparisons of the 16S rRNA gene, *groEL* and *rpoB BLAST results*

Comparative sequence analysis, using housekeeping genes, reveal the evolutionary events that have occurred in bacterial strains, and it gets significant information when using them as marker molecules to identify new bacterial strains. In this study, it was possible to identify some of the strains, using the three genes: 16S rRNA, *groEL* and *rpoB* (appendix 3 - Table 1). The strains CCUG 25812 and CCUG 32331 were identified as "*S. nidrosiense"* and respective "*S. infantis"* species, using all three genes. The results also showed identification of the strains CCUG 32132 as *S. nakanonensis* for the genes 16S rRNA and *groEL,* while the highest match, using the *rpoB* gene identified it as *S. oralis.*

Analysing the second-best hit also showed very close similarities- range when comparing with the first best hit. The small differences between ranges of the first- and second-best hits make it difficult to decide definitive identifications (Kilian et al.,2008) and it happened in this study. Using the *groEl* gene, there were a total of eighteen strains identified as *S. mitis,* apart from it, seven strains were also identified using 16S rRNA. When looking to the second best hit to 16S rRNA, there were four strains (CCUG 28754, CCUG 33690, CCUG 42636, CCUG 44763) that could be also identified as *S. mitis* because the similarity ranges between the first and the second values were very close and higher than 99.5%. Those four strains were identified as "*S. symci"* (first best hit), a species that is not yet validly published, but Killian et al. (2025) in a recent phylogenetic analysis research, found that it belongs to the same cluster as the species *S. mitis*. An extended comparison

including the strains results of the *rpoB,* only one strain was identified as *S. mitis*, which may represent that some strains are novel species that until now, were not identified.

The *rpoB* gene*,* is considered more discriminative when identifying bacterial strains in species level, presenting high mutation rate, such as preventing sequence heterogeneity, while most bacteria contain multiple copies of 16S rRNA, decreasing its resolution and interfering with the diagnostic routines when analysing the clinical bacterial strains (Bivand et al., 2024).

The results of the *rpoB*, showed that 7 strains were less than 97.0% similar to another species, which can represent novel species, thereby the recognition of novel bacterial species requires similarities range lower than 97% (Glazunova et al., 2009). Those strains identified with lower range in this study to rpoB are: *S. parasanguinis, S. oralis," S.vulneris", S. infantis, S. nakanonensis and S. hohhotensis.* The other identified strains with similarities ranged higher that 97% are represented by *S. toyakuensis, S. chosunensis, S. mitis, S. ilei, S. infantis, S. nidrosiensis, S. nakanonensis, S. koreensis, S. dentalis*, *S. pneumoniae and Streptococcus hohhotensis*. The genetic similarities of SMG contribute to most lineages in the group form clusters presenting small and limited distinction between them (Kilian et al., 2025).

The genetic diversity in the studied strains agreed with the taxonomy complexity described in the literature when identifying strains of SMG.

3.5 Identification of the strains as SMG at species level

The comparison of the sequences of the housekeeping genes 16S rRNA*, groEL and rpoB*, using BioNumerics and BLASTn provides results with higher accuracy than with the method used when the strains were received and archived in the CCUG collection. The results with highest identity scores and how close the first and second hits were in the alignment in BLASTn in all three genes was considered when suggesting an identification to each strain. According to the results, 23 strains are novel species; species which yet have not been described.

The 16S rRNA gene to CCUG 19074 was more than 99.0% similar to several species. Neither the *rpoB* nor the *groEL* sequences had high similarity (greater than 97.0%) to any known species. Based on this, this strain probably belongs to a novel species.

The 16S rRNA gene to CCUG 25812 strain also had a similarity higher than 99.0% to several species. The *rpoB* sequence was more than 97.0% similar to "*S. nidrosiense*" species, and the second hit was not close, and the same species was also the best hit for *groEL*, although with a lower similarity. This led to the identification as "*S. nidrosiense*"

The 16S rRNA gene to CCUG 26924 is more than 99.0% similar to several species. The *rpoB* sequence is 97.7% similar to the first hit, S. ilei, and the values between the first and second hits were not close to each other. The *groEL* sequence was less than 97.0% similar to several species, with values very close between the first and second-best hits. Based on this analysis, this strain possibly belongs to the species *S. ilei*.

The 16S rRNA gene to CCUG 27740 strain was more than 99.0% similar to several species. The *rpoB* sequence was more than 97.0% similar to several species, with values very close to the first- and second-best hits. The *groEL* sequence was less than 97.0% similar, with values very close to the first- and second-best hits. Based on it, this strain probably belongs to a novel species. The same was seen to the following strains: CCUG 28754, CCUG 31489, CCUG 31557, CCUG 35276, CCUG 35278, CCUG 35580, CCUG 33690, CCUG 36755, CCUG 37308, CCUG 42636, CCUG 44969 and CCUG 49591 strains.

The 16S rRNA gene to CCUG 27741 was more than 99% similar to several species. Further the *rpoB* sequence was more than 97.0% similar to several species and the values of the first- and second-best hits were very close, while the *groEL* sequence was more than 97.0% similar to *S.*

mitis, the first best hit and between the first- and the second-best hits the similarity values were not close. This strain belongs possibly to the species *S. mitis*.

The 16S rRNA gene to CCUG 32108 strain was more than 100.0% similar to several species and the values between the first and second hits were not close to each other. Both *groEL* and *rpoB* presented more than 97% similarities to "*S. lingualis*". Further, the values between the first- and second-best hits were not so close to each other. Based on this assumption, it is probably the species *S. lingualis*.

The 16S rRNA gene to CCUG 32132 strain presented more than 99.0% similarity to several species, while both *groEL* and *rpoB* sequences were less than 97.0% similar to several species, considering this strain probably belongs to a novel species. The same validation was done to CCUG 32601, CCUG 36753 and CCUG 55622B strains.

The 16S rRNA gene to CCUG 32331 strain was 100.0% identical to first hit and the values between the first and second hit were not close to each other. Further, *the rpoB* sequence was more than 97.0% similar to the first hit and the values to the first- and second-best hits were not close to each other, while the *groEL* sequence was less than 97.0% similar to the first best hit and between the first- and the second-best hits the similarity values were not close. Based on the analysis above, added to the results where the first hits to both *groEl* and *rpoB* sequences showed to belong to the species *S. infantis*, this strain belongs possibly to the species *S. infantis.*

The 16S rRNA gene to CCUG 32466 strain was more than 99.0% similar to several species. The *rpoB* sequence was more than 97.0% also similar to several species and the values between the first and second hits were very close. The *groEL* sequence was less than 97.0% similar and the values were not close to each other, which made difficult to define a possible species to this strain. This strain probably belongs to a novel species.

The 16S rRNA gene to CCUG 33514 strain presented more than 99.0% similarities to several species. The *rpoB* sequence was more than 97.0% also similar to several species and the values between the first and second hits were very close to each other, leading to consider it as a novel species. The Sanger sequencing of this strain to the *groEL* sequence didn't produce a good quality sequence.

The 16S rRNA gene to CCUG 36639 strain was more than 99.0% similar to several species and the values between the first- and second-best hits were very close. The *rpoB* sequence was more than 97.0% similar to the first hit and the values to the second hit were not close. The *groEL* sequence was less than 97.0% similarity to the first hit and the values between the first and second hits were not close to each other. The "*S. nidrosiense*" came up in both *groEL* (1st hit) and *rpoB* (2nd hit) and a comparison between those genes were done, but the values were lower than 97.0% (93.2%) to "*S. nidrosiensis*", leading to consider it as novel species.

The 16S rRNA gene to CCUG 39096B strain was more than 99.0% similar to several species and the values to the first- and second-best hits were very close. The *rpoB* sequence was more than 97.0% similar to *S. chosunense* at the first hit and the values between it and to the second hit were not close. The *groEL* sequence was more than 97.0% similarity to the first hit and the values between the first and second hits were very close to each other, leading to consider it possibly to belong to the species *S. chosunense*.

The 16S rRNA gene of CCUG 39210 was less than 99.0% similar to the species *Granulicatella adiacens*. This species doesn't belong to the SMG and could be due to contamination during the procedures. The quality of the sequence was not good, and it was possible to see that part of the sequence showing double tops (Figure 14). The *rpoB* and *groEL* sequences were not like *G. adiacens*. *RpoB* was less than 97% also similar to any species and the values between the first and second hits were very close. The *groEL* gene was more than 97.0% similar

and the values also were very close to each other, which made difficult to define a possible species to this strain, which possibly could be a novel species.

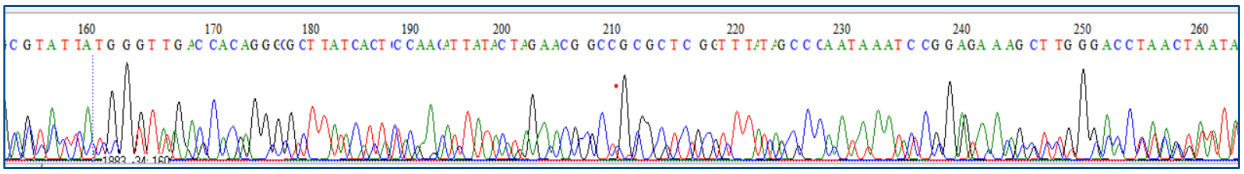

Figure 14. Chromatogram to CCUG 39210 to 16S rRNA gene showing contamination in the form of double peaks.

The 16S rRNA gene to CCUG 41450 strain was more than 99.0% similar to the first hit *S. cristatus*, and the values between the first- and second-best hits were not close to each other. The *groEL* sequence was 95.5% similar to the first hit *S. cristatus* and the values between the first- and second-best hits were not close. Considering that the first hit to both 16SrRNA and *groEl* sequences was found in the species *S. cristatus*; this strain possibly belongs to the species *S. cristatus*. Sanger sequencing of *rpoB* gene sequence to this strain didn't produce sequence with good quality, and it could therefore not be included in the analysis.

