Genomic and metagenomic analysis of microbial agents causing infective endocarditis

Iversen, Katrine Højholt

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Genomic and metagenomic analysis of microbial agents causing infective endocarditis

Katrine Højholt Iversen

December, 2019

Technical University
of Denmark

Katrine Højholt Iversen, MSc.
Department of Health Technology, Section for Bioinformatics, Technical University of Denmark.
Genomic and metagenomic analysis of microbial agents causing infective endocarditis

Supervisors
Anders Gorm Pedersen, Ph.D., Professor.
Department of Health Technology, Section for Bioinformatics, Technical University of Denmark.

Simon Rasmussen, Ph.D., Associated Professor.
Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

Xiaohui Chen Nielsen, MD, Ph.D., Senior Consultant
Department of Clinical Microbiology, Slagelse Hospital, Denmark

Jens Jørgen Christensen, MD, Ph.D., Clinical Professor
Department of Clinical Microbiology, Slagelse Hospital, Denmark.
Institute of Clinical Medicine, University of Copenhagen, Denmark.

Assessment committee
Gisle Alberg Vestergaard, Ph.D., Associate Professor.
Department of Health Technology, Section for Bioinformatics, Technical University of Denmark.

Mogens Kilian, Ph.D., Professor Emeritus.
Depardment of Biomedicine - Research and Education, Aarhus University, Denmark

Nathan Wales, Ph.D., Lecturer.
Department of Archaeology, University of York, United Kingdom.

Front cover art:

Dental disease, File ID: 131078420, by Guniita. License: Royalty free

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

— Marie Curie

Preface

This thesis was prepared at the Department of Health Technology, Section for Bioinformatics, at the Technical University of Denmark (DTU). The presented work consists of four peer-reviewed papers and one manuscript in review, produced during the period 2015 - 2019.

This thesis was carried out under the supervision of associate professor Simon Rasmussen, professor Anders Gorm Pedersen, clinical consultant Xiaohui Chen Nielsen, and Clinical Professor Jens Jørgen Christensen.

The Ph.D. was funded by the Hjerteforeningen (grant number 15-R99-A6040-22951) and by Technical University of Denmark

Lyngby, December 2019
Katrine Højholt Iversen

Summary

Infective endocarditis (IE) is an infection of the inner layer of the heart or the heart valves. Gram-positive cocci of the genera *Staphylococcus*, *Streptococcus*, and *Enterococcus* account for more than 80% of all IE cases. In rare cases, other microbial agents, such as *Aerococcus* spp. can cause the infection as well. IE is a relatively rare disease with two to ten incidents per 100,000 individuals per year in the general population world-wide. However, the mortality of IE is between 15-20%, and one-year mortality approaches 40%. Therefore, IE is a severe infectious disease, as it involves long term hospitalization and heart surgery in 50% of the cases.

This thesis includes five studies, and its main focus is IE caused by Mitis group streptococci (MGS). The majority of the bacteria in this group are commensal colonizers of the human oral cavity, where they can have beneficial effects on the oral health of the host. However, this group of oral bacteria can escape their niche and in rare cases, cause infectious diseases such as IE.

The MGS are closely related and naturally competent, which enables them to engage in recombination with closely related species. This complicates correct species identification, which is a crucial part of the diagnosis of IE. Additionally, species identification is important for correct antibiotic treatment, prognosis, and potential underlying predispositions. In cases of recidive of IE, it is also important to establish if the infection is due to treatment failure or whether it is caused by an infection by another species. The first study in this thesis compares different molecular phylogenetic approaches to identify the most suitable method for correct species identification using whole-genome sequencing of 80 MGS strains isolated from IE patient samples.

MGS consist of several different species. Among those are the commensal *Streptococcus mitis* and *Streptococcus oralis*, which are closely related with the pathogen *Streptococcus pneumoniae*. *S. pneumoniae* is one of the leading causes of fatal infections in children world-wide. As the majority of MGS live as

commensals in symbiosis with the host, the virulence potential of these species is not fully established. Study two and three in this thesis approaches this issue. Study two investigated genes that could be involved in the virulence of 40 *S. mitis* and *S. oralis* strains. All strains were isolated from IE patients and therefore had the ability to cause infections. In the third project, MGS genomes isolated from IE patients were compared with strains isolated from the oral cavity of healthy individuals with the aim to illuminate genetic differences that could explain the shift in pathogenicity.

Study four investigated the virulence potential of MGS persevered in a 5,700-year-old chewed birch pitch, i.e., ancient "chewing gum". As the pitch had been chewed, the oral microbiota of a western hunter-gatherer woman had been preserved inside the lump. The ancient streptococcal DNA was identified and compared to DNA from modern streptococci. The pathogenic potential of the ancient *Streptococcus* DNA was investigated to see if the ancient MGS contained more or fewer virulence genes than the modern MGS.

Study five in this thesis aimed to identify virulence genes in strains from another, though rare, IE-associated genus, *Aerococcus* spp. The two species *Aerococcus sanguinicola* and *Aerococcus urinae* can cause urinary tract infections, sepsis, and, more rarely, IE. No whole-genome comparisons and genomic characterizations have previously been performed on *A. sanguinicola* and *A. urinae*. In this study, we were able to identify several virulence genes associated with human disease.

Collectively, the thesis' first four studies provided an insight into the nature of MGS. For correct species identification, we showed that multi-locus sequence analysis involving seven genes provided enough genetic variability to generate distinct phylogenetic clusters. However, in cases where the hospitals have implemented whole-genome sequencing, species identification using CGI Phylogeny might be the fastest and most accurate approach. We furthermore identified genes that could be associated with virulence; genes associated with host colonization and modulation of the host immune system were identified in the *S. mitis*, *S. oralis*, *S. sanguinins*, *S. gordonii*, ancient *Streptococcus* DNA, and in the two *Aerococcus* species. The same level of virulence genes was identified in the ancient *Streptococcus* DNA as in modern oral samples. Interestingly, we found no genetic differences between strains isolated from patients with IE and the oral cavity of healthy individuals. Altogether, our findings illustrate the complexity of the role of these bacteria in IE etiology. While they all, most likely, carry a pathogenic potential, there might be other factors than the genetic composition that determines whether the bacterium can successfully establish an infection.

Dansk resumé

Infektiøs endocarditis (IE) er en infektion i det inderste lag af hjertet eller i hjerteklapperne. Gram-positive kokker fra slægterne *Staphylococcus*, *Streptococcus* og *Enterococcus* udgør 80-90% af alle tilfælde af IE. I sjældne tilfælde kan andre bakterier, såsom *Aerococcus* spp. forårsage IE. IE er en relativ sjælden sygdom med to til ti tilfælde per 100,000 individer per år i den almene befolkning på verdensplan. Dog er dødeligheden af IE mellem 15-20% og et års-dødeligheden kan være så høj som 40%. Derfor er IE en meget alvorlig infektionssygdom, som kræver langvarig behandling og hjerteoperation i 50% af tilfældene. Denne afhandling har sit hovedfokus på IE forårsaget af Mitis gruppe streptokokker (MGS) indeholder fem videnskabelige studier. Størstedelen af de bakterier, der tilhører MGS, er harmløse orale kommensaler, hvor deres tilstedeværelse kan have en gavnlig effekt på deres vært. I nogle tilfælde kan denne gruppe af orale bakterier dog undslippe deres niche og i sjældne tilfælde forårsage infektionssygdomme såsom IE.

MGS er nært beslægtede og er naturligt kompetente, hvilket gør dem i stand til at rekombinere med andre nært beslægtede arter. Dette vanskeligør en korrekt artsidentifikation, som ellers er vigtig for diagnose af IE. Derudover er artsidentifikationen vigtig for at tilrettelægge en korrekt antibiotikabehandling, vurdere prognosen og identificering af potentielle underliggende prædispositioner. I tilfælde af tilbagefald hos patienten er det også vigtigt at undersøge, om den tilbagevendende infektion skyldes behandlingssvigt, eller om den er forårsaget af en ny infektion med en anden bakterieart. Det første studie i denne afhandling sammenligner forskellige molekylære fylogenetiske fremgangsmåder for at undersøge hvilken en, der er den bedst egnede fremgangsmåde til korrekt artsidentifikation under anvendelse af helgenom-sekventering af 80 MGS-stammer isoleret fra IE-patiener.

MGS består af flere forskellige arter. Blandt disse er de kommensale *Streptococcus mitis* og *Streptococcus oralis*, som er nært beslægtede med den pato-

gene *Streptococcus pneumoniae*. *S. pneumoniae* er en af de førende årsager til dødelige infektioner hos børn på verdensplan. Da størstedelen af MGS er harmløse og lever i symbiose med værten, er virulenspotentialet for disse arter ikke fuldt ud etableret. Studie to og tre i denne afhandling beskræftiger sig med dette emne. Studie to undersøgte gener, der kunne være involveret i virulensen af 40 *S. mitis* og *S. oralis* stammer. Alle stammer blev isoleret fra IE-patienter og havde derfor evnen til at forårsage infektion. I det tredje studie blev MGS-genomer isoleret fra IE-patienter sammenlignet med stammer isoleret fra mundhulen hos raske individer. Formålet med dette var at belyse de genetiske forskelle, der kunne forklare skiftet i patogenicitet.

Studie fire undersøgte virulenspotentialet af MGS, der var bevaret i en 5.700 år gammelt tygget birkebeg, dvs. et forhistorisk "tyggegummi" lavet af beg (birketjære). Da birkebegen blev tygget, blev det orale mikrobiom fra en vestlig jæger-samlerkvinde bevaret inde i klumpen. Det gamle streptokok-DNA blev identificeret og sammenlignet med DNA fra moderne streptokokker. Virulenspotentialet for det gamle *Streptococcus* DNA blev undersøgt for at se, om de gamle MGS indeholdt flere eller færre potentielle virulensgener end de moderne MGS.

Målet med studie fem i denne afhandling var at identificere virulensgener IE-associerede bakterier af slægten *Aerococcus* spp.. *Aerococcus sanguinicola* og *Aerococcus urinae* arterne er bakterier, der kan forårsage urinvejsinfektioner, sepsis og i sjældne tilfælde IE. Der er ikke tidligere blevet udført nogen helgenom-sammenligninger eller genomisk karakterisering af *A. sanguinicola* og *A. urinae*. I dette studie identificerede vi virulensgener, der er forbundet med sygedomme i mennesker.

Samlet set giver afhandlingens fire første studier et indblik i MGS-arter. For korrekt artsidentifikation viste vi, at multi-locus-sekvensanalyse, der involverede syv gener, indeholdte tilstrækkelig genetisk variabilitet til at generere distinkte fylogenetiske klynger. På de hospitaler hvor helgenomsekventering allerede er implementeret, vil artsidentifikation ved hjælp af CGI Phylogeny imidlertid være den hurtigste og mest nøjagtige fremgangsmåde. Vi identificerede endvidere gener, der kunne være forbundet med virulens; gener, der er forbundet med vækst i værtens væv samt interaktion med værtens immunsystem blev identificeret i *S. mitis*, *S. oralis*, *S. sanguinins*, *S. gordonii*, gammel *Streptococcus* DNA og i de to *Aerococcus* arter. Det samme niveau af virulensgener blev identificeret i det gamle *Streptococcus* DNA som i moderne orale prøver. Interessant nok fandt vi ingen genetiske forskelle mellem stammer isoleret fra patienter med IE og mundhulen hos raske individer. Studierne i denne afhandling illustrerer kompleksiteten af disse bakteriers rolle i IE-etiologi. De

analyserede bakterier indeholder sandsynligvis alle et patogent potentiale, men der er sandsynligvis yderligere faktorer end den genetiske sammensætning, der bestemmer, om bakterien kan etablere en effektiv infektion.

Acknowledgements

My list of acknowledgments is long, as I have had the luxury of being surrounded by many wonderful people when working on my Ph.D.

First of all, I want to express my gratitude to my supervisor, **Simon Rasmussen**, for all our scientific discussions and your guidance throughout the project. It has been amazing working with you, and I am extremely thankful for your never-ending patience, support, and encouragement when things were difficult. I look forward to continuing our collaboration in the future. Furthermore, I would like to thank my co-supervisors **Xiaohui Chen Nielsen** and **Jens Jørgen Christensen** for letting me in on this project. Without your engagement in this project and the excellent data you provided, this Ph.D. project would probably not had become a reality. I truly appreciate your great insights into the clinical and microbial point of view of things, and that you always take your time to give valuable inputs to my work. I also want to thank my supervisor, **Anders Gorm Petersen**, for taking over the administrative part of the Ph.D. and giving me help and guidance on phylogenetic subjects regarding my projects.

A big thanks to the former Metagenomics Group at DTU and to the Rasmussen Group at KU for all the scientific discussions, constructive feedback, and input you have given me. Thanks to the administrative staff of DTU Bioinformatics and now DTU Health Tech for keeping things running. In connection to that, special thanks go to **Dorthe Kjærsgaard**, **Christel Wagner**, and **Inger Vibeke Dorph Hansen** for your help and forthcomingness in all the administrative issues I have had.

I was granted the great opportunity to join the Evolutionary Genomics group and their exiting scientific work in ancient genomics. Thank you **Hannes**

Schroeder, **Theis Zetner Trolle Jensen**, and **Jonas Niemann** for sharing your project with me and for teaching me so much about ancient genomics data. It has been a great experience and pleasure working with you.