The 16S rRNA gene to CCUG 42984 presented less than 99.0% similarity to any species (the best match was to *S. infantis* with 98,2%). The *rpoB* and *groEL* sequences were less than 97.0% similar to any species, which probably makes this strain a novel species.

The 16S rRNA gene to CCUG 44763 is more than 99.0% similar to several species. The *rpoB* sequence is 100.0% similar to the first hit *S. toyakuensis*, and the values between the first and second hits were not close to each other. The *groEL* sequence was less than 97.0% similar to several species, with values very close between the first and second-best hits. Based on this analysis, this strain possibly belongs to the species *S. toyakuensis*.

3.6 Phylogenetic tree

The phylogenetic tree to the gene 16S rRNA, *groEL* and *rpoB* was done using BioNumerics. (Appendix 4 - Figures 1, 2 and 3). A part of the species *Granulicatella adiacens*, that does not belong to SMG, the phylogenetic tree to the gene 16S rRNA, showed small clusters with several strains with very similar sizes, confirming their genetical closely related relationship described in earlier studies (Jensen et al, 2016; Gevers et al., 2005; Keller et al., 2013).

Lower in the tree is possible to observe the species *"S. shenyangsis", S. australis* and *S. fermentans* distinguishing of the rest of the species and presenting percent similarity lower than 99.0%. Lower in the tree it is shown twelve species with very high and close identity rates: *S. oralis, S. toyakuensis*, *S. infantis, S. cristatus, "S. lingualis", S. nakanonensis*, "*S. nidrosiense", S. pneumoniae," S. symci", S. hohhotensis and S. mitis* clustering in the same clade with" *S. bouchesdurhonensis*". Lower in the tree it is also possible to see species clustered alone: "*S. lingualis2, S. cristatus, S. fermentans, "S. shenyangsis", S. infantis.*

Table 4. Results of phylogenetic tree to the gene 16S rRNA.

CCUG code	Identification	CCUG code	Identification	CCUG code	Identification
CCUG 19074	*Streptococcus sp*	CCUG 32331	*S. infantis*	CCUG 36755	*Streptococcus sp*
CCUG 25812	*Streptococcus sp*	CCUG 32466	*Streptococcus sp*	CCUG 37308	*Streptococcus sp*
CCUG 26924	*Streptococcus sp*	CCUG 32601	*"S.shenyangsis"*	CCUG 39096B	*Streptococcus sp*

Table 4. Continued. Results of phylogenetic tree to the gene 16S rRNA.

CCUG code	Identification	CCUG code	Identification	CCUG code	Identification
CCUG 27740	*S. mitis*	CCUG 33514	*Streptococcus sp*	CCUG 39210	*Granulicatella adiacens*
CCUG 27741	*S. mitis*	CCUG 33690	*Streptococcus sp*	CCUG 41450	*S. cristatus*
CCUG 28754	*Streptococcus sp*	CCUG 35276	*S.nakanonensis*	CCUG 42636	*Streptococcus sp*
CCUG 31489	*S.hohhotensis*	CCUG 35278	*Streptococcus sp*	CCUG 42984	*Streptococcus sp*
CCUG 31557	*S. mitis*	CCUG 35580	*Streptococcus sp*	CCUG 44763	*Streptococcus sp*
CCUG 32108	*"S. lingualis"*	CCUG 36639	*Streptococcus sp*	CCUG 44969	*Streptococcus sp*
CCUG 32132	*S. oralis*	CCUG 36753	*"S. nidrosiense"*	CCUG 49591	*S. mitis*
				CCUG 55622B	*Streptococcus sp*

The phylogenetic tree of gene *groEL* showed three species in individual clusters represented by *S. mitis, S. dentalis and "S. lingualis"* (Table 5).

Table 5. Results of phylogenetic tree to the gene *groEL*

CCUG code	Identification	CCUG code	Identification	CCUG code	Identification
CCUG 19074	*Streptococcus sp*	CCUG 32331	*Streptococcus sp*	CCUG 36755	*Streptococcus sp*
CCUG 25812	*Streptococcus sp*	CCUG 32466	*Streptococcus sp*	CCUG 37308	*Streptococcus sp*
CCUG 26924	*Streptococcus sp*	CCUG 32601	*Streptococcus sp*	CCUG 39096B	*S. dentalis*
CCUG 27740	*Streptococcus sp*	CCUG 33514	----------	CCUG 39210	*Streptococcus sp*
CCUG 27741	*S. mitis*	CCUG 33690	*Streptococcus sp*	CCUG 41450	*Streptococcus sp*
CCUG 28754	*Streptococcus sp*	CCUG 35276	*Streptococcus sp*	CCUG 42636	*Streptococcus sp*
CCUG 31489	*Streptococcus sp*	CCUG 35278	*Streptococcus sp*	CCUG 42984	*Streptococcus sp*
CCUG 31557	*Streptococcus sp*	CCUG 35580	*Streptococcus sp*	CCUG 44763	*Streptococcus sp*
CCUG 32108	*"S. lingualis"*	CCUG 36639	*Streptococcus sp*	CCUG 44969	*Streptococcus sp*
CCUG 32132	*Streptococcus sp*	CCUG 36753	*Streptococcus sp*	CCUG 49591	*Streptococcus sp*
				CCUG 55622B	*Streptococcus sp*

The phylogenetic tree to gene *rpoB* was quite similar with *groEL*'s phylogenetic tree. It is divided in small clusters. The tree represents seven species with similarities higher than 97.0%: *S. nakanonensis, S. toyakuensis, "S. nidrosiensis", S. infantis, S. ilei, S. chosunense,* and *S. dentalis* (Table 6 and Appendix 4-Figure 3).

Table 6. Results of phylogenetic tree to the gene *rpoB*

CCUG code	**Identification**	**CCUG code**	**Identification**	**CCUG code**	**Identification**
CCUG 19074	*Streptococcus sp*	**CCUG 32331**	*Streptococcus sp*	**CCUG 36755**	*Streptococcus sp*
CCUG 25812	*"S. nidrosiensis"*	**CCUG 32466**	*Streptococcus sp*	**CCUG 37308**	*Streptococcus sp*
CCUG 26924	*S. ilei*	**CCUG 32601**	*Streptococcus sp*	**CCUG 39096B**	*S. chosunense*
CCUG 27740	*Streptococcus sp*	**CCUG 33514**	----------	**CCUG 39210**	*Streptococcus sp*
CCUG 27741	*Streptococcus sp*	**CCUG 33690**	*Streptococcus sp*	**CCUG 41450**	*Streptococcus sp*
CCUG 28754	*Streptococcus sp*	**CCUG 35276**	*S. nakononensis*	**CCUG 42636**	*Streptococcus sp*
CCUG 31489	*Streptococcus sp*	**CCUG 35278**	*Streptococcus sp*	**CCUG 42984**	*Streptococcus sp*
CCUG 31557	*Streptococcus sp*	**CCUG 35580**	*Streptococcus sp*	**CCUG 44763**	*S. toyakuensis*
CCUG 32108	*S. dentalis*	**CCUG 36639**	*Streptococcus sp*	**CCUG 44969**	*Streptococcus sp*
CCUG 32132	*S. infantis*	**CCUG 36753**	*Streptococcus sp*	**CCUG 49591**	*Streptococcus sp*
				CCUG 55622B	*Streptococcus sp*

Thus, in total, we were able to identify eight strains to the species level, one using only 16S gene sequencing (CCUG 41450) and one also using *rpoB* (CCUG 26924) or *groEl* (CCUG 27741). Five strains were possible to identify by comparisons between the tree genes, taking in account the analysis of the best hit in each gene. A total of 23 strains remain unidentified and probably belong to not yet described species (table7).

In the Department of Infectious Diseases another ongoing project in parallel with this study to identify species of SMG, based on published whole genome sequences from other studies. The genomes were compared, using dDNA-DNA hybridization and ANI. The sequences used in this study were compared with the sequence database from this project, and in total eight of the unidentified strains, were identified as the same novel species that were identified, using ANI and dDDH: CCUG 31557, CCUG 44763, CCUG 31489, CCUG 33514, CCUG 32601, CCUG 55622B (similarity using *rpoB*). While CCUG 27741, CCUG 44763, CCUG 31489, CCUG 31557, CCUG 35580, CCUG 36201, CCUG 33514 and 55622B with similarity using *groEl* gene sequence**.** In the phylogenetic trees these strains are seen as "GV genomespecies" to gene rpoB, while "GVgenomes" to gene *groEL.*

Table 7. Final identification of 31 clinically relevant Strains of SMG

CCUG code	Identification	CCUG code	Identification	CCUG code	Identification
CCUG 19074	*Streptococcus sp*	CCUG 32331	*S. infantis*	CCUG 36755	*Streptococcus sp*
CCUG 25812	*S. nidrosiens*	CCUG 32466	*Streptococcus sp*	CCUG 37308	*Streptococcus sp*
CCUG 26924	*S. ilei*	CCUG 32601	*Streptococcus sp*	CCUG 39096B	*S. chosunense*
CCUG 27740	*Streptococcus sp*	CCUG 33514	*Streptococcus sp*	CCUG 39210	*Streptococcus sp*
CCUG 27741	*S. mitis*	CCUG 33690	*Streptococcus sp*	CCUG 41450	*S. cristatus*
CCUG 28754	*Streptococcus sp*	CCUG 35276	*Streptococcus sp*	CCUG 42636	*Streptococcus sp*
CCUG 31489	*Streptococcus sp*	CCUG 35278	*Streptococcus sp*	CCUG 42984	*Streptococcus sp*
CCUG 31557	*Streptococcus sp*	CCUG 35580	*Streptococcus sp*	CCUG 44763	*S. toyakuensis*
CCUG 32108	*S. lingualis*	CCUG 36639	*Streptococcus sp*	CCUG 44969	*Streptococcus sp*
CCUG 32132	*Streptococcus sp*	CCUG 36753	*Streptococcus sp*	CCUG 49591	*Streptococcus sp*
				CCUG 55622B	*Streptococcus sp*

4. Discussion

The identification of the bacterial strains in this study used a Multi-Locus Sequence Analysis (MLSA) scheme designed with a set of three housekeeping genes: 16S rRNA, *groEL* and *rpoB*. It was set up on 31 strains, archived at CCUG and previously classified as *S. mitis* or *S. mitis* complex or *S. mitis* group.

The cut-off values used during the species level sequence identification, was 99.0% for the 16S rRNA gene, while for *groEL* and *rpoB* it was 97.0% (Glazunova et al., 2009). It means, the cut-off values are gene-dependent and that the gene's evolution rate has an influence on the values. It is known that the 16S rRNA contains more conserved regions when compared to the genes *groEL* and *rpoB*. Janda et al. (2007) suggested a cut-off higher than 99.0% for the 16S rRNA gene sequence similarity for species identification, while Adékambi et al. (2003) suggested that the cut-off value for *rpoB* gene sequence similarity should be decided after analysis of the size of sequence fragment.

Kilian et al. (2025) described in earlier research that some bacterial strains of clinical relevance such as *S. pneumoniae, S. pseudopneumoniae* and *S. mitis* were difficult to distinguish, using analysis based only on 16S rRNA gene, which is in accordance with this study, where eight strains to the 16S rRNA gene were more than 99.0% similar to the species *S. mitis* as first best hit and at the same time these strains were also more than 99.0% similar to other species such as *"S. bouchesdurhonensis"* (5 strains), *S. gwangjuensis* (1 strain) and *"S. symci"* (2 strains).

Further, there were 20 more strains presenting both first- and second-best hits with more than 99% similarities with several species. It is due to the structure of the 16S rRNA gene, which is highly conserved, with highly similar sequences, making it impossible to distinguish species within SMG with high accuracy (Glaeser and Kämpfer, 2015; Janda and Abbott, 2007). The 16S rRNA high conserved regions is explained by the position of it into the 30S ribosomal subunit, an essential

region, where proteins are synthesized (Noller & Gutell, 2022; Janda & Abbott, 2007). There are many advantages in using a set of housekeeping genes, such as *rpoB* and *groEL*. The *rpoB* can provide greater sequence divergence, which means that it makes the species-level identification more discriminative than the 16S rRNA gene does (Adékambi et al., 2009; Drancourt et al., 2004; Imai et al., 2020). In one study developed by Drancourt et al. (2004) the discriminative capacity of the *rpoB* gene was shown for the *Streptococcus* species.

MLSA represents a more accurate, reliable, and reproducible method to identify bacterial strains of SMG compared to 16S rRNA gene sequencing alone (Keller et al., 2013; Imai et al., 2020). MLSA has higher resolution at the species level, because it comprises several loci that accurately reflect phylogenetic relationships of closely related bacterial strains, such as SMG (Jensen et al., 2016; Gevers et al., 2005; Keller et al., 2013).

If the MLSA does not work as expected, the whole-genome sequence-based tools such as core-genome clustering, nanopore sequencing plus tools like Kraken2 or WIMP (What′s in? my Pot - bioinformatic tool to analyse data from MinIon device), or ANI (Average Nucleotide Identity) analyses can be used to determine the bacterial strains identification. (Sadowy & Hryniewicz, 2020; Imai et al., 2020). Facklam et al. (2002) for example, showed in their study the difficulties in distinguishing species of *S. mitis* and *S. pneumoniae,* due to them having very similar 16S rRNA gene sequences. The nine hypervariable regions (V1-V9) of 16S rRNA gene are useful to classify at the genus or family level, but unfortunately it lacks the needed resolution for distinguishing closely related species within SMG (Tindall et al., 2010), having significant implications in microbial taxonomy, once that the high similarities of 16S rRNA gene sequences cannot be used for taxonomy or diagnostics approaches (Imai et al., 2020; Kilian et al., 2008). Kilian et al. (2008) also found that the variable regions in 16S rRNA gene had no distinctions in closely related species like *S. pseudopneumoniae, S. mitis,* and *S. pneumoniae,* where the variables regions in 16S rRNA gene, have highly conserved sequences that contribute to decreasing the discriminatory power at the level species within SMG.

To analyse homology within a bacterial genus or family using only 16S rRNA gene in studies of clinical microbiota and even environmental purposes is not recommended because of the risks to incorrect and misleading results (Větrovský & Baldrian, 2013). The 16S rRNA gene conserved regions are observed in all bacteria and generally it is used for universal primers in PCR amplification (Weisburg *et al*., 1991).

The evolutionary history of the housekeeping genes comprises different rates and it can reflect on the results of this study where the hits found to 16S rRNA, *groEL* and *rpoB* genes showed to be different in several strains. Earlier studies have shown that sequence comparisons of genes into *Streptococcus* mitis revealed horizontal gene transfer (HGT) events into this group, which can explain the high similarity between the species and the different match to different genes (Hakenbeck et al., 2001; Salvadori et al., 2019; Kilian et al., 2008). The hypervariable regions of 16S rRNA gene do not contribute to distinguish closely related species of SMG. Its highly conserved regions make 16S rRNA unappropriated to identify SMG species at level species (Glazunova et al., 2009; Kilian et al., 2009; Jensen et al., 2016). According to Stackebrandt et al. (2002), the gene 16S rRNA gene can be used to identify and recognize novel species at the genus level, only if the similarity value between the known and the novel species is higher or equal 97% (Stackebrandt et al., 2002).

The *groEL* gene can be more affected by recombination during evolutionary events, when compared to 16S rRNA and *rpoB* genes and it can lead to more diversity than it is known to (Król et al, 2021; Glazunova et al., 2009). On the other hand, the *rpoB* gene has lower recombination rates than *groEL* and it may explain why it was found to have different matches compared to *groEL*.

Moreover, the *rpoB* gene is more stable marker to use when identifying SMG species at species level. At the same time, it is known that some species, such as *S. pneumoniae* and *S. mitis* can cluster too close to each other, where the diversity revealed by *groEL* can be missed when using *rpoB* to identify SMG members (Bivand et al., 2024; Salvadori et al., 2019).

A study done by Kilian et al. (2025), using a whole-genome phylogeny-based method, was allowed to explore phylogenetic comparison between isolates of SMG, where it was able to find the same tendency as seen in the phylogenetics trees in this study. Kilian et al., 2025, conclude that *S. mitis* has later *heterotypic synonyms*, represented by the following species: *S. toyakuensis*, *S. chosunensis*, *S. gwangjuensis*, and *S. hohhotensis.* Further, they considered that the species "not validly published:" "*S. shenyangsis*", "*S. symci*" and "*S. vulneris*" fit in better in *S. mitis,* while suspecting that the species "*S. bouchesdurhonensis*" is a mix of two different species; *S. mitis* and *S. pseudopneumoniae* (Kilian et al., 2025). This is due to that a single reference sequence for each species was used, not considering the diversity of individual species (Kilian et al, 2025).

It gave substance to our results, considering that *S. mitis* and "*S. bouchesdurhonensis*" of gene 16S rRNA clustered in the same clade, being included in a group of eight closest species in the tree represented by *S. nakanonensis, S. hohhotensis, S. toyakuensis, S. pneumoniae, "S. symci" and "S. nidrosiensis"*. For the gene *groEL* it was also seen that *S. mitis* was not close to any identified species, but in the same cluster as *S. pneumoniae, S. pseudopneumoniae*, which confirm the close relatedness between these species.

As expected, the phylogenetic tree for the gene 16S rRNA showed low resolution and shorter distances between strains, becoming difficult to distinguish the species from each other, due to the highly conservative nature of it.

Some strains in this study have the nomenclature status at List of Prokaryotic named with standing in Nomenclature (LPSN) as "not validly published". There is a small quantity of "valid published" names of bacteria registered at the International Code of Nomenclature of Prokaryotes (ICNP) and it is common that the "non-validly published names" often show up in high-quality phylogenomic or molecular analyses when identifying bacterial strains (Wambbui et al., 2021; Oren, 2024). It means that several microorganisms are still uncultured, or their name has not been validly published and that it is allowed to use species with "not validly published" status to clarify the species limitations, where in some situations, can contribute to adjustment in important limiting values, that provide accurate identification of unknown species (Wambui et al., 2021).

5. Conclusion

Accurate identification of clinically relevant bacterial strains is crucial in clinical infection diagnostics, for epidemiological surveillance and for formulating guidelines aimed at reducing the global health burden caused by *Streptococcus* mitis group (SMG) infectious diseases. Advances in DNA sequencing and analysis methodologies, improve our understanding of SMG evolutionary dynamics, contribute to a more robust taxonomic framework and emphasize the necessity of exploring the functional capacities of these organisms to improve clinical management in both human and veterinary settings (Jensen et al., 2016).

It is universally recognized that correct diagnosis enables appropriate treatment, thereby reducing the emergence of antimicrobial resistance (AMR), once the well documented higher resistance profiles among SMG members compared to other Viridans group streptococci (VGS) is rising (Sadowy, 2020). Therefore, a rapid and reliable species-level identification of microorganisms is essential for increasing effective therapeutic decision-making and public health interventions.

This thesis demonstrates that a Multi-Locus Sequence Analysis (MLSA) approach, based on three conserved housekeeping genes, such as 16S rRNA, *groEL*, and *rpoB*, significantly improves the species-level resolution within *Streptococcus* mitis group. While 16S rRNA gene alone has limited discriminatory power due to minimal interspecies divergence (Glazunova et al., 2009), the combined use of *rpoB* and *groEL* improves resolution, where a minimum interspecies divergence can be actual for *rpoB* and for *groEL* (Glazunova et al., 2009). These findings align with earlier research demonstrating that single-gene approaches are not enough when identifying SMG species, while multi-locus approaches offer superior reliability (Imai et al., 2020; Jensen et al., 2016).