I want to thank all my fellow Ph.D. students and office mates that I have the pleasure of working with and next to. Among those I want to highlight **Louise Hesselbjerg Rasmussen**, **Derya Carkaci**, **Christian Salgård Jensen**, **Franziska Klinke**, **Kosai Al-Nakeeb Jakob Nybo Nissen**, **Jose Juan Almagro Armenteros**, **Rosa Allesøe**, **Joachim Johansen**, **Anor Ingi Sigurdsson**, **Leonardo Cubuccio**, **Henry Webel Marie Louise Jespersen**, and **Kristoffer Niss**. Thank you all for our great discussions, both in lunch and coffee breaks, as well as in the office.

A special thanks to **Janne Marie Moll** for your valuable inputs and constructive feedback on this thesis. Also thanks to **Marie**, **Jakob**, **Joachim**, **Anor**, and **Jose**, for your feedback.

Finally, I genuinely like to thank and appreciate my family and friends for their endless support in my daily life and my academic career. You always pick me up when things get tough, and you help, guide, and distract me when I am stuck in my own head - thank you **Christian, Elly, Karen, Anders, Ea, Anne, Sara, Janne, Kitt** and **Yasemin**.

Publications Included in This Thesis

This thesis is based on the following five articles:

1. Louise H. Rasmussen, Rimtas Dargis, **Katrine Højholt**, Jens Jørgen Christensen, Ole Skovgaard, Ulrik S. Justesen, Flemming S. Rosenvinge, Claus Moser, Oksana Lukjancenko, Simon Rasmussen, and Xiaohui C. Nielsen†. Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of Mitis group streptococci.
   European Journal of Clinical Microbiology & Infectious Diseases,35(10): 1615–1625, October 2016.

2. Louise H. Rasmussen*, **Katrine Højholt***, Rimtas Dargis, Jens Jørgen Christensen, Ole Skovgaard, Ulrik S. Justesen, Flemming S. Rosenvinge, Claus Moser, Oksana Lukjancenko, Simon Rasmussen, and Xiaohui C. Nielsen†. In silico assessment of virulence factors in strains of Streptococcus oralis and Streptococcus mitis isolated from patients with Infective Endocarditis.
   Journal of Medical Microbiology, 66(9): 1316–1323, September 2017.

3. **Katrine Højholt Iversen***, Louise Hesselbjerg Rasmussen*, Kosai Al-Nakeeb, Jose Juan Almagro Armenteros, Christian Salgård Jensen, Rimtas Dargis, Oksana Lukjancenko, Ulrik Stenz Justesen, Claus Moser, Flemming S. Rosenvinge, Xiaohui Chen Nielsen, Jens Jørgen Christensen†, and Simon Rasmussen†. Similar genomic patterns of clinical infective endocarditis and oral isolates of *Streptococcus sanguinis* and *Streptococcus gordonii*.
   The manuscript is in peer-review in Scientific reports.

4. Theis ZT Jensen*, Jonas Niemann*, **Katrine Højholt Iversen***, Anna K Fotakis, Shyam Gopalakrishnan, Mikkel H S Sinding, Martin R Ellegaard, Morten E Allentoft, Liam T Lanigan, Alberto J Taurozzi, Sofie Holtsmark Nielsen, Michael W Dee, Martin N Mortensen, Mads C Christensen, Søren A Sørensen, Matthew J Collins, Tom Gilbert, Martin Sikora, Simon Rasmussen, and Hannes Schroeder†. Stone Age "chewing gum" yields 5,700 year-old human genome and oral microbiome.
Manuscript accepted in Nature Communications, October 29th, 2019.

5. Derya Carkaci*, **Katrine Højholt***, Xiaohui Chen Nielsen, Rimtas Dargis, Simon Rasmussen, Ole Skovgaard, Kurt Fuursted, Paal Skytt Andersen, Marc Stegger, and Jens Jørgen Christensen†. Genomic characterization, phylogenetic analysis, and identification of virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* strains isolated from infection episodes.
Microbial Pathogenesis, 112: 327–340, November 2017.

* Authors contributed equally
† Corresponding author

Additional Publications

The following papers are not included in this thesis:

1. Peter de Barros Damgaard, Rui Martiniano, Jack Kamm, J. Víctor Moreno-Mayar, Guu sKroonen, Michaël Peyrot, Gojko Barjamovic, Simon Rasmussen, Claus Zacho, Nurbol Baimukhanov, Victor Zaibert, Victor Merz, Arjun Biddanda, Ilja Merz, Valeriy Loman, Valeriy Evdokimov, Emma Usmanova, Brian Hemphill, Andaine Seguin-Orlando,Fulya Eylem Yediay, Inam Ullah, Karl-Göran Sjögren, **Katrine Højholt Iversen**, Jeremy Choin, Constanza de la Fuente, Melissa Ilardo, Hannes Schroeder, Vyacheslav Moiseyev, Andrey Gromov, Andrei Polyakov, Sachihiro Omura, Süleyman Yücel Senyurt, Habib Ahmad, Catriona McKenzie, Ashot Margaryan, Abdul Hameed, Abdul Samad, NazishGul, Muhammad Hassan Khokhar, O. I. Goriunova, Vladimir I. Bazaliiskii, John Novembre, Andrzej W. Weber, Ludovic Orlando, Morten E. Allentoft, Rasmus Nielsen, Kristian Kristiansen, Martin Sikora, Alan K. Outram, Richard Durbin, and Eske Willerslev. "The first horse herders and the impact of early Bronze Age steppe expansions into Asia."
Science 29 Jun 2018: Vol. 360, Issue 6396, eaar7711: DOI: 10.1126/science.aar7711

2. Ashot Margaryan*, Daniel Lawson* , Martin Sikora*, Fernando Racimo*, Simon Rasmussen, Ida Moltke, Lara Cassidy, Emil Jørsboe, Andrés Ingason, Mikkel Pedersen, Thorfinn Korneliussen, Helene Wilhelmson, Magdalena Buś, Peter de Barros Damgaard, Rui Martiniano, Gabriel Renaud, Claude Bhérer, J. Víctor Moreno-Mayar, Anna Fotakis, Marie Allen, Martyna Molak, Enrico Cappellini, Gabriele Scorrano, Alexandra Buzhilova, Allison Fox, Anders Albrechtsen, Berit Schütz, Birgitte Skar, Caroline Arcini, Ceri Falys, Charlotte Hedenstierna Jonson, Dariusz Błaszczyk, Denis Pezhemsky, Gordon Turner-Walker, Hildur Gestsdóttir,

Inge Lundstrøm, Ingrid Gustin, Ingrid Mainland, Inna Potekhina, Italo Muntoni, Jade Cheng, Jesper Stenderup, Jilong Ma, Julie Gibson, Jüri Peets, Jörgen Gustafsson, **Katrine Højholt Iversen**, Linzi Simpson, Lisa Strand, Louise Loe, Maeve Sikora, Marek Florek, Maria Vretemark, Mark Redknap, Monika Bajka, Tamara Pushkina, Morten Søvsø, Natalia Grigoreva, Tom Christensen, Ole Kastholm, Otto Uldum, Pasquale Favia, Per Holck, Raili Allmäe, Sabine Sten, Símun Arge, Sturla Ellingvåg, Vayacheslav Moiseyev, Wiesław Bogdanowicz, Yvonne Magnusson, Ludovic Orlando, Daniel Bradley, Marie Louise Jørkov, Jette Arneborg, Niels Lynnerup, Neil Price, M. Thomas Gilbert, Morten Allentoft, Jan Bill, Søren Sindbæk, Lotte Hedeager, Kristian Kristiansen, Rasmus Nielsen†, Thomas Werge†, Eske Willerslev†. "Population genomics of the Viking world."
In review in Nature. Posted on bioRxive July 17, 2019.

3. Theis Zetner Trolle Jensen, Arne Sjöström, Anders Fischer, Erika Rosengren, Liam Thomas Lanigan, Ole Bennike, Kristine Kurzow Richter, Kurt J. Gron, Meaghan Mackie, Morten Fischer Mortensen, **Katrine Højholt Iversen**, Alberto John Taurozzi, Jesper Olsen, Hannes Schroeder, Nicky Milner, Mikkel Sørensen, Matthew James Collins. "Barbed bone point chronology reveals a radiocarbon hiatus, at 10.2 ka, during the Early Mesolithic in southern Scandinavia."

   Manuscript in preparation

Abbreviations

A. sanguinicola	*Aerococcus sanguinicola*
A. urinae	*Aerococcus urinae*
S. australis	*Streptococcus australis*
S. azizii	*Streptococcus azizii*
S. cristatus	*Streptococcus cristatus*
S. gordonii	*Streptococcus gordonii*
S. infantis	*Streptococcus infantis*
S. mitis	*Streptococcus mitis*
S. mutans	*Streptococcus mutans*
S. oralis	*Streptococcus oralis*
S. parasanguinis	*Streptococcus parasanguinis*
S. perioditis	*Streptococcus perioditis*
S. pneumoniae	*Streptococcus pneumoniae*
S. pseudopneumoniae	*Streptococcus pseudopneumoniae*
S. salivarius	*Streptococcus salivarius*
S. sanguinis	*Streptococcus sanguinis*
Y. pestis	*Yersinia pestis*

A	Adenosine
C	Cytosine
G	Glutamine
T	Tyrosine
DNA	DeoxyriboNucleic Acid
AUC	Area Under Curve
FN	False negatives
FP	False positives
FPR	False positive rate
HMP	Human Microbiome Project
HOMS	Human Oral Microbiome Samples
IE	Infective Endocarditis
MALDI-TOF-MS	Matrix-Assisted Laser Desorption Ionization Time-Of-Flight Mass Spectrometry
MCC	Matthews Correlation Coefficient
MGS	Mitis Group Streptococci
PMD	Post-Mortem Damage
ROC	Receiver Operating Characteristics
SNV	Single-Nucleotide Variant
TN	True negatives
TP	True positives
TPR	True positive rate
VFDB	Virulence Factor Database

List of Figures

List of Tables

Table of contents

Mitis Group Streptococci and Infective Endocarditis

Infective endocarditis (IE) is an inflammation of the endocardial surface of the heart and is caused by infectious organisms [1]. Gram-positive cocci of the genera *Staphylococcus*, *Streptococcus*, and *Enterococcus* account for more than 80% of all IE cases, and the infection is often caused by the patient's own commensal strains [2, 3]. This disease is fairly rare; however, the in-hospital mortality rate can be as high as 40% [4, 5]. In spite of improvements in healthcare and higher awareness of hygiene, the incidence of IE has remained constant over the past two decades [6, 7]. This thsis will cover IE caused by oral streptococci belonging to the Mitis group streptococci (MGS), as oral streptococci, together with staphylococci, are the leading cause of IE [1]. Normally, the oral streptococci species are commensals with beneficial effects in the host. A high abundance of strains belonging to the MGS species is associated with a healthy oral cavity without the development of carries [8, 9]. However, MGS can escape from their oral niche and cause a variety of infectious complications, including IE [3]. This section will contain an introduction to the oral microbiome that the above-mentioned *Streptococcus* species inhabit. The virulence of the oral streptococci and their ability to cause infective endocarditis will also be introduced.

1.1 The Oral Microbiome

The oral microbiome is a mixed-species biofilm that covers the oral cavity, and it is composed of several different microorganisms, including viruses, fungi, and bacteria [10]. In many respects, the oral cavity is an extreme environment. The resident microbes have to cope with variations in numerous environmental

factors, such as temperature levels, oxidative stress, and strong hydrodynamic - and mechanical forces caused by food consumption, chewing, and talking [11]. The oral cavity consists of several anatomical sites of both hard and soft tissue: tongue, saliva, throat, tonsils, cheek, palate, gums, supragingival dental plaque, and subgingival dental plaque (Figure 1.1) [12]. The different anatomical sites in the oral cavity have their unique microenvironments that drive the evolution of a diverse microbiome [13].

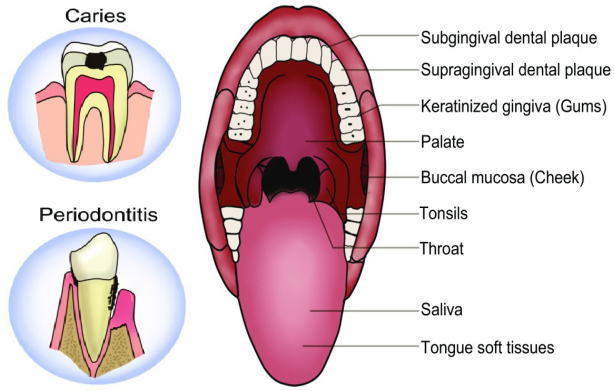

Figure 1.1. The different habitat sites of the oral cavity: tongue, saliva, throat, tonsils, cheek, palate, gums, supragingival dental plaque, and subgingival dental plaque. The two most commen oral diseases: dental caries and periodontal disease, are also illustrated. Modified from Gao et al. [12], link to Creative Commons licens: http://creativecommons.org/licenses/by/4.0/

The oral microbiome is estimated to have around 700 different species and is the second most diverse microbiome site in the human body [10, 14, 15]. Among the most important and predominant oral genera, we find: *Actinomyces*, *Streptococcus*, *Neisseria*, *Veillonella*, *Poryphyrommonas*, and *Selenomonas*. The oral streptococci have been found to be the most common species at all sites in the mouth. They are known to be involved in the initial colonization of the oral cavity and make up over 80% of the early biofilm [13, 16, 14]. Colonization of the oral cavity is initialized within a few hours after birth by bacteria being transferred from the mother, nurses, or environment [17]. Studies have shown

that *Streptococcus sanguinis* start colonizing the oral cavity by attaching the surface of the teeth when the teeth begin emerging around the age of six to nine months [18]. The oral microbiome seems to keep evolving from the pioneer colonization and throughout life. The composition of species appears to depend on the diet of the host as well as the health status of the host. Recent studies have shown that the diversity of oral microbiome changes concurrently with dramatic shifts in the diet: the adaptation of carbohydrate-rich diet in the Neolithic ($\approx 10,000$ years before present), and the more recent introduction of industrially processed flour and sugar around 1850 [19]. One study suggests that this significant shift in diet has played a central role in markedly decreasing the diversity of the oral microbiome and enabled the emergence of modern oral pathogens [19]. The increased consumption of domesticated cereals at the beginning of the Neolithic period is, furthermore, associated with an increase in the prevalence of dental calculus and other oral diseases as dental caries and periodontal disease [19, 12].