New protocols using WGS methods, considered to be most accurate to identify species of SMG, need to be developed, making it cheaper, faster and easier to implement in diagnosis laboratories, which certainly will save many lives worldwide.

Acknowledgements

I sincerely thank my supervisors, Dr. Edward R. B. Moore, Dr. Francisco Salvà Serra, and Dr. Liselott Svensson, for their guidance and valuable feedback throughout this project. I am grateful to the CCUG lab members—Susanne, Sofia, Christel, Elisabeth, Maria and Hanna-Sophia —and to the staff of the Department of Clinical Microbiology at Sahlgrenska University Hospital, especially Bea, Leonarda, Marianela, and Victor, for generously support and sharing their knowledge and for the warm welcome. My thanks also go to my study partner, Georgia Mesohoriti, for her support, positivity, and teamwork during the laboratory work, and to Catharina Olsson for all information and availability. Finally, I am deeply thankful to my husband Leif, my children Beatriz and Viktor, and my friends, especially Gisele Brandão and Carlos Gustavo Regis, for their patience, encouragement, and unwavering support.

6. References

Adékambi, T., Drancourt, M., & Raoult, D. (2009). The rpoB gene as a tool for clinical microbiologists. *Trends in microbiology*, *17*(1), 37–45. https://doi.org/10.1016/j.tim.2008.09.008

Anonymous (2021). Public Health England: UK Standards for Microbiology Investigations Identification of *Streptococcus* species. Enterococcus species and morphologically similar organisms, ID 4 | Issue number: 4 |1–31. Downloaded from https://www.rcpath.org/static/ce35b1b6-9d79-4125-a0f64ba28e9a3584/UK-SMI-ID-4i4-Identification-of-Streptococcus-species-Enterococcus-species-and-morphologically-similar-organisms-September-2021.pdf, 2025-04-12

Belman, S., Chaguza, C., Kumar, N., Lo, S., & Bentley, S. D. (2022). A new perspective on ancient Mitis group streptococcal genetics. *Microbial genomics*, *8*(2), 000753. https://doi.org/10.1099mgen.0.000753

Bishop, C. J., Aanensen, D. M., Jordan, G. E., Kilian, M., Hanage, W. P., & Spratt, B. G. (2009). Assigning strains to bacterial species via the internet. *BMC biology*, *7*, 3. https://doi.org/10.1186/1741-7007-7-3

Bivand, J. M., Dyrhovden, R., Sivertsen, A., Tellevik, M. G., Patel, R., & Kommedal, Ø. (2024). Broad-range amplification and sequencing of the *rpoB* gene: a novel assay for bacterial identification in clinical microbiology. *Journal of clinical microbiology*, *62*(7), e0026624. https://doi.org/10.1128/jcm.00266-24

Bloch, S., Hager-Mair, F. F., Andrukhov, O., & Schäffer, C. (2024). Oral streptococci: modulators of health and disease. Frontiers in cellular and infection microbiology, 14, 1357631. https://doi.org/10.3389/fcimb.2024.1357631

Chait, E., Page, G., & Hunkapiller, M. (1988). Battle of the DNA sequencers. *Nature*, *333*(6172), 477–478. https://doi.org/10.1038/333477a0

Chun, S., Huh, H. J., & Lee, N. Y. (2015). Species-specific difference in antimicrobial susceptibility among viridans group streptococci. *Annals of laboratory medicine*, *35*(2), 205–211. https://doi.org/10.3343/alm.2015.35.2.205

Crossley, B. M., Bai, J., Glaser, A., Maes, R., Porter, E., Killian, M. L., Clement, T., & Toohey-Kurth, K. (2020). Guidelines for Sanger sequencing and molecular assay monitoring. *Journal of veterinary diagnostic investigation : official publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *32*(6), 767–775. https://doi.org/10.1177/1040638720905833

Deibel, R.H., & Seeley, H.W. Jr. (1974). Family II. *Streptococcaceae*. In: R.E. Buchanan and N.E. Gibbons (eds.), *Bergey's Manual of Determinative Bacteriology*, (eighth edition., pp. 490-515). The Williams & Wilkins Co, Baltimore.

Drancourt, M., Roux, V., Fournier, P. E., & Raoult, D. (2004). rpoB gene sequence-based identification of aerobic Gram-positive cocci of the genera Streptococcus, Enterococcus, Gemella, Abiotrophia, and Granulicatella. *Journal of clinical microbiology*, *42*(2), 497–504. https://doi.org/10.1128/JCM.42.2.497-504.2004

Doern, C. D., & Burnham, C. A. (2010). It's not easy being green: the viridans group streptococci, with a focus on pediatric clinical manifestations. *Journal of clinical microbiology*, 4*8*(11), 3829–3835. https://doi.org/10.1128/JCM.01563-10

Facklam, R. (2002). What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. *Clinical Microbiology Reviews*, *15*(4), 613–630. https://doi.org/doi:10.1128/cmr.15.4.613-630.2002

Fredricks, D. N., & Relman, D. A. (1998). Improved amplification of microbial DNA from blood cultures by removal of the PCR inhibitor sodium polyanetholesulfonate. *Journal of clinical microbiology*, *36*(10), 2810–2816. https://doi.org/10.1128/JCM.36.10.2810-2816.1998

Gao, X. Y., Zhi, X. Y., Li, H. W., Klenk, H. P., & Li, W. J. (2014). Comparative genomics of the bacterial genus *Streptococcus illuminates* evolutionary implications of species groups. *PloS one*, *9*(6), e101229. https://doi.org/10.1371/journal.pone.0101229

Gevers, D., et al. (2005). Multilocus sequence typing of *Streptococcus* species: Re-evaluating the evolutionary history of *Streptococcus pneumoniae* and related species. *International Journal of Systematic and Evolutionary Microbiology*, 55(4), 1697-1703. https://doi.org/10.1099/ijs.0.63419-0

Glaeser, S. P., & Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology, 38*(4), 237–245. https://doi.org/10.1016/j.syapm.2015.03.007

Glazunova, O. O., Raoult, D., & Roux, V. (2009). Partial sequence comparison of the rpoB, sodA, groEL and gyrB genes within the genus *Streptococcus*. *International journal of systematic and evolutionary microbiology*, *59*(Pt 9), 2317–2322. https://doi.org/10.1099/ijs.0.005488-036).

Global Burden of Disease (GBD) 2021 Lower Respiratory Infections and Antimicrobial Resistance Collaborators (2024). Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990-2021: a systematic analysis from the Global Burden of Disease Study 2021. *The Lancet. Infectious diseases*, *24*(9), 974–1002. https://doi.org/10.1016/S1473-3099(24)00176-2

Global Burden of Diseases (GBD) 2019 Antimicrobial Resistance Collaborators (2022). Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet (London, England)*, *400*(10369), 2221–2248. https://doi.org/10.1016/S0140-6736(22)02185-7

Gonzales-Siles, L., Salvà-Serra, F., Degerman, A., Nordén, R., Lindh, M., Skovbjerg, S., & Moore, E. R. B. (2019). Identification and capsular serotype sequetyping of *Streptococcus pneumoniae* strains. *Journal of medical microbiology*, *68*(8), 1173–1188. https://doi.org/10.1099/jmm.0.001022

Gu, W., Miller, S., & Chiu, C. Y. (2019). Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. Annual review of pathology, 14, 319–338. https://doi.org/10.1146/annurev-pathmechdis-012418-012751

Gunn B. A. (1984). Chocolate agar, a differential medium for gram-positive cocci. *Journal of clinical microbiology*, *20*(4), 822–823. https://doi.org/10.1128/jcm.20.4.822-823.1984

Hakenbeck, R., Balmelle, N., Weber, B., Gardès, C., Keck, W., & Saizieu, A. d. (2001). Mosaic Genes and Mosaic Chromosomes: Intra- and Interspecies Genomic Variation of Streptococcus pneumoniae. Infection and Immunity, 69(4), 2477-2486. https://doi.org/doi:10.1128/iai.69.4.2477-2486.2001

Khatoon, H., Chavan, D., Anokhe, A., & Kalia, V. (2022). Catalase Test: A Biochemical Protocol for Bacterial Identification.

Hossain, Z. (2014). Bacteria: *Streptococcus.* In: Encyclopedia of Food Safety, Motarjemi, Y., Ed.; Academic Press: Waltham, MA, USA, Elsevier; pp. 535–545.

Imai, K., Nemoto, R., Kodana, M., Tarumoto, N., Sakai, J., Kawamura, T., Ikebuchi, K., Mitsutake, K., Murakami, T., Maesaki, S., Fujiwara, T., Hayakawa, S., Hoshino, T., Seki, M., & Maeda, T. (2020). Rapid and Accurate Species Identification of Mitis Group Streptococci Using the MinION Nanopore Sequencer. *Frontiers in cellular and infection microbiology*, *10*, 11. https://doi.org/10.3389/fcimb.2020.00011

Ishii N. (2017). GroEL and the GroEL-GroES Complex. *Sub-cellular biochemistry*, *83*, 483–504. https://doi.org/10.1007/978-3-319-46503-6_17

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, *45*(9), 2761–2764. https://doi.org/10.1128/JCM.01228-07

Jensen, A., Scholz, C. F. P., & Kilian, M. (2016). Re-evaluation of the taxonomy of the Mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. *International journal of systematic and evolutionary microbiology*, *66*(11), 4803–4820. https://doi.org/10.1099/ijsem.0.001433

Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J., & Lewis, N. E. (2022). What are housekeeping genes?. *PLoS computational biology*, *18*(7), e1010295. https://doi.org/10.1371/journal.pcbi.1010295

Kawamura, Y., Whiley, R. A., Shu, S. E., Ezaki, T., & Hardie, J. M. (1999). Genetic approaches to the identification of the mitis group within the genus *Streptococcus*. *Microbiology (Reading, England)*, *145 (Pt 9)*, 2605–2613. https://doi.org/10.1099/00221287-145-9-2605

Keller, S. L., & Cole, M. A. (2013). The utility of MLSA in bacterial taxonomy and identification. *International Journal of Systematic and Evolutionary Microbiology*, 63(3), 675-680. https://doi.org/10.1099/ijs.0.038030-0

Khehra N, Padda IS, Swift CJ. Polymerase Chain Reaction (PCR) [Updated 2023 Mar 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK589663/

Kielkopf, C. L., Bauer, W., & Urbatsch, I. L. (2021). Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis of Proteins. *Cold Spring Harbor protocols*, *2021*(12), 10.1101/pdb.prot102228. https://doi.org/10.1101/pdb.prot102228

Kilian, M., Poulsen, K., Blomqvist, T., Håvarstein, L. S., Bek-Thomsen, M., Tettelin, H., & Sørensen, U. B.S. (2008). Evolution of Streptococcus pneumoniae and its close commensal relatives. *PloS one*, *3*(7), e2683. https://doi.org/10.1371/journal.pone.0002683

Kilian, M., Slotved, H. C., Fuursted, K., D'Mello, A., & Tettelin, H. (2025). Re-evaluation of boundaries of *Streptococcus mitis* and *Streptococcus oralis* and demonstration of multiple later synonyms of *Streptococcus mitis, Streptococcus oralis* and *Streptococcus thalassemiae:* description of *Streptococcus mitis* subsp. *carlssonii* subsp. nov. and emended description of *Streptococcus mitis*. *International journal of systematic and evolutionary microbiology*, *75*(3), 10.1099/ijsem.0.006704. https://doi.org/10.1099/ijsem.0.006704

Kilian, M., & Tettelin, H. (2019). Identification of Virulence-Associated Properties by Comparative Genome Analysis of *Streptococcus pneumoniae, S. pseudopneumoniae, S. mitis*, Three S. oralis Subspecies, and *S. infantis*. *mBio*, *10*(5), e01985-19. https://doi.org/10.1128/mBio.01985-19

Kralik, P., & Ricchi, M. (2017). A Basic Guide to Real Time PCR in Microbial Diagnostics: Definitions, Parameters, and Everything. *Frontiers in microbiology*, *8*, 108. https://doi.org/10.3389/fmicb.2017.00108

Król, J., Nowakiewicz, A., Błaszków, A., Brodala, M., Domagała, A., Prassol, A. N., Sławska, D., & Wojtynia, J. (2022). Genetic diversity of oral streptococci in the guinea pig as assessed by sequence analysis of the 16S rRNA and groEL genes. *Folia microbiologica*, *67*(2), 311–318. https://doi.org/10.1007/s12223-021-00936-3

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*, 28. https://doi.org/10.1186/1476-0711-10-28

López-Aladid, R., Fernández-Barat, L., Alcaraz-Serrano, V., Bueno-Freire, L., Vázquez, N., Pastor-Ibáñez, R., Palomeque, A., Oscanoa, P., & Torres, A. (2023). Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Scientific reports*, *13*(1), 3974. https://doi.org/10.1038/s41598-023-30764-z

Ludwig, W., Schleifer, K.H., and Whitman, W.B. (2009). Order II. *Lactobacillalles* ord. Nov. In: P. De Vos, G.M. Garrity, D. Jones, N.R. Krieg, W. Ludwig, F.A. Rainey, K.H. Schleifer, and W.B. Whitman (eds.), *Bergey's Manual of Systematic Bacteriology,* (second edition., pp. 464). Springer.

McDevitt, E., Khan, F., Scasny, A., Thompson, C. D., Eichenbaum, Z., McDaniel, L. S., & Vidal, J. E. (2020). Hydrogen Peroxide Production by *Streptococcus* pneumoniae Results in Alpha-hemolysis by Oxidation of Oxy-hemoglobin to Met-hemoglobin. *mSphere*, *5*(6), e01117-20. https://doi.org/10.1128/mSphere.01117-20

Mitchell J. (2011). *Streptococcus mitis*: walking the line between commensalism and pathogenesis. *Molecular oral microbiology*, *26*(2), 89–98. https://doi.org/10.1111/j.2041-1014.2010.00601.x

Murray, P (2018). Bacteria. In: Basic medical microbiology. (University of Maryland School of Medicine, Baltimore Maryland. Philadelphia), 6th Edition, Philadelphia, Elsevier, pp 10-21.

Noller, H. F., Donohue, J. P., & Gutell, R. R. (2022). The universally conserved nucleotides of the small subunit ribosomal RNAs. *RNA (New York, N.Y.)*, *28*(5), 623–644. https://doi.org/10.1261/rna.079019.12

O'Brien, K. L., Wolfson, L. J., Watt, J. P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E., Mulholland, K., Levine, O. S., Cherian, T., & Hib and Pneumococcal Global Burden of Disease Study Team (2009). Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet (London, England)*, *374*(9693), 893–902. https://doi.org/10.1016/S0140-6736(09)61204-6

Obradovic, J., Jurisic, V., Tosic, N., Mrdjanovic, J., Perin, B., Pavlovic, S., & Djordjevic, N. (2013). Optimization of PCR conditions for amplification of GC-Rich EGFR promoter sequence. *Journal of clinical laboratory analysis*, *27*(6), 487–493. https://doi.org/10.1002/jcla.21632

Ogier, J. C., Pagès, S., Galan, M., Barret, M., & Gaudriault, S. (2019). rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC microbiology*, *19*(1), 171. https://doi.org/10.1186/s12866-019-1546-z

Oren, A. (2024). On validly published names, correct names, and changes in the nomenclature of phyla and genera of prokaryotes: a guide for the perplexed. npj Biofilms and Microbiomes, 10(1), 20. https://doi.org/10.1038/s41522-024-00494-9

Rentschler, S., Kaiser, L., & Deigner, H. P. (2021). Emerging Options for the Diagnosis of Bacterial Infections and the Characterization of Antimicrobial Resistance. *International journal of molecular sciences*, *22*(1), 456. https://doi.org/10.3390/ijms22010456

Sadowy, E., Bojarska, A., Kuch, A., Skoczyńska, A., Jolley, K. A., Maiden, M. C. J., van Tonder, A. J., Hammerschmidt, S., & Hryniewicz, W. (2020). Relationships among streptococci from the mitis group, misidentified as *Streptococcus pneumoniae*. *European journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology*, *39*(10), 1865–1878. https://doi.org/10.1007/s10096-020-03916-6

Sadowy, E., & Hryniewicz, W. (2020). Identification of *Streptococcus pneumoniae* and other Mitis streptococci: importance of molecular methods. *European journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology*, *39*(12), 2247–2256. https://doi.org/10.1007/s10096-020-03991-9

Salvadori, G., Junges, R., Morrison, D. A., & Petersen, F. C. (2019). Competence in *Streptococcus pneumoniae* and Close Commensal Relatives: Mechanisms and Implications. *Frontiers in cellular and infection microbiology*, *9*, 94. https://doi.org/10.3389/fcimb.2019.00094

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sanseverino, Isabella & Navarro, Anna & Loos, Robert & Marinov, Dimitar & Lettieri, Teresa. (2018). State of the Art on the Contribution of Water to Antimicrobial Resistance. 10.2760/771124.

Schleifer, K.H.(2009) Phylum XIII: Firmicutes Gibbons and Murray 1978, 5 (Firmicutes (sic) Gibbons and Murray 1978, 5). In: De Vos, P., Garrity, G.M., Jones, D., Kreig, N.R., Ludvig, W., Rainey, F.A., Schleifer, K-H., Whitman, W.B. Bergey's manual of systematic bacteriology: The Firmicutes. Vol 3. Springer, New York, U.S.

Sherman J. M. (1937). THE STREPTOCOCCI. *Bacteriological reviews*, *1*(1), 3–97. https://doi.org/10.1128/br.1.1.3-97.1937

Scholz, C. F. P., Poulsen, K., & Kilian, M. (2012). Novel molecular method for identification of *Streptococcus pneumoniae* applicable to clinical microbiology and 16S rRNA sequence-based microbiome studies. *Journal of Clinical Microbiology, 50*(6), 1968–1973. https://doi.org/10.1128/JCM.00365-12

Sperling, J. L., Silva-Brandão, K. L., Brandão, M. M., Lloyd, V. K., Dang, S., Davis, C. S., Sperling, F. A. H., & Magor, K. E. (2017). Comparison of bacterial 16S rRNA variable regions for microbiome surveys of ticks. *Ticks and tick-borne diseases*, *8*(4), 453–461. https://doi.org/10.1016/j.ttbdis.2017.02.002

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., Nesme, X., Rosselló-Mora, R., Swings, J., et al. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology, 52*(4), 1043–1048.