The two most common oral diseases in humans are dental caries and periodontal diseases and are, ironically, caused by our natural oral microbiome (Figure 1.1) [12, 10]. These diseases are among the most prevalent chronic diseases of people worldwide and affect 60-90% of children and the vast majority of adults in the industrialized countries [20, 21]. Intake of high levels of carbohydrates increases the risk of developing dental caries [22]. To prevent the development of oral diseases, it has become common practice, especially in developed countries, to implement regular oral hygiene routines by brushing teeth and rinsing with mouthwash [10, 20].

1.2 Mitis Group Streptococci

As described in the previous section, the most common species in the oral cavity belong to the *Streptococcus* genus, more precisely the Mitis group of streptococci (MGS). Streptococci are gram-positive bacteria, belonging to the phylum Firmicutes and Lactobacillales order. The word Streptococcus comes from the Greek strepto (twisted) and coccus (spherical), drawn by the shape of the *Streptococcus* bacterium. Streptococci divides along its axis, forming chains or pairs of streptococci cells. The *Streptococcus* genome is between 1.6Mb and 2.4Mb in length, has a average GC content of approximately 39%, and encodes around 1,700-2,300 genes [23, 24].

Today, more than 100 streptococcal species are recognized [25], however the classification of the *Streptococcus* species has shown to cause great difficulty

[26, 27]. Several schemes have been introduced to standardize the groupings of the streptococcal species; even today scientist faces great challenges when allocating the strains, though having both pheno - and genotypic technologies available. Today the majority of the streptococcal species are classified and grouped based on 16S rRNA sequencing. The 16S rRNA phylogenetics can divide the bacteria into six major groups: Pyogenes, Bovis, Salivarius, Mutans, Anginosus, and Mitis (Figure 1.2) [25]. This thesis will mainly focus on the oral streptococci, belonging to the Mitis group.

1.2.1 Data Availability

The Human Oral Microbiome Database (HOMD), provides a body site-specific comprehensive database containing the prokaryotic species that are present in the human oral cavity [28]. It is possible to download other oral samples and compare the bacterial compositions with other metagenomics data. The database contains 37 species of the *Streptococcus* genus, of which 33 belong to the oral taxa, which makes a total of 122 strains of oral streptococci - all available at HOMD [28] (data from 6/9-2019 at www.homd.org). Within the MGS, we find 13 different species whereas all are publicly available as whole genome sequences at NCBI [29, 30] (data from 6/9-2019 at www.ncbi.nlm.nih.gov). The species included in this thesis are: *Streptococcus mitis*, *Streptococcus oralis*, *Streptococcus infantis*, *Streptococcus cristatus*, *Streptococcus gordonii*, *Streptococcus sanguinis*, *Streptococcus parasanguinis*, *Streptococcus pseudopneumoniae* and *Streptococcus pneumoniae*. The mentioned MGS strains belong to the group of oral streptococci and are known to be commensal, non-periodontal pathogenic bacteria [13].

1.2.2 Mitis Group Streptococci and Oral Health

Due to the high abundance and persistence of oral *Streptococcus* species in the oral microbiome, these bacteria have a high impact on the oral health of the host [31]. Individuals with a healthy oral cavity have shown to have a significantly higher prevalence of *S. sanguinis*, while individuals with dental caries have a significantly higher aboundance of *Streptococcus mutans* and almost no detectable levels of *S. sanguinis* [8, 9]. *S. mutans*, which is one of the species associated with for tooth decay is found to compete with the more beneficial species: *S. sanguinis* and *S. gordonii* [31, 32]. In addition to this, *S. mutans* can secrete bacteriocins that inhibit the growth of *S. sanguinis* and *S. gor-*

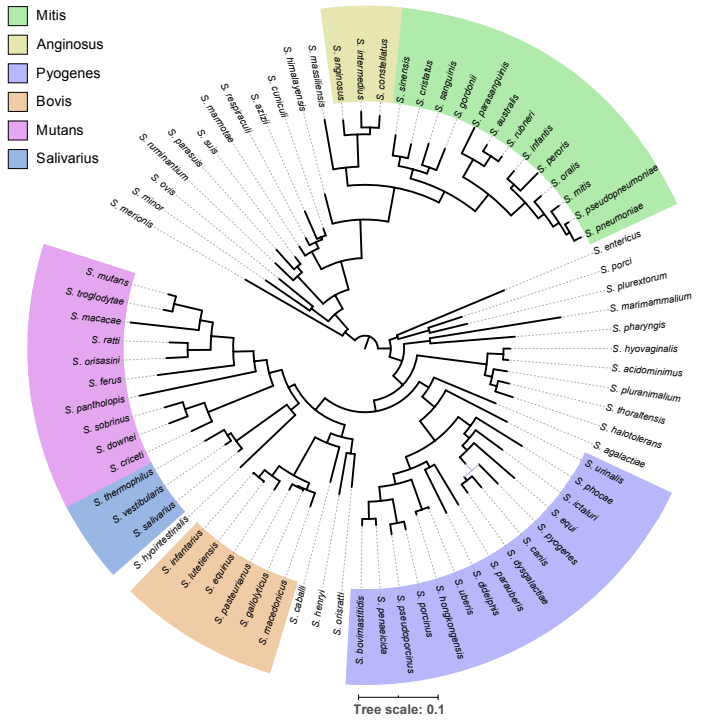

Figure 1.2. Phylogenetic tree illustrating the taxonomic relationship for 77 whole-genome sequenced species of *Streptococcus*. The tree is based on 81 core-gene sequences and constructed as described in Paper 1 and Paper 3 in this thesis. The coloured species in the phylogenetic tree indicates their group: Mitis (green), Anginosus (yellow), Pyogenes (purple), Bovis (orange), Mutans (pink), and Salivarius (blue). The species with no assigned colour do not belong to one of the six major groups. The strain names of the species and their BioProjcet number can be found in Supplementary Table S1.

donii, while *S. sanguinis* and *S. gordonii* inhibits the growth of *S. mutans* by the production of hydrogen peroxide (H2O2) [32, 25].

1.2.3 Mitis Group Streptococci as Opportunistic Pathogens

Even though the majority of MGS are commensal colonizers of the oral cavity, they are also opportunistic pathogens [3, 25]. One member of the MGS, *S. pneumoniae* is one of the most frequent microbial killers of humans [33], with about 14.5 million episodes of serious pneumococcal diseases and 826,000 deaths in children aged 1-59 months in 2000 [34]. *S. pneumoniae* is an asymptomatic colonizer of the nasopharynx. If the bacterium is not cleared by the immune system, it can spread into the lower airways and other organs and tissues [33]. *S. pneumoniae* is one of the leading causes of bacterial pneumonia, meningitis, and sepsis in children worldwide [34, 35]. People with a weakened immune system, as infants and elderly, have a higher risk of *S. pneumoniae* related infections [33]. As described above, especially *S. mitis* is closely related with *S. pneumoniae*. *S. mitis*, as well as many of the other oral commensals of MGS, can cause infectious diseases, such as bacteremia, septicemia, and infective endocarditis [3, 25].

1.3 Infective endocarditis

Infective endocarditis (IE) is an infection of the endocardial surface of the heart, usually the valves [1] (Figure 1.3). The infection is typically caused by bacteria that have entered the bloodstream, either through the oral cavity, the skin, or the intestines (Figure 1.4) [36]. Gram-positive cocci of the genera *Staphylococcus*, *Streptococcus*, and *Enterococcus* account for 80-90 % of IE cases [37, 1]. The current in-hospital mortality rate for patients with IE is 15–20%, with one-year mortality approaching 40% [5, 4]. It is estimated that untreated IE would have a mortality approaching 100% [38]. Despite improvements in medical health care, such as diagnosis, medical therapy, and surgical treatment, the mortality rate has not changed much over the past 25 years [5, 6, 4]. The number of incidents of IE is between two and ten per 100,000 individuals per year in the general population, and two-thirds of the cases of infective endocarditis occur in men [2, 6]. In 2001-2006 the median age of the patients suffering from IE was approximately 70 years, which is a considerable increase from 1980-1984, where the median age was 46 years [39, 40, 37].

The age of the patients, the key risk factor, and the microbiology of the disease vary across the world (Figure 1.4). In low-income countries, rheumatic heart disease remains the key risk factor for IE [37]. Rheumatic heart disease is most often caused by group A streptococci, and the patients are usually young

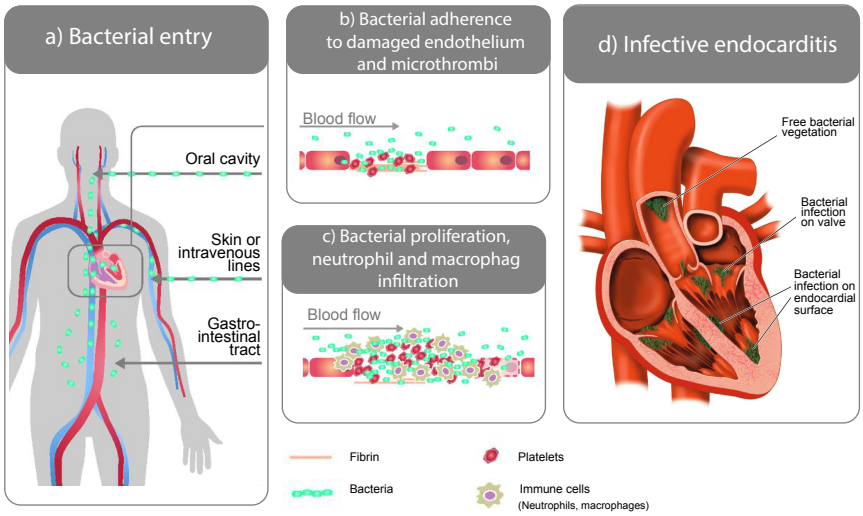

Figure 1.3. Visualisation for the pathogenesis of infective endocarditis. a) illustrates the three common bacterial entries: mouth, skin, skin and gastrointestinal tract. b) visualisation of the bacteria (green) adhering to damaged endothelium and platelets. c) the bacteria (green) starts the biofilmformation, which is infiltrated with immune cells. d) illustration of bacterial vegetation of the heart. The figure is from Thomas J Cahill *et al.* BMJ 2017;358:bmj.j3942 [41] and shutterstock, licens number: 4675800416221, with modifications.

adults [5]. Back in the early 1980s, this was also the key risk factor in high-income countries [37, 40]. Today, degenerative valve disease, diabetes, cancer, intravenous drug use, and congenital heart disease have replaced rheumatic heart disease as the major risk factors for IE in high-income countries. This shift has resulted in *Staphylococcus aureus* replacing streptococci as the most common cause of the disease in many areas of the world (Figure 1.4) [37, 5, 1].

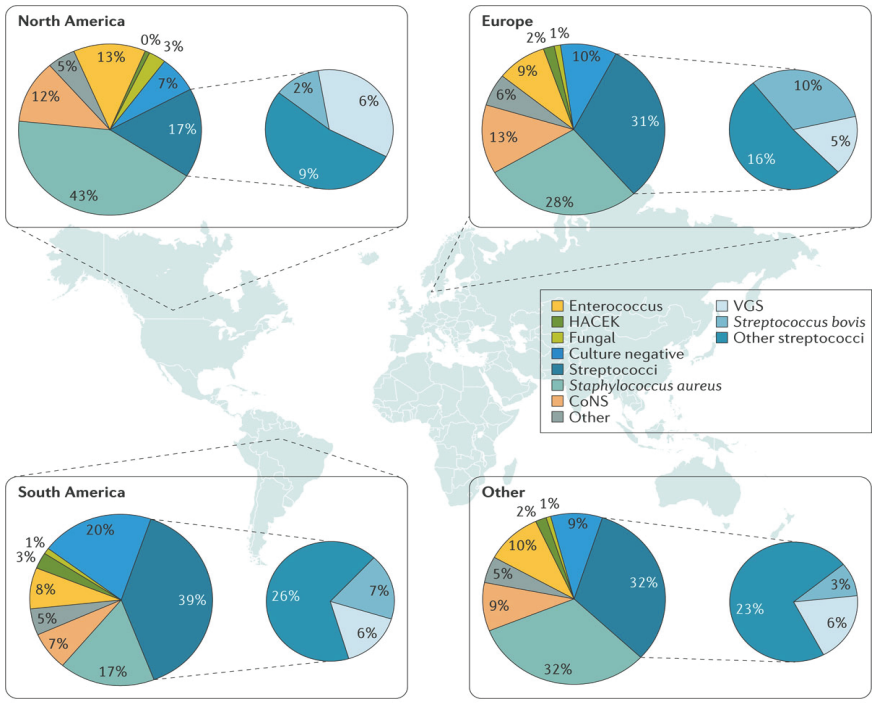

Figure 1.4. Geographically differences of the causative pathogens involved in infective endocarditis across the globe. Data from Murdoch *et al.* [5]. Source: Holland *et al.*, [1].

1.3.1 Diagnosis and Treatment

Infective endocarditis was first described in 1885 by Wiliam Osler. In his seminar "The Gulstonian Lectures, on Malignant Endocarditis", William Osler remarked, "*Few diseases present greater difficulties in the way of diagnosis, difficulties which in many cases are practically insurmountable.*" [42]. More than 130 years later, the diagnosis of infective endocarditis still causes great difficulties [37]. The majority of individuals suffering from infective endocarditis present fever, often associated with systemic symptoms of chills, poor appetite, and weight loss [43, 38]. Also, symptoms as heart murmurs are found

in IE patients. However, in elderly and immunocompromised patients, fever is often absent [43]. The unspecific disease pattern often leads to a delayed diagnosis with a latency period of 20 days [38]. The presence of bacteria in the blood, together with the symptoms mentioned above, is often an indicator of IE. But in some cases, the patient has already received antibiotic treatment before blood samples are collected, which can delay the diagnosis of IE even further [2, 43].

Today the modified Duke criteria are widely used as a tool to diagnose IE [2]. The Duke criteria contain two major criteria: positive blood culture with typical IE pathogens, and positive echo-cardiogram; and six minor criteria including fever and predisposition. Based on the Duke criteria, patients suspected for IE are divided into three categories: definite IE, possible IE, and rejected IE [2, 43]. When infective endocarditis has been diagnosed, it is, in most cases, treated with antibiotics for four to six weeks [1]. Approximately 50% of patients with IE undergoes surgery [2]. Surgery is associated with a higher long-term survival rate; however, the patients with IE are often elderly and extremely sick with multi-system diseases, making surgery very difficult if not impossible [44, 43, 2].

1.3.2 Identification of Streptococci Involved In Infective Endocarditis

Correct species identification is crucial for the diagnosis of infective endocarditis. Additionally, the species identification is important, for correct antibiotic treatment, prognosis, and eventual underlined predispositions [38, 45]. In the case of recurrence of IE, it is important to clarify whether the new IE episode is due to a relapse of the old infection or due to a new infection with a new species. As described in the previous section, species within the Mitis group streptococci (MGS) are closely related, and their ability to exchange genetic material does not simplify the species identification. Different methods can be used in the clinic to make a species identification; single gene sequence, multigene sequences, as well as whole-genome sequence.

1.3.2.1 Taget Gene Identification

Initially, a positive blood culture is necessary to be able to isolate and culture the bacteria. When the bacteria are isolated, sequencing of the 16S rRNA gene can be performed, and based on this; a phylogenetic tree can be reconstructed. Studies have shown that *S. pneumoniae*, *S. oralis*, and *S. mitis* exhibit a very

high sequence similarity within the 16S rRNA gene. The 16S rRNA gene similarity is estimated to be >99% even though they only show a 60% similarity across the whole genome [27]. The 16S rRNA analysis is therefore known to contain a level of uncertainty, and can not always distinguish between the different MGS.

Multi Locus Sequence Analysis (MLSA) is an alternative to the 16S rRNA analysis and has proven to be a fairly fast and low-cost method. It targets seven house-keeping genes that have been identified in the viridans group of streptococci [26]. The seven house-keeping genes are: *rpoB, sodA, ppaC, pyk, map, pfl,* and *tuf*. Using this multi-locus approach, it is possible to distinguish the MGS on species-level, this approach, therefore useful for MGS species identification [26].

1.3.2.2 Identification Based on Whole Genome Sequencing

Whole-genome sequencing (WGS) provides more information about the bacterial genome compared to MLSA and 16S rRNA gene analysis. Today, WGS is available as a method that provides information about the entire bacterial genome (see also section 3.1 for more information). WGS has shown great potential for being used for clinical diagnostics and is applied at some hospitals [46, 47, 48, 49]. However, WGS requires a more extended bioinformatics analysis than the prementioned alternatives, which can be time-consuming and, therefore, expensive. The bioinformatic challenges can be a reason that WGS is not yet widely used as a clinical bacterial diagnostic application [46]. WGS provides a great potential for correct species identification and outbreak detection of infectious diseases by phylogenetic reconstruction [46, 50]. Several different analysis software has been developed to ease the genomic analysis, such as CLC Genomics Workbench [51], Rapid Annotation using Subsystem Technology (RAST) [52], and CSI Phylogeny [50]. Additionally, it is possible to extract specific genes such as house-keeping genes as for MLSA or core-genes from the bacterial genome for species identification or phylogenetic reconstruction.

Virulence of Mitis Group Streptococci

To gain a deeper insight into how the Mitis group streptococci (MGS) can attach and infect the heart valves, it is important to understand how the bacteria grow in their normal niche.

2.1 Biofilm formation

Streptococci are found in natural environments often growing upon surfaces. Many species of streptococci that colonize mammals exist naturally within biofilm communities [25].

The biofilm formation is a complex process, where attachment to the surface is the first step. In the oral cavity, a saliva film containing albumin, glycoproteins, acidic proline-rich proteins, mucins, sialic acids, and other compounds covers the tooth surface [53, 54]. These compounds provide attachment sites for the oral streptococci [53, 25, 54]. Bacterial surface-exposed proteins usually meditate adherence to the host tissue. Oral streptococci can attach to the surface of the oral cavity in various ways, some of which are listed below [25, 55]:

- Direct binding to the host; MGS have been shown to adhere to glycoproteins and glycolipids in the host.
- Long-range adherence using pili or fibrin presented on the surface of the bacterium. These can penetrate mucus or adhere to surface proteins.
- Fibrinogen interaction using a molecular bridge between extracellular matrix proteins presented by the host and the bacteria.

Fibrils and pili, known as fimbriae, are particularly important for the growth and survival in the oral cavity [25, 6]. When the first streptococci have attached in the oral cavity, the development of a complex community of various bacteria can begin [25]. This development of the community involves coaggregation and coadherence of oral bacteria, and if undisturbed, the community can develop into a stratified, complex biofilm [13].

2.2 Mitis group streptococci as pathogens

The mechanisms described above are important for the development of surface attachment and biofilm formation in the oral cavity. However, many of these mechanisms are also known to contribute to virulence and persistence of the bacteria. Microbial infection can be a complicated affair, and it is difficult to say, which mechanism meditates the first step of the infection [56, 55]. Some of the critical steps involved in the formation of an infection of the heart are described in the following sections. Table 2.1 gives a few examples of which genes could be involved in the different mechanisms.

Table 2.1. List of a few of the genes that could be involved in the different steps of infective endocarditis: host modulation, adherence, platelet binding, and encapsulation. The table lists the overall function, the gene names, and a short description of the potential virulence genes. The table also lists which of the studies included in this thesis have identified the following potential virulence genes. Last, the table list references on the genes.

Evasion of the host immune system and colonisation				
Overall function	**Genes involved in the process**	**Description**	**Identified in**	**References**
Ig binding and cleavage	*igA*	Interacts and modulates the immune system	Paper 2 & 3	[57, 58]
Adhesion	*psaA*	Contributes to bacterial adherence and virulence	Paper 2, 3, & 4	[59]
	pavA	Adherence to epithelial and endothelial cells	Paper 2, 3, & 4	[60, 61]
	lmb	Meditates attachment to human laminin	Paper 2, 3, 4, & 5	[62]
Platelet binding	*eno*	Binding plasminogen and prevents phagocytis	Paper 3 & 4	[25]
	sspA	Adhesion and platelet agregation	Paper 3	[63]
Capsulation	CPS locus genes	Prevent phagocytosis by blocking interactions between phagocytic cells and the bacteria	Paper 2, 3, 4, & 5	[64, 65]

2.2.1 Adherence

Bacterial adherence to the host tissue often results in tissue malfunction or destruction and is a critical step in the pathogenic process [55]. The adher-

ence is often meditated by bacterial surface-exposed proteins [55]. Especially fimbriae are considered as potential virulence factors as they meditate the adherence between the bacterial cell and the host tissue. In different studies, pili have been associated with mediating adhesion to host tissue as lungs, tonsils, intestines, and human endothelial cells [25]. Additionally, the ability to bind the fibronectin has been associated with colonization of the oral cavity as well as platelet adhesion.

2.2.2 Platelet Adhesion

Platelet adhesion have shown to be an important virulence factor for IE as it facilitates the binding to damaged heart tissue [56, 66, 67]. Platelets, also called thrombocytes, are small cells produced in the bone marrow and are a component of blood [68]. Platelets circulate in the blood vessels and interact with leukocytes and endothelial cells, and function as an important inflammatory sensor [68]. If blood vessels are damaged, the platelets are activated and initiate a coagulation cascade [68]. The adherence of streptococci to platelets is an important step in the colonization of damaged heart valves [25, 3]. The bacteria can bind directly to surface receptors of the platelet, or indirectly by using proteins, which bind both the platelet and the bacteria (this is usually occurring with plasma proteins) [66]. When the bacteria have attached the platelet, the platelet can respond by activation and the secretion of their granule content [66]. The secreted proteins, e.g., cytokines, play a key role in virulence as they meditate endothelial damage [66]. The activation of the platelets by MGS can also lead to the formation of blood clots in the blood vessels as a result of the released cytokines [25, 1].

2.2.3 Host Modulation

As the MGS have entered the bloodstream, they are exposed to the immune system of the host. IgA is an immunoglobulin that is an important part of the human immune system, where it provides antibody defense of the mucosal surfaces [57]. Several MGS species have shown to contain Ig-binding proteins, which enables them to interact and modulate the immune system [69]. Some MGS can cleave IgA receptors of the host immune system. This mechanism enables the bacteria to escape detection and hinder an immune response provided by the immunoglobulins [25, 3, 57]. In addition to this, some species of MGS are found to be relatively tolerant of different antimicrobial peptides [3]. They are furthermore able to modulate the expression of interleukin-8, which

is a proinflammatory chemokine [3]. Periodontal disease often involves tissue destruction, which is largely mediated by the host inflammatory response, and not by the oral pathogen itself. All these mechanisms are crucial factors of the virulence of the MGS [25, 3].

2.2.4 Encapsulation

The capsular polysaccharide (CPS) is another important factor for the evasion and modulation of the immune system. The encapsulation has shown to protect the bacteria against phagocytosis and introduce a survival advantage in human blood [64]. CPS is furthermore highly associated with virulence, as nonencapsulated strains of *S. pneumoniae* have shown to be almost completely avirulent [65]. More than 90 serologically distinct pneumococcal have been recognized in *S. pneumoniae* [70]. As *S. pneumoniae* and the rest of the MGS species are highly related, some of these CPS genes are also found in the oral species [64], including the species investigated in Paper 2 and Paper 3 of this thesis. In addition to this, it is shown that *S. mitis* can uptake and integrate CPS genes donated from *S. pneumoniae* [64]. This genetic competence is an important virulence factor and is described in further detail in the next section.

2.3 Horizontal Gene Transfer

As described in the sections above, MGS have a reservoir of different mechanisms enabling them to cause an infection. Additionally, MGS have shown natural genetic transformability; they are able to take up as well as release DNA [71, 72]. There are in general three different ways a bacterium can take up and integrate DNA into its chromosome [11]:

- Cojungation; direct transfer doner and recipient cells in a DNase intensive manner. Plasmids or conjugative transposons often transfer the DNA.

- Transduction; transfer of genomic DNA by a bacteriophage. The bacteriophage package the host DNA into the head of the phage and then injects the DNA into another bacterial cell.

- Transformation; uptake of extracellular DNA from the environment. The DNA is often released from dead bacterial cells.

The genetic competence of streptococci has shown to be an important mechanism for acquiring genes involved in biofilm formation, adherence, and resistance to host immune systems and therefore contribute to their virulence potential [72, 73, 11, 64]. The gene exchange across the related species in the MGS, as well as the bacterial species in the surrounding biofilm, ensures a diversity within the individual species [74]. Gene exchange of virulence genes between *S. mitis*, *S. oralis*, *S. infantis* and *S. pneumoniae* is well known, and it is believed that the gene exchange gives the commensal pathogens an advantage in their adaptation to the host econiches and modulation of immune responses [64, 74, 11]. Several studies have suggested that genetic recombination has a greater importance for the pathogenic potential of a bacterium, than mutations of specific genes [75, 76, 77]. The MGS are known to have an open pan-genome, as several new genes are added to the total gene pool when an additional strain is sequenced [23, 24] (see Section 3.8 for more information). The adaptation of new genes and gene exchange between the related streptococci species contribute to the accessory genes in the open pan-genome. The inter-species recombination can also lead to a sort of inter-mediate species [78]. The genetic competence and high degree of genomic flexibility of MGS may also explain some of the taxonomic complexity within this group, due to the flexibility of their genomes [26, 78, 11].

DNA - Biology and Technology

In 1953, Watson and Crick solved the three-dimensional structure of DNA, working from crystallographic data produced by Rosalind Franklin and Maurice Willkins [79]. DNA, or deoxyribonucleic acid, is built from nucleic acids. The nucleic acids are composed of modular units of a sugar-phosphate backbone, each attached to a molecule called a nucleobase, together called a nucleotide. There exist four different nucleotides: adenosine, tyrosine, glutamine, and cytosine (Figure 3.1 A). The order and composition of these four different nucleotides make up the genetic code and contain the biological information of all living organisms [80].

Until 1965, it was not possible to read a nucleic acid sequence. In that year, Robert Holley and colleagues produced the first whole nucleic acid sequence of tRNA from *Saccharomyces cerevisiae* [79]. From 1965 till now, the sequencing technology has undergone a tremendous change, moving from sequencing short oligonucleotides to millions of bases [79].