Straume, D., Stamsås, G. A., & Håvarstein, L. S. (2015). Natural transformation and genome evolution in *Streptococcus pneumoniae*. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, *33*, 371–380. https://doi.org/10.1016/j.meegid.2014.10.020

Tindall, B. J., Rossello-Mora, R., & Klenk, H. P. (2010). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, 8(5), 322-332. https://doi.org/10.1038/nrmicro2316

Versmessen, N., Mispelaere, M., Vandekerckhove, M., Hermans, C., Boelens, J., Vranckx, K., Van Nieuwerburgh, F., Vaneechoutte, M., Hulpiau, P., & Cools, P. (2024). Average Nucleotide Identity and Digital DNA-DNA Hybridization Analysis Following PromethION Nanopore-Based Whole Genome Sequencing Allows for Accurate Prokaryotic Typing. *Diagnostics (Basel, Switzerland)*, *14*(16), 1800. https://doi.org/10.3390/diagnostics14161800

Větrovský, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PloS one*, *8*(2), e57923. https://doi.org/10.1371/journal.pone.0057923

Wambui, J., Cernela, N., Stevens, M. J. A., & Stephan, R. (2021). Whole Genome Sequence-Based Identification of *Clostridium estertheticum* Complex strains supports the need for taxonomic reclassification within the species *Clostridium estertheticum*. *Frontiers in Microbiology, 12*, Article 727022. https://doi.org/10.3389/fmicb.2021.727022

Wei, K., Zhang, T., & Ma, L. (2018). Divergent and convergent evolution of housekeeping genes in human-pig lineage. *PeerJ*, *6*, e4840. https://doi.org/10.7717/peerj.4840

Wei, Y., Sturges, C. I., & Palmer, K. L. (2023). Human Serum Supplementation Promotes *Streptococcus mitis* Growth and Induces Specific Transcriptomic Responses. *Microbiology spectrum*, *11*(3), e0512922. https://doi.org/10.1128/spectrum.05129-22

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of bacteriology*, *173*(2), 697–703. https://doi.org/10.1128/jb.173.2.697-703.1991

World Health Organization (2019) Pneumococcal conjugate vaccines in infants and children under 5 years of age: WHO position paper. Wkly Epidemiol Rec 94:85–104r

Internet Sites

Ahmed Abdullah. (2015, November 17). *Are erythrocytes lysed during alpha hemolysis?* Biology Stack Exchange. Bangladesh, Downloaded from https://biology.stackexchange.com/questions/40711/are-erythrocytes-lysed-during-alpha-hemolysis, 2025-02-25

APHL (2025) Association of Public Health Laboratories United States, Downloaded from https://www.aphl.org, 2025-05-23

BioMérieux (2025) United States, Downloaded from (https://www.biomerieux.com/corp/en/our-offer/clinical-products/api-id-strip-range-apiweb.html), 2025-09-03

BSC (2025) Barcelona, Supercomputing Center - Centro Nacionale de Supercomputación Spain, Downloades from https://compss-doc.readthedocs.io/, 2025-08-08

Berkley Library (2025) Berkerley Library University of California United States, Downloaded from https://guides.lib.berkeley.edu/, 2025-08-08

CCUG (2024) Culture Collection University of Gothenburg Sweden, Downloaded from https://www.ccug.se, 2024-02-18

Gendex (2025) Genome Diagnostics Netherlands, Downloaded from https://www.gendx.com, 2025-05-17

NIH (2025) National Library of Medicine - National Center for Biotechnology Information (NCBI) United States, Downloaded from https://www.ncbi.nlm.nih.gov, 2025-08-08

List of Prokaryotic names with Standing in Nomenclature (LPSN) Germany, Downloaded from https://lpsn.dsmz.de/text/introduction, 2025-08-30

Artificial Intelligence

OpenAI. (2025). *ChatGPT* (Version GPT-5) [English]. https://chat.openai.com/chat

Julius AI. (2025). *Julius AI* [AI writing assistant]. https://www.julius.ai/

Appendix 1- Popular Science summary

Bacteria are small living organisms, and they are found everywhere: on plants, in animals, in soil and water. The quantity of bacteria living in and on the surface of human bodies are circa 1–3% of its total weight! Yes, it is true! Together, these communities live in different parts of our body as skin, airways, mouth and our gut. It is why they are a part of the human microbiota.

But are bacteria truly our friends, or are they enemies? The answer is: both.

The friendly bacteria are essential for our survival. They help us digest food, break down carbohydrates, produce vitamins, and support our immune system. There is a dependent partnership between bacteria and living organisms. To keep this functional partnership strong, we need to have a healthy lifestyle, such as regular physical activities, and having a healthy diet rich in fruits and vegetables. To sleep well is also included. These habits not only improve our wellbeing but also maintain our microbiota in balance. In this way, bacteria act like loyal friends, helping us every day.

However, there is another side to the story, where bacteria are responsible for causing serious infectious diseases. Streptococcus mitis group (SMG), for example, is a group of bacteria considered living peacefully in the mouth and respiratory system (ex: *Streptococcus mitis and Streptococcus pneumoniae*). But if they reach the bloodstream, they can cause life-threatening infections such as endocarditis, an inflammation of the heart. Worldwide, infections caused by SMG are responsible for millions of deaths each year. What makes them especially challenging is their ability to exchange genetic material, creating new variants that are harder to detect with standard lab tests. When doctors cannot correctly identify the bacteria, patients will be treated with wrong antibiotics, which can lead to complications, long-term health problems, or even death.

So, how can we fight back?

Improving bacterial identification techniques is a start to future solution, where researchers and clinicians can identify the exact species causing an infection and thereafter, they can choose the most effective treatments. This means faster recovery for patients, fewer days spent in hospitals, reduced chances of recurrence, which means overall lower healthcare costs.

In my project, I studied the taxonomic complexity of the Streptococcus mitis group. The findings confirmed how difficult it is, especially these species belonging to SMG. My research also revealed novel species, and each new discovery improves our knowledge and brings us closer to more accurate diagnoses and better treatments for patients worldwide.

Appendix 2 - PCR templates to the genes 16S rRNA, *groEL and rpoB*

Table 1. PCR schedules to 16S rRNA results when adjusting the amount of DNA in the samples.

CCUG Accession number PCR Concentrations	**19074**	**25812**	**26924**	**27740**	**27741**	**28754**	**31489**	**31557**	**32108**	**32132**
PCR 1	-	-	-	-	-	-	-	-	-	-
PCR 2										
1:20						-		-		
1:50						-		-		
1:100						-		-		
PCR 3										
1:100		+							-	
1:1000		+							+	
PCR 4										
1:100	-	+	+	+	+	+	-	-	+	+
PCR 5										
1:1	+						+			

	32331	**32466**	**32601**	**33514**	**33690**	**35276**	**35278**	**35580**	**36639**	**36753**	**36755**
PCR 1											
1:10		-	-	-	-	-	-	-	-	-	-
PCR 2											
1:20	-	-									
1:50	-	-									
1:100	-	-									
PCR3											
1:100			-				+				
1:1000			+				+				
PCR4											
1:100	+	-	+	+	+	+	+	+	+	+	-
PCR 5											
1:1				+							

	37308	**39096B**	**39210**	**41450**	**42636**	**42984**	**44763**	**44969**	**49591**	**55622B**	
PCR 1											
1:2	+	+	+	+	+	+	+	+	+	+	
PCR 5											
1:1			+								

+ = Positive PCR - = Negative PCR +- = Weak bands.

OBS: Negative PCR and weak bands on gel electrophoresis required new PCR performance.

Table 2. PCR schedule groEL gene results when adjusting the amount of DNA in the samples.

CCUG assession PCR nr Concentrations	**19074**	**25812**	**26924**	**27740**	**27741**	**28754**	**31489**	**31557**	**32108**	**32132**
PCR 1										
1:1						+-				
1:2							+-			
1:3								+-		+
PCR 2										
1:1						+-	+-	+-		+
1:2						+-	+-	+-		+

Table 2. Continued. PCR schedule groEL gene results when adjusting the amount of DNA in the samples.

PCR 3											
1:100	-	-	-	-	-				-	-	
PCR 4											
1:10	+	+	+								
1:100	-	-	-								
1:1000											
PCR 5											
1:10			+	+	+				+		
PCR 6											
1:10	+	+	+	+	+				+	+	
	32331	**32466**	**32601**	**33514**	**33690**	**35276**	**35278**	**35580**	**36639**	**36753**	**36755**
PCR 1											
1:1		+									
1:2						+		-			
1:3	+		+-				+-		+-	+	+
PCR 2											
1:1	+	+	+			-		+-	-	-	-
1:2	+	+	+			-		+-	-	-	-
PCR3											
1:20	-			-	-			-			
1:100											
PCR4											
1:10									+		
1:100									-		
PCR 5	+			+	+			+		+	
1:10											
PCR 6											
1:10	+	+		+	+			+		+	
	37308	**39096B**	**39210**	**41450**	**42636**	**42984**	**44763**	**44969**	**49591**	**55622 B**	
PCR 1											
1:1										+	
1:2	+-	+-	+	+-	-	+--	+	+	+-		
1:3											
PCR2											
PCR3											
PCR 4											
1:10	+										
1:100	-										
PCR 5											
1:10		+		+	+	+			+		
PCR 6											
1:10	+	+				+			+		
1:50					+						

OBS: Strains CCUG 35514 didn't work for *groEL* gene by Sanger sequencing. The provided sequence was of bad quality although the positive PCR

+ = Positive PCR - = Negative PCR +- = Weak bands. OBS: Negative PCR amplicons and weak band on gel electrophoresis requires new PCR performance.

Table 3. PCR schedules to rpoB gene results when adjusting the amount of DNA in the samples.

CCUG Accession nr PCR Concentrations	**19074**	**25812**	**26924**	**27740**	**27741**	**28754**	**31489**	**31557**	**32108**	**32132**	
PCR 1											
1:50				+	+-	+	+-	+	+	+	
PCR 2											
1:50	-	+									
1:80					+		+				
PCR 3											
1:100	-		+								
	32331	**32466**	**32601**	**33514**	**33690**	**35276**	**35278**	**35580**	**36639**	**36753**	**36755**
PCR 1											
1:50	+-	+	-						-		+
PCR 2											
1:50				+	+	-	-	+		+	
1:80	+		+						+		
PCR3											
1:100						+	-				
PCR4											
1:10							+				
	37308	**39096B**	**39210**	**41450**	**42636**	**42984**	**44763**	**44969**	**49591**	**55622B**	
PCR 1											
1:2											
PCR 2											
1:50	-	+									
1:80											
PCR 3											
1:100	+		+	+	+	+	+	+	+	+	
OBS: Strains 32466 and 41450 didn't work for *rpoB* gene by Sanger sequencing. The provided sequences were of bad quality although the positive PCR											

+ = Positive PCR - = Negative PCR +- = Weak bands. OBS: Negative PCR amplicons and weak band on gel electrophoresis requires new PCR performance.