3.1 The Era of DNA Sequencing

Today it is possible to sequence a whole genome of a single organism, sequence a whole environmental sample containing a large span of different species, and even produce real-time sequencing from a single cell [79]. The rapid development in sequencing technology has lead to the emergence of several competing sequencing companies [83]. This competition and the increased demand for DNA sequencing have led to a decrease in sequencing cost [79]. A widely used sequencing technology is the high-throughput sequencing technology (also called next-generation sequencing (NGS)), which can sequence a large amount of DNA in a relatively short time. The data of this thesis are sequenced with Illumina, the largest company in the sequencing industry, which has a large

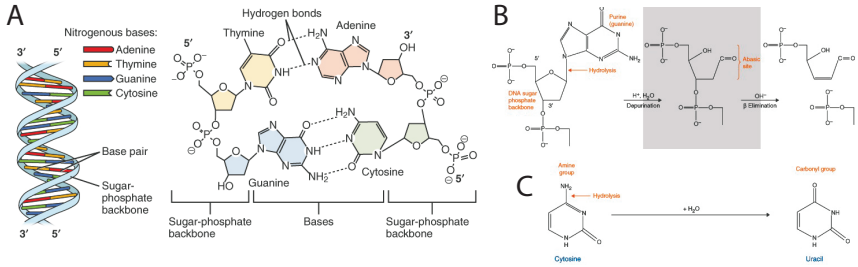

Figure 3.1. Visualization of the DNA structure and DNA damage. A) The DNA is composed of two complementary strands. The double helix and the four different nucleotides: adenosine, tyrosine, glutamine, and cytosine, which make up the DNA molecule. B) depurination of DNA, which causes fragmentation. The *N*-glycosyl bond between a sugar and an adenine or guanine residue is cleaved, which causes an abasic site. The β elimination then results in a fragmentation of the DNA strand. C) Deamination of cytosine to uracil leads to the miscoding in ancient DNA. The deamination results in a C to T substitution and a G to A substitution on the complementary strand. Figure A is from OpenStax [81], Figure B and C is from Dabney et al. [82], license number: 15702006661738057.

market share [84]. The Illumina sequencing workflow is visualized in Figure 3.2, and can, in general terms, be explained in the following steps:

- **Library Preparation**: The sequencing libraries are typically created by fragmenting the DNA and adding specialized adapters to both ends of the DNA. The adapters contain complementary sequences that allow the DNA fragments to bind to the flow-cell [83].

- **Cluster Generation**: The prepared libraries are loaded into a flow-cell and placed on the sequencer. The clusters of DNA fragments are amplified in a process called cluster generation, resulting in millions of copies of double-stranded DNA. Clusters are denatured, the reverse strand is washed away, and sequencing by synthesis begins

- **Sequencing by Synthesis**: this process is initiated by binding primers with the adapter sequences. The primers are designed to be complementary to the adapter sequences. A polymerase and a mixture of four differently colored fluorescent reversible dye terminators are added, where a reversible terminator blocks incorporation of the next base. The fluorescent signal indicates which nucleotide has been added, and the terminator is cleaved so the next base can bind [79, 83].

- **Paired-end Sequencing**: for double-stranded DNA, it can be an advantage to sequence both the forward and reverse DNA strands. After the sequencing of the forward strand, the sequencing product is washed away. The DNA fragment then folds over and bind the flow-cell, forming a bridge. DNA polymerase then extends flow-cell, forming a double-stranded bridge. The two strands are separated, and the original forward strand is removed. The reverse read is then obtained by sequencing by synthesis, as described above. To ensure that the read pair do not overlap, the DNA fragments are prepared, so the fragments are longer than the sequenced reads. The length of the DNA fragment is also called the insert size. The paired-end reads, and the insert size is visualized in Figure 3.3 1.

- **Output**: The final output is based on base-calling on the processed emission wavelength of the fluorescent nucleotides. The base-calling can be reformatted into a text file containing the DNA sequences as raw reads. The raw reads text file is also called a fastq file. Each sequencing read has four lines in a fastq file: a header (begins with a '@'), the DNA sequence, a '+', and then the encoded quality score of each sequenced nucleotide. Each quality symbol corresponds to a quality value, and the quality value encoding depends on the sequencing technology.

The data in this thesis is based on single isolate sequencing and metagenomic sequencing, which will be explained briefly in the following sections.

3.2 Isolate Sequencing

The sequencing of a single organism comes in handy when investigating the genome of one specific organism. It is important to mention that it is not sequencing of a single cell, but many clones of the same bacterium. In this thesis, we worked with whole-genome sequences of bacteria that cause infective

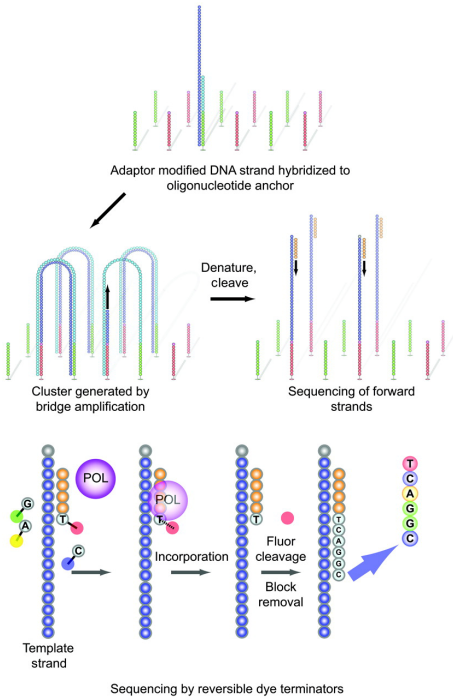

Figure 3.2. Scheme of Illumina DNA sequencing. The fragmented DNA is bound to the flow-cell by hybridization. Bridge amplification generates amplified clusters generation. Clusters are denatured and cleaved; sequencing by synthesis is initiated with addition of primer, polymerase (POL) and 4 reversible dye terminators. Reprinted from Voelkerding *et al.* [83], with permission from publisher.

endocarditis. The principle was to extract a blood sample containing bacteria causing an infection, inoculating and harvesting them. DNA was extracted from the bacterial cells and sequenced as described above (section 3.1).

3.3 Metagenomic Sequencing

Metagenomics is the study of DNA sampled from the environment or a host. Using the metagenomic approach, all DNA of the community is sequenced, where all the organisms of the community are sequenced in parallel. Over the last decades, various environments have been metagenomic sequenced: marine environments, soil, fermentation reactors, and specific niches within the bodies of humans and animals [85].

The DNA from an environmental sample is sequenced without any culturing in the lab. This means that a great portion of bacteria that is uncultivable can be sequenced, while the majority of RNA viruses will remain unsequenced [86, 87]. The DNA is extracted from the sample using a DNA extraction kit and then sequenced as described in section 3.1. There has been considerable research into techniques of how to reconstruct whole genomes from metagenomic sequencing data, this can, for example, be done by metagenomic binning without any reference genomes developed by Nielsen *et al.*, [88]. Examples of available metagenomic binning tools are MetaBAT [89], CONCOCT [90], and VAMB [91].

3.4 Ancient DNA

The introduction of metagenomic sequencing has made DNA sequencing easier for organisms that can not be cultivated [86]. DNA is termed ancient DNA when it is extracted from ancient long-dead individuals, animals, or microorganisms. This provides an unique insight into ecological and evolutionary processes through time [92, 93, 94]. The metagenomic sample included in this thesis is of ancient origin.

When an organism dies, its DNA slowly becomes degraded [95]. Under favorable conditions as low-temperature and high salt concentration, the ancient DNA can be preserved, and the degradation process can be stopped [82, 95]. However, the DNA will still be affected, as a result of post-mortem DNA decay. Over time the DNA destabilizes and breaks - leading to small DNA fragments [95]. A likely cause of the destabilization is depurination, in which the N-glycosyl bond between a sugar and an adenine or guanine residue is cleaved (Figure 3.1 B). Ancient DNA is also hydrolyzed over time, which results in the deamination of cytosine to uracil (Figure 3.1 C). The cysteine to uracil substitution leads to a misencoding under sequencing, where the complementary nucleotide for uracil is tyrosine. Ancient DNA, therefore, has a considerable

amount of C to T substitutions [82]. The complementary substitutions of G to A is like-vise observed, in paired-end sequencing of ancient DNA [82]. The fragmentation and deamination rates are deepened on several factors; low temperature, stable pH, or encapsulation of the DNA can slow down the damage process [82, 94, 95]. Besides the issues of fragmentation and deamination in ancient DNA, there is also the factor of environmental contamination of the sample. It is, therefore, essential to check for modern (non-damaged) DNA in the sample.

3.5 Prepossessing of Sequencing Data

As described in Section 3.1, non-biological sequences, adapters are added to the DNA fragments during the library preparation step. When the DNA is sequenced, those adapters become a part of the sequenced data. It is, therefore, highly important to remove those non-biological sequences, as they are not a part of the biological sample, and they can be highly prevalent. It is furthermore important to check the quality of the final raw reads. Low quality in sequencing can mean the introduction of sequencing errors [96]. The general threshold of sequencing quality is a Phred quality score of 20, corresponding to a probability of error of no more than 1 percent [97]. Two other important measures to validate in NGS data, is the sequencing depth, also called the depth of coverage, and the genome coverage also called the breadth of coverage [96]. The depth of coverage represents the average number of times that each position in the sequenced genome is covered by a high-quality base read [96]. The depth of coverage can be determined by aligning the sequence reads to a representative reference, or it can be calculated based on the assumed genome size by the equation [96]:

$$Depth\ of\ coverage = \frac{L * N}{G} \tag{3.1}$$

where L is the read length N is the number of reads and G is the expected genome size.

The breadth of coverage, on the other hand, is a measure of the percentage of the genome that is sequenced a given number of times [96]. For example, can a genome have a breadth of coverage of 95% with a at a minimum depth of coverage of ten reads. The breadth of coverage is calculated by aligning the sequencing reads to a representative reference and calculate the percentage of covered positions of the genome.

When working with ancient data, it might be an advantage to remove the first and last two - to - five base-pairs, since the majority of the DNA damage is located in the ends of the DNA fragment. The DNA damage patterns can be evaluated using mapDamage [98].

3.6 *de novo* Assembly

When the raw data is preprocessed, the fragmented DNA, represented as reads, can be assembled into long stretches of DNA, which represents the genome. The assembly itself can be described as a giant puzzle with millions of pieces [99]. Therefore, it is impossible to assemble a genome manually and requires advanced mathematical models and a substantial amount of computer power [99]. The general workflow of the *de novo* assembly is to identify shared regions of overlapping reads. The overlapping reads are pieced together into longer stretches of DNA, called contigs (contiguous sequences). If the reads are paired-end sequences, the contigs can be assembled further using library insert size information [99]. The assembled contigs result in longer stretches of sequences, called scaffolds (Figure 3.3 2).

Several assembly algorithms have been developed; however, for short sequence Illumina data, the majority has been based on the de Bruijn graph approach [99]. This approach uses k-mers to build a graph of overlapping sequences (Figure 3.3 3). The reads are split into smaller pieces of $L - k + 1$ k-mers, where L is the read length, k is the k-mers size. The k-mers are connected if the two k-mers completely overlap except for one nucleotide at each end [100]. By going through the optimal path in the graph, the de Bruijn graph is transformed into sequences. The choice of k-mer value can have great effects on the construction of the de Bruijn graph and the resulting assembly: small values of k will result in a more tangled graph due to more repeats will be collapsed together. Large values of k may in low coverage regions fail to detect overlaps of reads resulting in a more fragmented graph [101]. Some assemblers, such as SPAdes [101], can construct a multisized de Bruijn graph using a varied set of k-mer-values. The final assembly represents the whole genome of the isolated bacterium or multiple species in a metagenomic sample. This assembly can, among others, be used for further analysis, such as gene prediction and phylogenetic reconstruction.

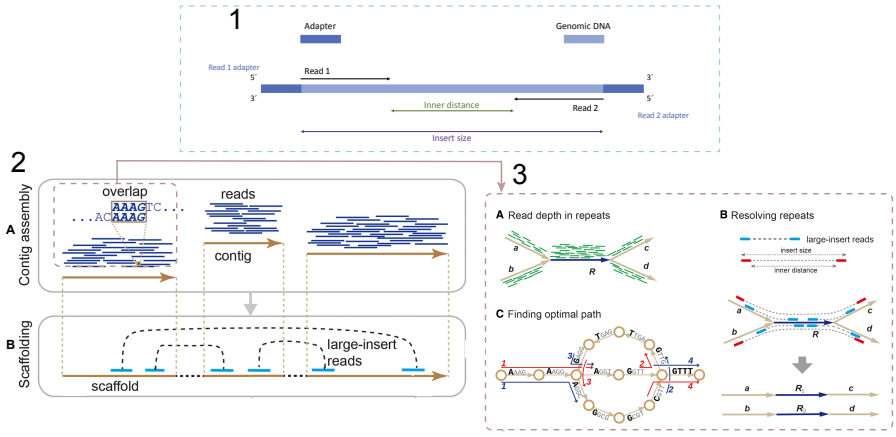

Figure 3.3. The general workflow of a *de novo* assembly of a whole genome. 1) illustrates the principle of insert size in paired-end sequencing. 2A) the reads are overlapped to construct the contigs. 2B) Using insert size information, the contigs can be assembled into scaffolds. 3) visualization of a de Bruijn graph that is used to assemble the contigs. 3A) shows an example of a small genome, 3B) each read from the genome, represents a node in the graph, and the edges represent the alignments between the reads. 2C) the reads are split into all possible k-mers, 3D) the assembly algorithm constructs a de Bruijn graph by representing all k-mer prefixes and suffixes as nodes and then drawing edges that represent k-mers having a particular prefix and suffix. Figure 3.3 2 is from Sohn *et al.* [99] with modifications, Licence number: 4694661395690, publisher tax id: GB125506730. Figure 3.3 3 is from Compau *et al.* [100].

3.7 Molecular Phylogenetics

As DNA sequencing has become more and more widely used, phylogenetics is applied in almost every branch of biology [102]. Phylogenetics can be used to show the relationship among species, between paralogs in a gene family, to describe histories of populations and evolution. Phylogenetics can furthermore

be used for tracking the epidemiological dynamics of pathogenic outbreaks or a reemerging infection in a patient [50]. It is also widely used in metagenomics, to classify the sequences of the sample [102]. In this thesis, phylogenetics was used to classify different species of Mitis group streptococci (MGS), as well as to identify the phylogenetic relationship between disease-causing bacteria and bacteria isolated from healthy individuals. The following section will, therefore, focus on bacterial phylogenetic reconstruction.

A phylogenetic tree contains nodes and branches (Figure 3.4). Nodes are connected by branches. There are two types of nodes in a phylogenetic tree: Internal nodes, which correspond to ancestral taxa, and external nodes, which correspond to the observed taxa investigated in the respective study. Each branch represents a separation event from the previous node [102]. The length of a branch typically represents the number of changes (substitutions) that have occurred between the two connected nodes. For rooted trees (as illustrated in Figure 3.4, the total distance is calculated by summing up the horizontal branch-lengths between two external nodes of interest [103], for other trees the, the total distance between two external nodes of interest, is calculated by summing up all branch-lengths. Phylogenetic trees can be rooted or unrooted. For rooted trees, the root represents the oldest point in the tree and corresponds to the theoretical common ancestor [104]. It is therefore required to have some ancestral knowledge about the taxa of the phylogeny, to be able to infer rooted phylogeny. An outgroup can represent the root of the tree, if it is less related to all the ingroup taxa, that the ingroup taxa are related to each other. If such taxa are unknown, or not a part of the data, the tree can be rooted in the middle of the tree, also called midpoint rooting. The phylogenetic tree can also be represented without a root: unrooted phylogeny, which can be beneficial then reconstructing the phylogeny of related species [102, 104].

3.7.1 Phylogenetic Approaches

The reconstruction of phylogeny can be rather complicated and computationally demanding [103]. Several mathematical models and algorithms have been developed and can be used for different approaches to infer phylogeny. One of these statistical models is called maximum likelihood, which is widely used, as it gives good models of molecular evolution [102]. As the name suggests, the maximum likelihood is a method that estimates parameters of a probability distribution generates a maximized likelihood function resulting in a model with the highest probability [102, 105]. Generally, maximum likelihood tree estimation requires two optimization steps: optimization of branch-length to calculate

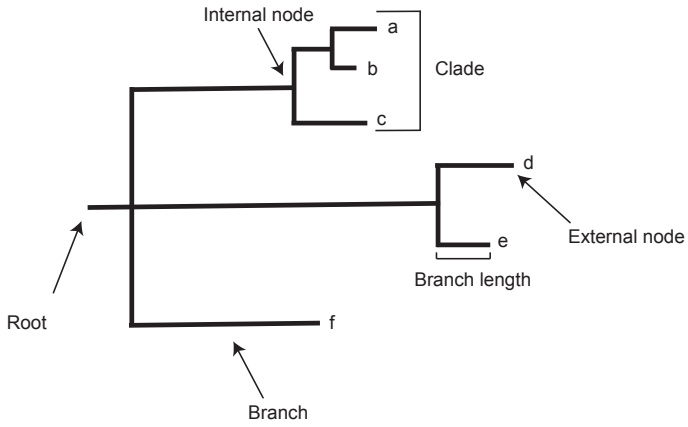

Figure 3.4. Phylogenetic tree illustrating the placement of an root, branches, internal - and external nodes.

the tree score for each candidate tree, and a search in the tree space for the maximum likelihood tree [102]. The model parameters can, among others, include nucleotide frequencies, nucleotide substitutions rates, tree topology, and branch-lengths. Different substitution models can be specified for the probability calculations. However, most of those assume independent evolution of sites in the sequence, meaning that the likelihood is the result of probabilities for different sites [102]. Available tools which uses maximum likelihood to generate phylogeny include RAxML [106], PhyML [107], and GARLI [108]. The input data is often given by an alignment of either DNA or amino acid sequences. These alignments can, among others, be generated by MUSCLE [109]. The input alignments are often based on 16S rRNA sequences or alignments of one of several genes of evolutionary interest (also called marker genes). Even tough 16S rRNA gene sequences can be highly useful in regards to bacterial classification, it has low statistical power at the species level and often fails to separate closely related taxa [110, 111]. The more genes that are included in the phylogenetic analysis, the more statistical power is there in the final phylogenetic tree. Recent studies have therefore shifted towards using core-gene phylogeny or whole-genome phylogeny [110, 112, 50]. By reconstructing

phylogeny based on whole-genome assemblies or reads, the full genomes can be used for the single-nucleotide variant (SNV) calling. The online web tool CSI phylogeny has been developed to infer phylogeny based on whole genomes to identify infectious disease outbreaks [50]. Different approaches have been used to reconstruct core-gene phylogeny, among others ortholog gene sequence clustering or clustering based on gene functions [113, 114, 115, 112]. In this thesis, the pipeline PanFunPro [115] was used to identify the core-genome of species within MGS strains. The pipeline predicts genes in the assembled genomes, identifies functional domains on amino acid levels, and group them into protein families, based on the functional profile. The core-genome can be identified by using protein families. The sequences of those protein families were aligned and used for phylogenetic reconstruction. In the next section, functional domains, and the principle behind a pan - and core-genome is briefly described.

3.8 The Pan - and Core-Genome

As described above, the core-genome can be used for reconstructing phylogeny. As the name suggests, the core-genome consists of core-genes. A core-gene is a gene that is present in all analyzed samples in the data. Biological datasets often consist of multiple samples, each containing a whole-sequenced genome. The pool of genes from all samples represents the pan-genome. Within the pan-genome, the core-genome can be identified (Figure 3.5). The core-genome can be identified using sequence similarities by aligning sequences using BLAST [116], detecting orthology within genes using OrthoMCL [114, 23] or clustering genes using CD-hit [117]. Identifying homology between genes can be of great importance in understanding the functionality and evolution of the genes. However, sequence similarities do not necessarily imply that the proteins perform the same biochemical function; all sites in a protein sequence are not equally important and, therefore, not equally conserved [118, 119, 120].

In this thesis, functional domain architectures were used for the characterization of the complete genomic reservoir (the pan-genome). To establish as well as validate the core-genome, several genomic comparison approaches were applied on the pan-genome: identification of functional domain architecture, gene-sequence clustering, and multiple alignments of similar genes. As functional domains are units of a protein, the gene sequences were translated into amino acid sequences. These will be referred to as protein sequences, even though they originate from whole-genome sequencing and not protein sequencing.

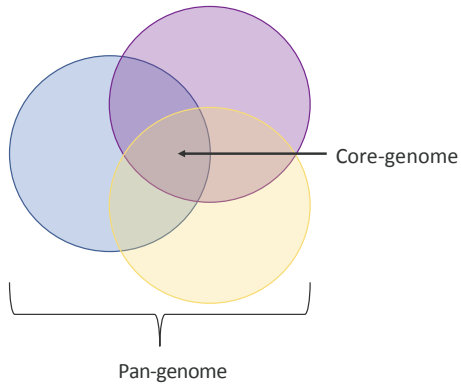

Figure 3.5. Visualisation of the pan - and core-genome of three genomes. The center where all three circles overlap represents the core-genome. The pan-genome represents the total gen-pool, including the core-genome.

3.8.1 Functional Domains

When comparing multiple homologous protein sequences, some regions in the sequence are more conserved than others [118, 119]. These conserved regions are often referred to as protein domains, which are fundamental units of the structure and evolution of the proteins [121, 122]. A protein can contain one or more domains, and the domain architecture has great importance for the tertiary structure and, therefore, also the function of the protein [123]. Each functional domain is evolutionary and functional independent and can fold into its own stable, compact tertiary structure or fold [122]. One shared protein domain across different genes, does not necessarily mean that the genes carry out the same protein function [124]. It is the collection and composition of the functional domains that should classify the overall function of the protein [124, 125]. In this thesis, the proteins were divided into protein families based on domain architecture. Thus, a protein family consisted of protein sequences that all contained the same functional domains and in the same order (Figure 3.6). The functional domains were identified and divided into protein families using the pipeline PanFunPro [115]. In the pipeline, the functional domains were

predicted based on the translated gene sequences using three Hidden Markov Models (HMM) collections: PfamA [126], TIGRFAM [127], and Superfamily [128]. The translated gene sequences were scanned against the HMM collection using InterproScan software [129].

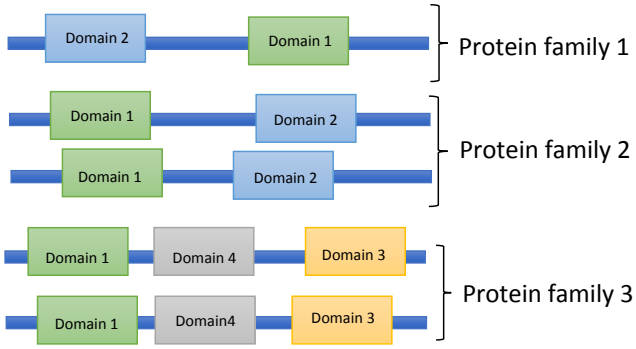

Figure 3.6. Visualisation of functional protein domains present in a gene. Each colored box represents a protein domain in a gene, and the color indicate the different domain families. Based on the domain content and order, each gene can be assigned into a protein family.

3.9 Machine Learning

Machine learning is a term for techniques and models designed to identify and learn patterns from data. These techniques are not bound to biological data but are applicable across various purposes and fields of science, such as economics, text mining, and image recognition [130, 131, 132]. Machine learning is now a part of our everyday life, as models, among others, can filter our emails and detect spam [133]. We meet machine learning every time we do web-searches, and machine learning has even enabled the introduction of self-driving cars [130].

Genomics is one of the most important domains in bioinformatics, and the number of sequences available has been increasing exponentially and keeps on rising [134, 135]. This huge amount of biological data has introduced a whole new problem: how to handle all that data and how to extract the essential information to produce useful results [135, 134]. The combination of the availability of enormous computational power and large datasets has increased the interest in and the development of machine learning [134]. The general principle in machine learning is to recognize patterns using features and different mathematical models. Features are usually numeric, and each feature explains the data. Based on the specific patterns that are found in the features, machine learning models can predict the data to belong to different categories or clusters, depending on a supervised or unsupervised learning approach. In this thesis, supervised learning was applied to the model, as the input data was labeled.

When using machine learning on biological data, it is often computational modeling of biological networks. In this thesis, Random Forest modeling was applied to genomic data in an attempt to separate bacteria isolated from patients with an infection from the same bacterial species isolated from healthy individuals. [130].

3.9.1 Random Forest

Random Forest is a robust and powerful classifier which is widely used for bioinformatics [136]. As the name suggests, Random Forest consists of a series of decision tree models. A decision tree is a graph which consists of nodes and branches. Each node represents attributes in a group that is to be classified, and each branch represents a value that the node can take. Each tree will make a decision, and the graph will expand until some criteria are met. That criterion could be a maximum in-depth or until the input data is well separated. The final decision of the model is a majority vote for the collected trees (Figure 3.7) [137, 136]. Random Forest is also known as an ensemble classifier. An ensemble is a method consisting of multiple sub-models, results of which are aggregated into a final decision. In the Random Forest approach, each decision tree is a model, and all the trees have a very low correlation with each other as it is given different features to model on. This makes the Random Forest machine learning approach very powerful as uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. In some cases, a prediction in a tree will be wrong, but because of the low correlation, the errors from one tree will not affect the predictions in the

other trees. However, for this to be true, an actual signal in the input features must be present, and predictions made by the individual trees need to have a low correlation.

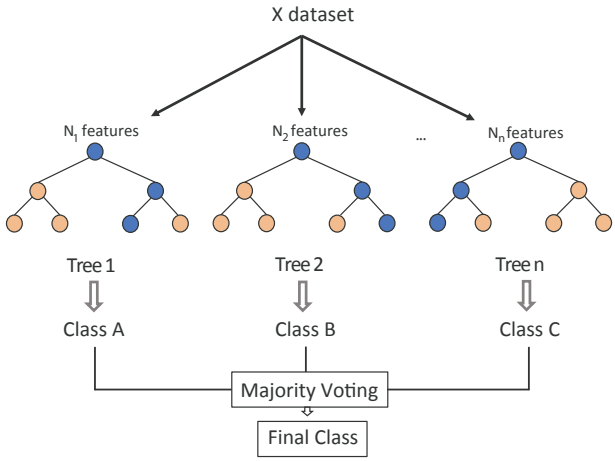

Figure 3.7. A simplified visualization of a Random Forest model with multiple decision trees. The tree is drawn upside down, with its root at the top. The root represents the full dataset, which is split into different sample subset, using a chosen feature that best separates samples. A majority vote describes the final prediction of the classification.