Appendix 3 – BLAST results

Table 1. BLAST results of the housekeeping genes 16S rRNA, groEL and rpoB (First best hit)

CCUG Strains Code	Scientific Name 16S	BLASTn % Ident	Scientific name groEl	BLASTn % Ident	Scientific Name rpoB	BLASTn % Ident
19074	*S. hohhotensis*	100.00%	*S. pseudopneumoniae*	94.32%	*S. hohhotensis*	96.28%
25812	*"S. nidrosiense"*	99.86%	*"S. nidrosiense"*	95.38%	*"S. nidrosiense"*	98.14%
26924	*S. fermentans*	99.44%	*S. mitis*	95.51%	*S. ilei*	97.71%
27740	*S. mitis*	100.00%	*S. mitis*	95.65%	*S. toyakuensis*	98.55%
27741	*S. mitis*	100.00%	*S. mitis*	100.00%	*S. toyakuensis*	98.71%
28754	*"S. symci"*	99.86%	*S. mitis*	94.72%	*S. toyakuensis*	97.99%
31489	*S. hohhotensis*	100.00%	*S. mitis*	96.70%	*S. toyakuensis*	98.57%
31557	*S. mitis*	100.00%	*S. mitis*	96.97%	*S. hohhotensis*	98.14%
32108	*"S. lingualis"*	100.00%	*"S. lingualis"*	98.42%	*"S. dentalis"*	99.71%
32132	*S. nakanonensis*	99.73%	*S. nakanonensis*	91.45%	*S. oralis*	95.70%
32331	*S. infantis*	99.05%	*S. infantis*	94.44%	*S. infantis*	97.35%
32466	*"S. fermentans*	99.40%	*"S. koreensis"*	95.51%	*"S. koreensis"*	97.45%
32601	*"S. shenyangsis"*	99.29%	*S. mitis*	95.38%	*"S. vulneris"*	96.85%
33514	*S. toyakuensis*	99.73%	----------	---------	*S. mitis*	98.14%
33690	*"S. symci"*	99.72%	*S. mitis*	94.60%	*S. toyakuensis*	97.99%
35276	*S. mitis*	99.85%	*S. pseudopneumoniae*	94.19%	*S. nakononensis*	98.85%
35278	*S. toyakuensis*	99.59%	*S. mitis*	96.04%	*S. hohhotensis*	98.28%
35580	*S. mitis*	99.86%	*S. mitis*	96.04%	*S. pneumoniae*	97.21%
36639	*"S. nidrosiense"*	99.84%	*"S. nidrosiens"*	93.26%	*S. chosunense*	97.42%
36753	*"S. nidrosiense"*	100.00%	*"S. nidrosiens"*	92.48%	*S. oralis*	95.54%
36755	*S. mitis*	99.73%	*S. mitis*	95.51%	*S. toyakuensis*	97.42%
37308	*"S. nidrosiense"*	99.32%	*S. mitis*	96.44%	*S. hohhotensis*	98.28%
39096B	*"S. nidrosiense"*	99.32%	*S. mitis*	96.44%	*S. chosunense*	98.42%
39210	*Granulicatella adiacens*	98.78%	*S. dentalis*	97.88%	*S. parasanguinis*	94.93%
41450	*S. cristatus*	100.00%	*S. cristatus*	91.51%	----------	--------
42636	*"S. symci"*	99.86%	*S. mitis*	94.72%	*S. toyakuensis*	97.99%
42984	*S. infantis*	98.17%	*"S. nidrosiens"*	94.45%	*S. infantis*	95.27%
44763	*"S. symci"*	99.86%	*S. mitis*	96.70%	*S. toyakuensis*	100.00%
44969	*S. mitis*	99.73%	*S. mitis*	96.97%	*S. nakanonensis*	97.42%
49591	*S. mitis*	100.00%	*S. mitis*	96.17%	*S. hohhotensis*	98.71%
55622B	*S. hohhotensis*	99.85%	*S. mitis*	96.93%	*S. nakononensis*	96.56%

Table 2. BLAST results for 16S rRNA. groEL and rpoB genes. and the similarities rates.

Obs: same colour = same species

16S rRNA ***Species Scientific name*** **(quantity)** **Strains Code**	**% identity Referens sequence** **GenBank**	***groEl*** ***Species Scientific name*** **(quantity)** **Strains**	**% identity Referens sequence** GenBank	***rpoB*** ***Scientific Name*** **(quantity)** **Strains**	**% identity Referens sequence** GenBank
S. hohhotensis (3) CCUG19074 CCUG31489 CCUG55622B	99.5- 100%	*S. pseudopneumoniae* (2) CCUG19074. CCUG35276	94.2% - 94.3%	*S. hohhotensis* (6) CCUG19074. CUG331557, CUG35278, CCUG37308 CCUG44969, CUG55622B	96.3% 98.3 - 98.7%
S. toyakuensis (2) CCUG33514 CCUG35278	99.6% 99.8%			*S. toyakuensis* (8) CCUG27740, CCUG27741 CCUG28754, CCUG31489 CCUG33690, CCUG36755 CCUG42636, CCUG44763	97.4%- 100%
S. nakanonensis (1) CCUG32132	99.7%	*S. tigurinus* (1) CCUG32132	91.5%	*S. nakanonensis* (2) CCUG35276, CCUG44969 CCUG55622B (1)	97.4% - 98.9% 96.6%
"S. nidrosiense" (5) CCUG25812, CUG33639, CCUG36753, CUG37308 CCUG39096B	99.3% - 100.0%	*"S. nidrosiense"* (4) CCUG25812, CUG33639, CCUG63753, CUG42984	92.5% - 95.4%	*"S. nidrosiense"* (1) CCUG25812	98.4%
S. mitis (8) CCUG27740, CCUG27741 CCUG31557, CCUG35276 CCUG35580, CCUG36755 CCUG44969, CCUG49591	99.7% - 100.0%	*S. mitis* (18) CCUG26924,CCUG27740, CCUG27741, CCUG28754 CCUG31489,CCUG31557, CCUG32601, CCUG33690, CCUG35278, CCUG35580, CCUG37308, CUG39096B, CCUG42636, CCUG44763, CCUG55622B, CUG36755, CCUG44969, CCUG49591	94.7% -100%	*S. mitis* (1) CCUG33514	98.1%

Table 2. Continued. BLAST results for 16S rRNA. groEL and rpoB genes. and the similarities rates.

S. infantis (2) CCUG32331, CCUG32984	96.4%-98.2%	*S. infantis* (1) CCUG32331	94.4%	*S. infantis* (2) CCUG42984, CCUG32331	95.3% - 97.4%
"S. shenyangsis" (1) CCUG32601	99.3%	*S. dentalis* (1) CCUG3921	97.9%	*S. dentalis*(1) CCUG32108	99.7%
"S. lingualis"(1) CCUG32108	100.0%	*"S. lingualis"* (1) CCUG32108	98.4%	*"S. vulneris"* (1) CCUG32601	96.9%
Granulicatella adiacens (1) CCUG39210	98.8%			*S. ilei* (1) CCUG26924	97.7%
S. fermentans (2) CCUG26924, CCUG32466	99.4%			*S. pneumoniae*(1) CCUG 35580	97.2%
"S. symci" (4) CCUG28754, CCUG33690 CCUG42636, CCUG44763	99.7 - 99.9%			*S. chosunense* (2) CCUG36639, CCUG39096B	97.4 -98.4%
				S. parasanguinis (1) CCUG39210	94.9%
				S. oralis (2) CCUG32132, CCUG36753	95.5%-95.7%
		"S. koreensis" (1) CCUG32466	95.6%	*"S. koreensis"* (1) CCUG32466	97.4%
S. cristatus (1) CCUG41450	100.0%	*S. cristatus* (1) CCUG41450	91.5%		

Appendix 4- Phylogenetic trees of the housekeeping genes 16S rRNA, groEL and rpoB