3.9.2 Redundancy

When using machine learning to solve biological questions, it is important to prepare and clean the input data: remove redundancy and encode non-numerical data points. If some features are very similar, or the features are binary encoded, redundancy can cause problems as the importance of the features can be misleading. The accuracy performance of the model might not be affected, but the computational time will increase the more redundancy there are in the data [138, 139].

3.9.3 Overfitting and Cross-validation

As described in the section above, it is essential to prepare the input data. It is furthermore crucial to avoid overfitting in the model. When the model is trained, it is in its learning phase, where it should be able to recognize patterns and categorize correctly on data that differ from the data it has been training on. Overfitting is when a model has been trained too well and is therefore not able to create a generalized pattern recognition, which consequently makes it lose its predictive power for unseen data [140, 139]. An overfitted model tends to have very high accuracy; however, it will not be able to perform well and generalize on previously unseen data, which results in unreliable results. Therefore the input data needs to be shuffled and divided into partitions for test and training [140, 139]. A model may never be tested on data that has been used for training. The shuffling is performed to ensure that each partition in the training and test set is well balanced; in this thesis, we used stratified partitions, which means that each partition contains approximately the same proportion of labels as in the original dataset [141].

Cross-validation is often used to detect overfitting. Here is the data divided into K-folds or partitions of approximately equal size (Figure 3.8). The model is trained on each of the $K-1$ folds, while the remaining fold is used for testing or validating the model. The training and testing are repeated K times; thus, each partition has been subject to validation. The test error can then be estimated by taking the average test error across K trials. In some cases, leave-one-out cross-validation is used. This cross-validation follows the same principle as K-fold cross-validation, but here is the number of folds equal to the number of data points in the dataset.

3.9.4 Performance

There are several ways to evaluate the performance of a classifier; in general, a prediction can be correct or wrong when classified on a test set. The correct and wrong predictions can then be divided into four categories [142]:

- True positives (TP): when the model correctly predicts the positive class.
- True negatives (TN): when the model correctly predicts the negative class.
- False positives (FP): when the model incorrectly predicts the positive class.

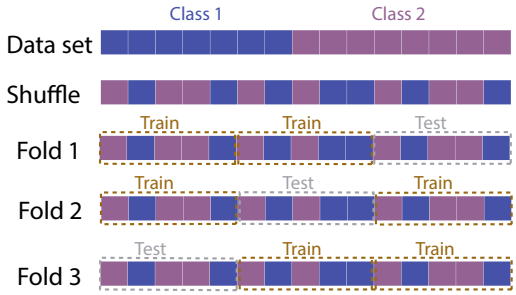

Figure 3.8. An example of 3-fold cross-validation of a reasonably well-balanced dataset. In this case, each square represents a data point or a sample. The data is shuffled in the first step as the order of the collected information seems biased. This bias is reduced by randomly shuffling the data. The shuffled data is then split up into three partitions, with approximately the same number of blue and purple data points - this is called stratified partitions. In each of the cross-validation folds, two of the partitions are trained, and one is tested.

- False negatives (FN): when the model incorrectly predicts the negative class.

Using the TP, TN, FP, and FN's different ratios can be calculated, which van explain the overall performance of the given classifier.

3.9.4.1 Accuracy

Using these prediction categories, the prediction accuracy of the final classifier can be calculated. This accuracy measurement is done by calculating the ratio of all the correct predictions to all the data points. For N data points the equation is [142]:

$$Accuracy = \frac{TP + TN}{N} \tag{3.2}$$

The accuracy score is between 0 and 1 and is an intuitive explanation of the performance, as it is the ratio of correct predictions (or percentage if multiplied

with 100) when tested on unseen data [142]. In the case of $TP + TN \approx N$, the model is close to perfect, which could imply an imbalanced dataset or overfitting.

3.9.4.2 ROC-curves and AUC

The Receiver Operating Characteristics (ROC) curve is a visual representation of the sensitivity and the 1−specificity [143]. The sensitivity is the ratio between the positive correctly classified samples and the total number of positive samples [142]. This ratio is also given as the True Positive Rate (TPR), and is given by [142]:

$$TPR = \frac{TP}{TP + FN} \tag{3.3}$$

The specificity is the ratio between correctly classified negative samples to the total number of the negative samples [142]. 1− specificity is also the calculation of the False Positive Rate (FPR), which is given by [142]:

$$FPR = \frac{FP}{TN + FP} \tag{3.4}$$

The ROC curve can be plotted, by calculating the TPR and FPR for different probability thresholds of a classifiers predictions [142]. By calculating the integral of the ROC curve, the Area Under Curve (AUC) is given. The AUC is a good evaluation measure of how well the model can separate the data into the correct classes; an AUC of ≈ 1 indicates a clear separation and a perfect performance of the classifier. At an AUC of ≈ 0.5, the model has no discrimination capacity to distinguish between the positive and negative class and is therefore as good as random guessing [142].

3.9.4.3 MCC

The Matthews Correlation Coefficient (MCC) is another metric that is useful for evaluating performance. The MCC is calculated by[142]:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{3.5}$$

And results in a number between -1 and 1, where -1 is an indication of anti-correlation (e.g. that it predict all 0 as 1 and vise versus). A MCC of 1 is an indication of an perfect classifier, while 0 indicates an uninformed (ramdom) classifier.

3.9.4.4 Small and Imbalanced Datasets

Random Forest does not take into account the possible class imbalance of the dataset. This can be an issue as class imbalance is known for increasing the bias towards the majority class [136]. Neither the accuracy of the AUC evaluation metrics account for imbalanced datasets, and can, therefore, be misleading. The MCC metric, on the other hand, works well with imbalanced datasets [144].

Research Objectives

This PhD thesis has the overall aim of understanding how oral commensal bacteria with only few identified toxins and virulence factors can be a successful pathogen in the human host.

To gain a deeper insight into this issue, the following hypotheses were formulated:

1. Can reconstruction and comparison of molecular phylogeny using different methodologies clarify the most-well suited approach for species identification?

2. Do pathogenic strains have specific genetic traits that separate them from non-pathogenic strains?

3. Is it possible to identify potential virulence and resistance genes in the Mitis group streptococci, *A. urinae*, and *A. sanguinicola* that can explain their ability to cause severe infections such as infective endocarditis?

4. Can the evolution of *Streptococcus* species be elucidated through the analysis of ancient data?

Paper 1, 2, and 3 are based on a dataset of 80 Mitis group streptococci collected from patients with verified infective endocarditis (IE). The strains were collected in the period 2006-2013 at three different Hospitals in Denmark: Capital Region of Denmark, Region Zealand, and Region of Southern Denmark. These three papers address Hypotheses 1, 2, and 3 by reconstructing the phylogeny of the 80 streptococci strains based on different methods (Paper 1). The second paper went into detail of several genes known to be associated with the virulence. This paper also applied hierarchical clustering on the pan-genome of strains belonging to *Streptococcus mitis* and *Streptococcus oralis* to investigate the different sub-groupings of *S. oralis* that was observed in Paper

1. Paper 3 focused on comparative genomics of pathogenic strains isolated from patients with IE and potentially pathogenic strains isolated from healthy individuals by the use of phylogenetics, clustering analysis, and random forest modeling.

Paper 4 is based on a unique dataset sampled from a 5,700-year-old pitch of chewed birch. From the birch pitch, it was possible to recover and sequence ancient DNA of the human genome and oral microbiome of the person that had been chewing on the birch pitch. The aim of this study was to show that these chewed birch pitches could help filling out some of the historical gaps that exist because of lack of human bone from specific time periods. The aim was also to test Hypothesis 4, by comparing ancient oral streptococci species with modern, to elucidate how, and if, the species have evolved through time.

Paper 5 is based on strain collection of 48 *Aerococcus sanguinicola* and *Aerococcus urinae*, all isolated from patients with verified IE, urinary tract infections, or sepsis. The strains were collected over two different time periods; 1984-2004 and 2010-2015. The aim of this study was, to test Hypothesis 3 by identifing virulence genes that could be associated with human disease. The study furhtermore aimed to gain a deeper insight into the pan - and core genome of the two species, using phylogenetic reconstruction and hierarchical clustering.

Paper 1

Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of Mitis group streptococci

Louise H. Rasmussen, Rimtas Dargis, **Katrine Højholt**, Jens Jørgen Christensen, Ole Skovgaard, Ulrik S. Justesen, Flemming S. Rosenvinge, Claus Moser, Oksana Lukjancenko, Simon Rasmussen, and Xiaohui C. Nielsen
European Journal of Clinical Microbiology & Infectious Diseases,35(10): 1615–1625, October 2016.

It is well known that the species identification of Mitis group streptococci can be especially challenging. It is furthermore essential to make the correct species identification in cases of life-threatening infections such as infective endocarditis. In this paper, we compare different molecular phylogenetic approaches to investigate how these methods perform with regard to accuracy of the species identification of 80 clinical strains of Mitis group streptococci isolated from patients with verified infective endocarditis.

Contribution: I (Katrine Højholt Iversen) contributed to the design and development of the phylogenetic analysis parts of the paper as well as to data analysis, interpretation of results, and construction of Figure 2 and Figure 3. I, furthermore, contributed to the revision and feedback on draft versions of the paper.

Link to paper:
https://doi.org/10.1007/s10096-016-2700-2

Paper 2

In silico assessment of virulence factors in strains of *Streptococcus oralis* and *Streptococcus mitis* isolated from patients with Infective Endocarditis

Louise H. Rasmussen*, Katrine Højholt*, Rimtas Dargis, Jens Jørgen Christensen, Ole Skovgaard, Ulrik S. Justesen, Flemming S. Rosenvinge, Claus Moser, Oksana Lukjancenko, Simon Rasmussen, and Xiaohui C. Nielsen.
Journal of Medical Microbiology, 66(9): 1316–1323, September 2017.

Streptococcus mitis and *Streptococcus oralis* are both members of the Mitis group streptococci and closely related to the pathogen *Streptococcus pneumoniae*. *S. mitis* and *S. oralis* are commensals of the human oral cavity and are, in some cases, able to enter the bloodstream and cause infective endocarditis. This study investigates the pangenome of 40 *S. mitis* and *S. oralis* genomes isolated from patients with IE. Potential virulence genes that could be involved in the parthenogenesis in these two species were identified in the pan-genome. Furthermore did the hierarchical clustering of the pan-genome show, that the three, recently reassigned, subspecies within the *S. oralis* species: subsp. *oralis*, subsp. *tigurinus*, and subsp. *dentisani* could be separated using this approach.

Contribution: I (Katrine Højholt Iversen) contributed to the design and development of the phylogenetic aspect of the paper and I performed the majority of the bioinformatics work. I furthermore contributed to the analysis and interpretation of results, and constructed Figure 1. I, furthermore, contributed to the revision and feedback on draft versions of the paper.

Link to paper:
https://doi.org/10.1099/jmm.0.000573

Paper 3

Similar genomic patterns of clinical infective endocarditis and oral isolates of *Streptococcus sanguinis* and *Streptococcus gordonii*

Katrine Højholt Iversen*, Louise Hesselbjerg Rasmussen*, Kosai Al-Nakeeb, Jose Juan Almagro Armenteros, Christian Salgård Jensen, Rimtas Dargis, Oksana Lukjancenko, Ulrik Stenz Justesen, Claus Moser, Flemming S. Rosenvinge, Xiaohui Chen Nielsen, Jens Jørgen Christensen, and Simon Rasmussen.
The manuscript is currently in review in Scientific Reports.

Streptococcus sanguinis and *Streptococcus gordonii* are members of the Mitis group streptococci and commensal colonizers of the human oral cavity with the ability able to cause infective endocarditis. In this paper, we compared the pan-genomes of 38 *S. sanguinis* and *S. gordonii* strains isolated from patients with infective endocarditis to 21 *S. sanguinis* and *S. gordonii* strains isolated from the oral cavity of healthy individuals. Our genomic comparisons based on phylogenetic reconstruction, hierarchical clustering, and random forest modeling revealed no significant differences between the two isolation groups. The presence of the specific, potential virulence genes did separate the species in two distinct clusters, but the two isolation groups seemed to contain similar virulence potential.

Contribution: I (Katrine Højholt Iversen) contributed to the design, planning, and development of the study. I carried out the analysis of data and interpretation of results. I furthermore constructed all tables and figures and I wrote the manuscript.

This manuscript is currently in review in Scientific Reports.

Abstract:
Streptococcus gordonii and *Streptococcus sanguinis* belong to the Mitis group streptococci, which mostly are commensals in the human oral cavity. Though they are oral commensals, they can escape their niche and cause infective endocarditis, a severe infection with high mortality. Several virulence factors important for the development of infective endocarditis have been described in these two species. However, the background for how the commensal bacteria, in some cases, become pathogenic is still not known. To gain a greater understanding of the mechanisms of the pathogenic potential, we performed a comparative analysis of 38 blood culture strains, *S. sanguinis* (n=20) and *S. gordonii* (n=18) from patients with verified infective endocarditis, along with 21 publicly available oral isolates from healthy individuals, *S. sanguinis* (n=12) and *S. gordonii* (n=9). Using whole genome sequencing data of the 59 streptococci genomes, functional profiles were constructed, using protein domain predictions based on the translated genes. These functional profiles were used for clustering, phylogenetics and machine learning. A clear separation could be made between the two species. No clear differences between oral isolates and clinical infective endocarditis isolates were found in any of the 675 translated core-genes. Additionally, random forest-based machine learning and clustering of the pan-genome data as well as amino acid variations in the core-genome could not separate the clinical and oral isolates. A total of 151 different virulence genes was identified in the 59 genomes. Among these homologs of genes important for adhesion and evasion of the immune system were found in all of the strains. Based on the functional profiles and virulence gene content of the genomes, we believe that all analysed strains had the ability to become pathogenic.