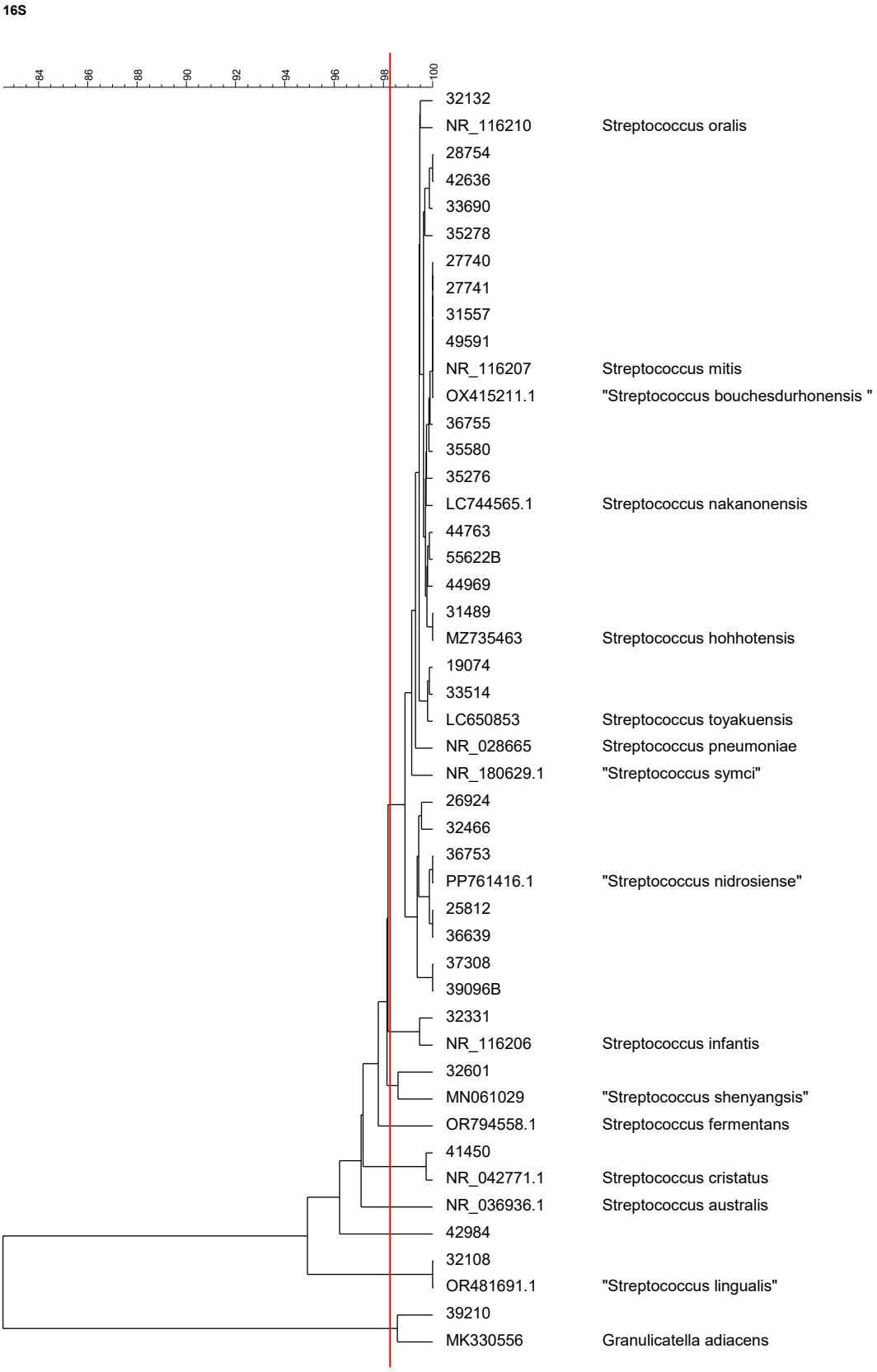

Figure 1 - Phylogenetic tree of gene 16S rRNA

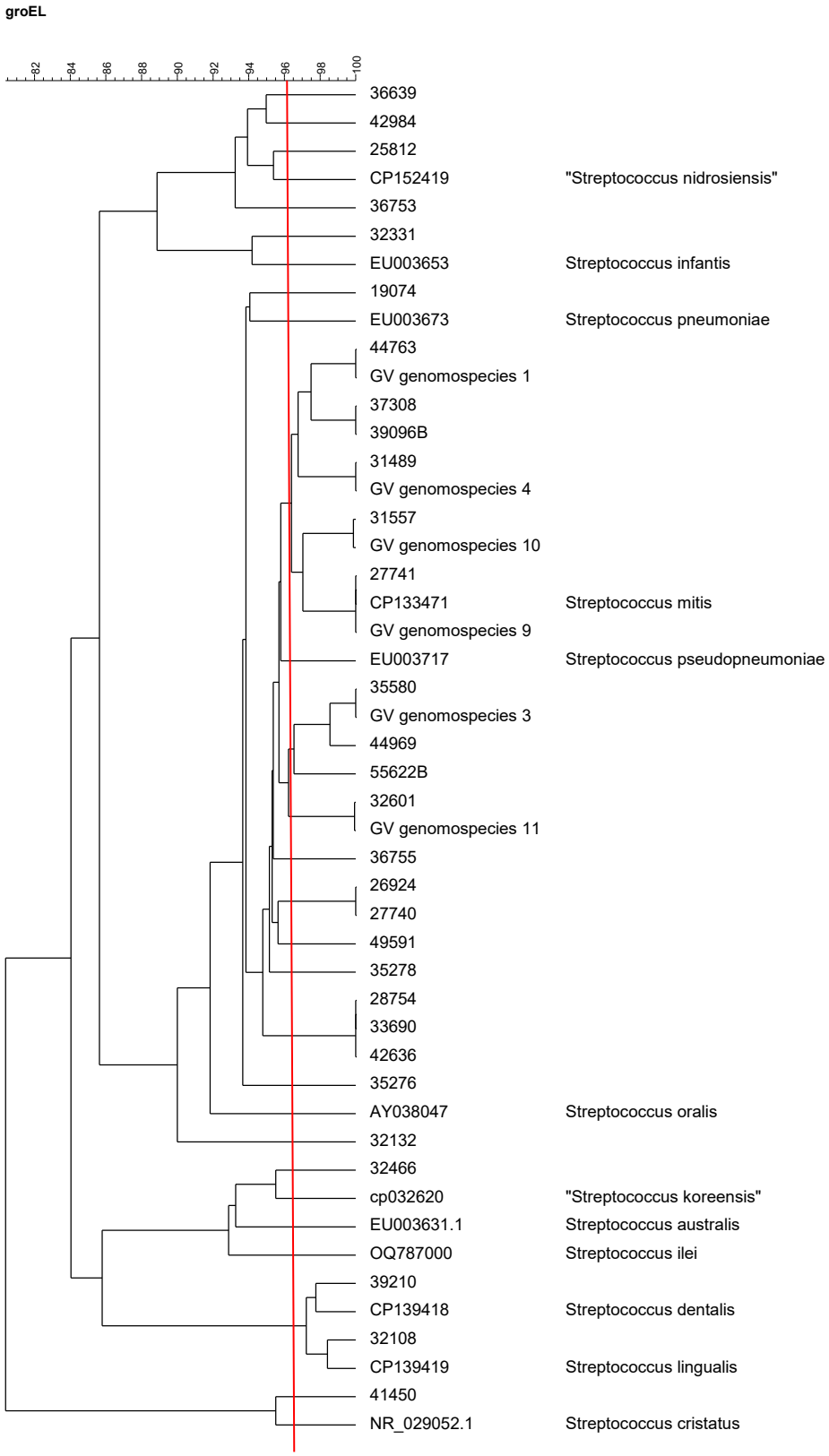

Figure 2 - Phylogenetic tree of gene *groEL*

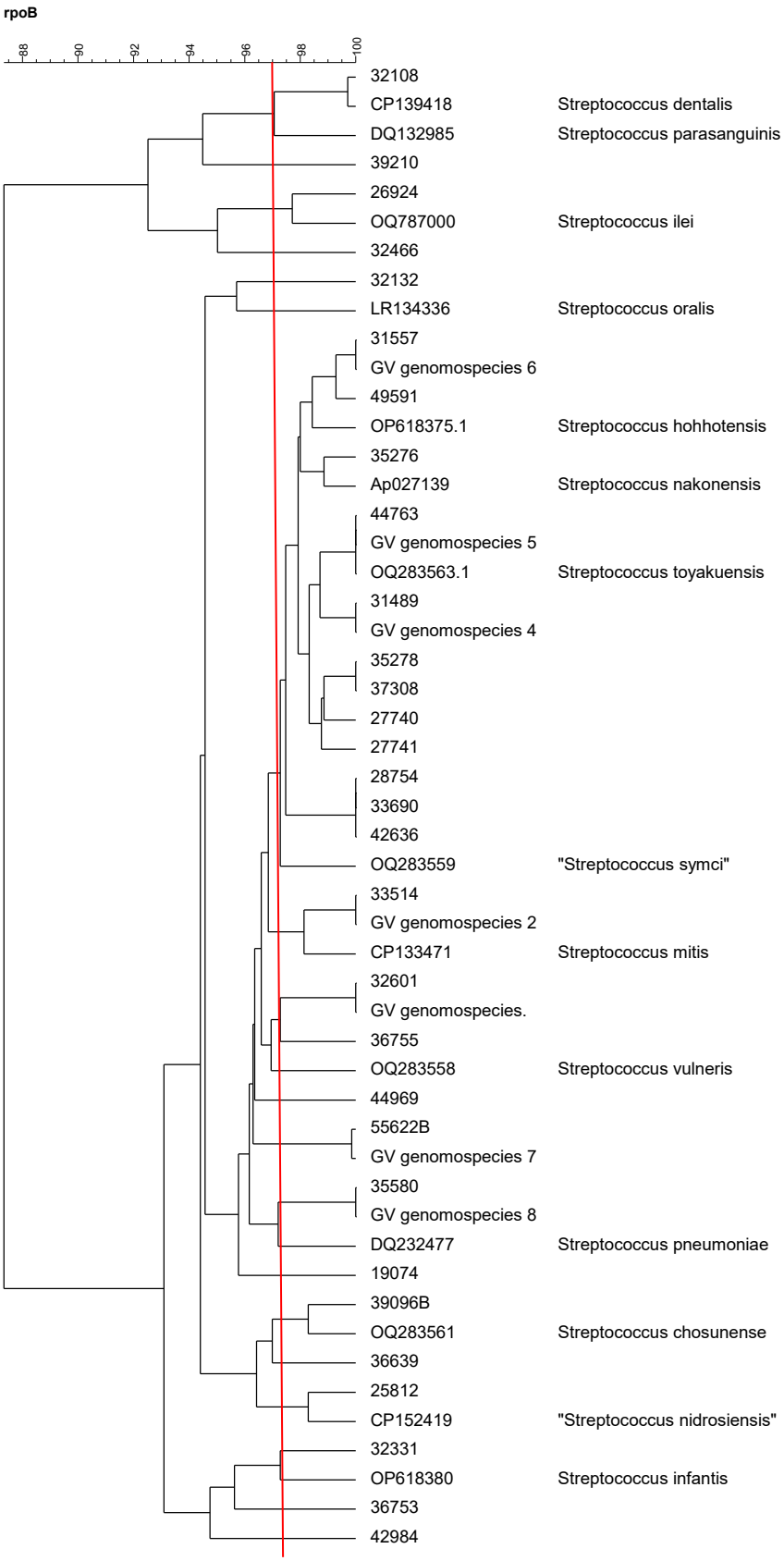

Figure 3 - Phylogenetic tree of gene *rpoB*