Link to paper: https://doi.org/10.1038/s41598-020-59549-4

Paper 4

Stone Age" chewing gum" yields 5,700-year-old human genome and oral microbiome

Theis ZT Jensen*, Jonas Niemann*, Katrine Højholt Iversen*, Anna K Fotakis, Shyam Gopalakrishnan, Mikkel HS Sinding, Martin R Ellegaard, Morten E Allentoft, Liam TLanigan, Alberto J Taurozzi, Sofie Holtsmark Nielsen, Michael W Dee, Martin N Mortensen, Mads C Christensen, Søren A Sørensen, Matthew J Collins, Tom Gilbert, Martin Sikora, Simon Rasmussen, and Hannes Schroeder.

Recovery of ancient DNA has enabled the scientific world to give a unique insight into the prehistory of Man. By extracting DNA from an ancient chewed birch pitch, we were able to recover the human genome, the oral microbiome, and food leftovers from the person that had chewed the 'gum'. The ancient bacterial reads could then be compared with modern bacteria to investigate the genetic differences that could have aroused due to evolution.

As several results were excluded from the paper due to space limitations, I have chosen to include some of them in the section "Additional Results of Paper 4", which can be found right after Paper 4.

Contribution: I (Katrine Højholt Iversen) contributed to the design and development of the bacterial aspect of the paper. I performed the analysis related to *Streptococcus* and the post-mortem DNA damage profile of the hazelnut. I contributed to the construction of Figure 4 and Supplementary figure 16, and to the writing and revision of the manuscript.

This manuscript is accepted in Nature Communications, but has not yet

been published.

Link to paper:
https://doi.org/10.1038/s41467-019-13549-9

Additional Results from Paper 4

This section includes results that were not included in the accepted paper of "Stone Age 'chewing gum' yields 5,700-year-old human genome and oral microbiome". Additionally, some of the results in this section are supplementary figures or data of the paper. It will be stated clearly in the figure texts whether the following results are a part of the accepted paper or not.

As previously explained, working with ancient genomic data can be a challenging task. Furthermore, species of MGS are highly similar at the sequence level. During this study, we learned that the separation of these highly similar species in an ancient metagenomic sample was very difficult. The low sequencing coverage, DNA damage, and highly similar sequences complicated the analysis. This was the main reason why many of the following results had to be excluded from the final paper. However, the results are still interesting and hold great potential for further studies.

Edit Distance and DNA damage

In the ancient pitch, we identified several species belonging to the MGS and to evaluate which MGS species that were present in the ancient sample, we generated edit distance plots (Figure 8.1) and looked at DNA damage patterns (Figure 8.2). The edit distance plots are created by mapping the ancient reads to a reference genome. The mismatches between the reference and the ancient reads are then counted and plotted. If the expected species is present in the sample, the highest count of mapped reads would have an edit distance of 0 (0 mismatches), the second-highest count of mapped reads would have an edit distance of 1 (1 mismatch) and so forth. The edit distance plots showed that the ancient reads had a higher affinity towards *Streptococcus viridans*, but several other MGS species were also likely to be present in the sample. However, it was clear that the edit distance distributions encountered a higher amount of mismatches than expected, which was likely due to the high degree of sequence similarity between closely related species.

The DNA damage plots give a reasonable estimation of the identified species is present in the ancient sample, and if the ancient reads are, in fact, ancient. When looking at the DNA damage patterns from the 5' (Figure 8.2), we could define that all investigated species presented post-mortem DNA damage. Based on the edit distance plots and the post-mortem DNA damage, we were not able to conclude which species were present and which were not. The indication of

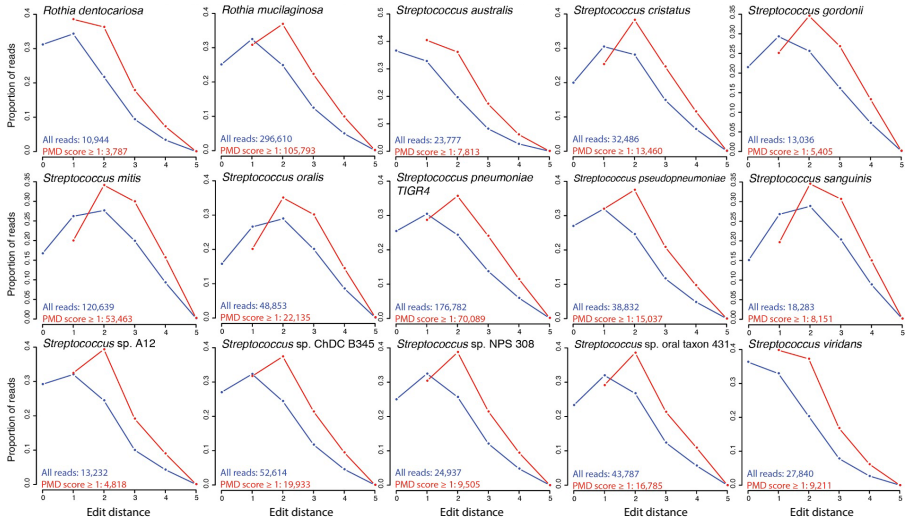

Figure 8.1. *Streptococcus* edit distances which is part of the **Supplementary Figure 10** of the accepted paper. Edit distance distributions of all reads (blue) and reads filtered for post-mortem damage ($PMD \geq 1$) (red) for bacterial taxa with >10,000 assigned reads recovered from the Syltholm pitch.

the presence of several *Streptococcus* species is more likely due to the high similarities among those species, than the fact that they were all present.

Phylogenetic Placement of the Ancient Gum

Even though several species of MGS were likely to be present in the sample, we wanted to reconstruct the phylogeny of the ancient sample. First, we mapped the ancient sample to a reference and reconstructed the phylogeny as described in Rasmussen *et al.* 2015 [145]. A total of 75 modern MGS genomes were included in the analyses and were a part of the final phylogenetic reconstruction. Figure 8.3 illustrates how important the choice of reference is for the phylogenetic reconstruction of these species. The trees showed a clear bias towards the chosen reference, as the ancient sample was placed in the same species cluster

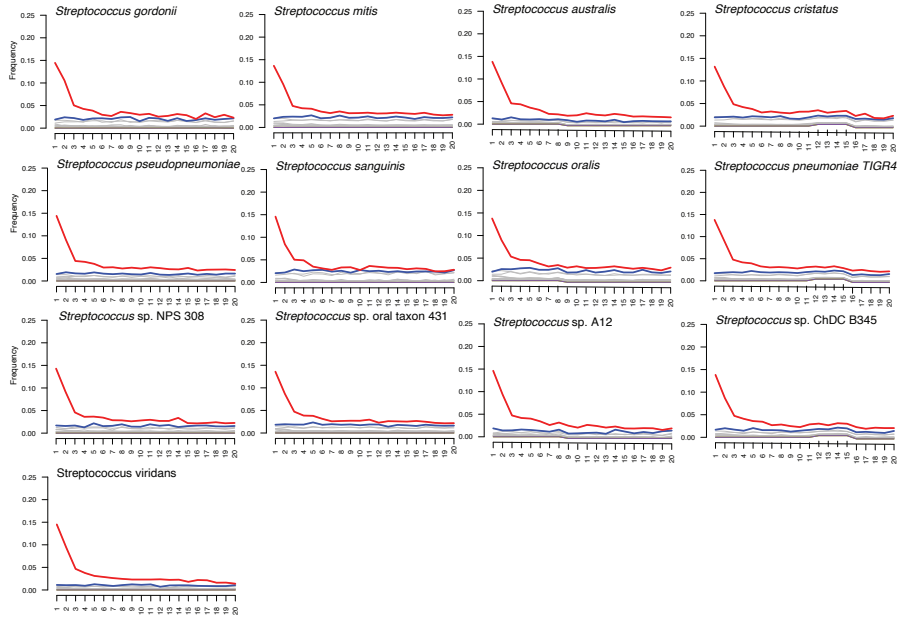

Figure 8.2. *Streptococcus* DNA damage patterns which are a part of **Supplementary Figure 8** in the accepted paper. The frequencies of all possible mismatches observed between the reference strain and the reads are reported in gray as a function of distance from 5' (first 25 nucleotides sequenced). The typical DNA damage mutations C-T (5') and G-A (3') are reported in red and blue, respectively.

as the reference. This method was therefore not suited for a sample with a mix of highly similar species.

To address this issue, we used a phylogenetic placement method [146] that places metagenomic reads onto the branches of a given reference tree. The reference tree was generated as described in Rasmussen *et al.* 2016 [147] and Iversen *et al.* [148] (Paper 1 and Paper 3). The ancient *Streptococcus* reads were placed on the *Streptococcus* reference tree using the evolutionary placement algorithm implemented in PaPaRa v. 2.543 with default settings and the option

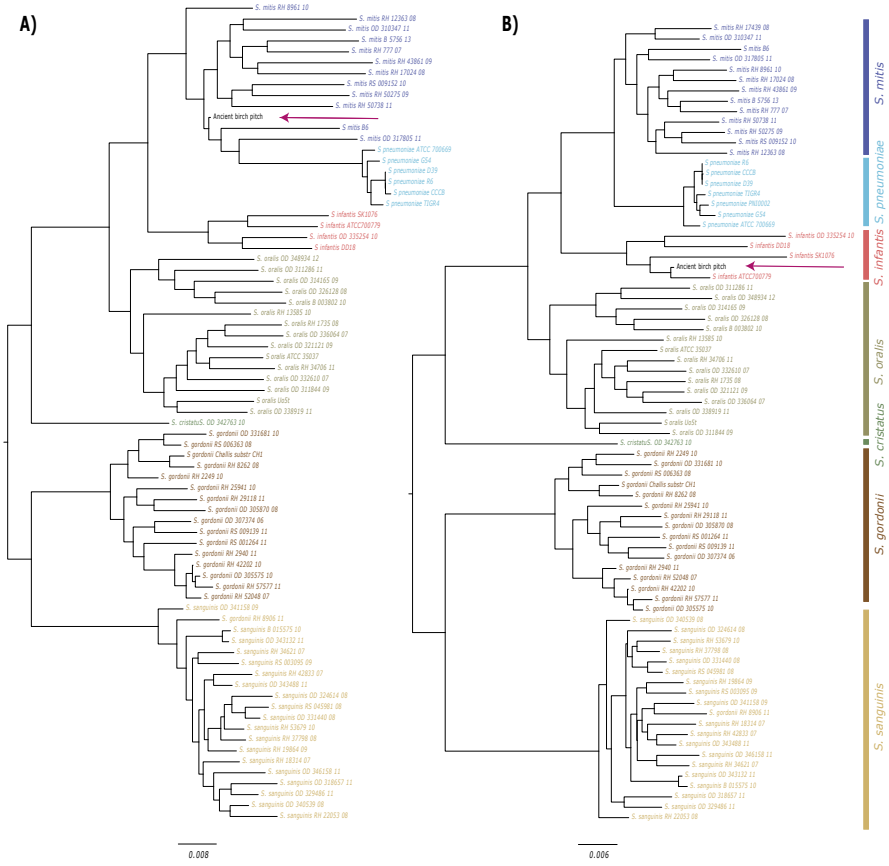

Figure 8.3. Phylogenetic reconstruction of the ancient pitch. The tree was constructed as described in [145]. We mapped the ancient reads together with 75 modern streptococci genomes to A) *S. mitis* RH8961 10 reference genome and B) the *S. infantis* ATCC700779 reference genome. Only the high confidence position were extracted for the phylogenetic reconstruction. **This figure is not a part of the accepted paper.**

-r. The placement of the ancient reads was then visualized with GAPPA [146] (Figure 8.4 A). To validate the method we used gargammel [149] to simulate 100,000 ancient DNA sequences from selected modern MGS strains, one for each species (Figure 8.4 B-J). We furthermore mapped the reads from three modern Human Oral Microbiome Samples (HOMS) to the reference (Figure 8.4 K-M).

We assigned the ancient reads to a *Streptococcus* reference tree based on core-genes from nine different species of the MGS. We found evidence for the presence of *S. infantis*, as well *S. mitis* and *S. pneumoniae* (Figure 8.4 A). Of those, *S. infantis* and *S. mitis* display continuously declining edit distance distributions, confirming the assignment (Figure 8.1). We also identified *S. pseudopneumoniae*, *S. cristatus* and *S. parasanguinis*, but based on simulations, these species were often assigned as false positives (Figure 8.4 B-J). The phylogenetic placement of the HOMS reads was similar to the phylogenetic placement of the ancient reads (Figure 8.4 K-M). This could suggest similarities between the ancient and modern oral microbiome.

Virulence Assessment of the Ancient Gum

While the majority of the species within the MGS are considered to be commensals of the oral cavity, some of them can cause infectious diseases, such as pneumonia and infective endocarditis. To investigate the virulence potential of the ancient *Streptococcus* spp., we tried to identify known virulence genes of modern *Streptococcus* strains in the ancient sample (Figure 8.5) (methods described in Paper 4). We compared the number of identified *Streptococcus* virulence genes in the ancient pitch with five modern HOMS, where the number of reads was included. We found that the number of identified virulence genes was higher in the ancient pitch, but when encountering the number of reads, the virulence potential of all the samples seemed to be similar.

The annotation of the 26 identified potential virulence genes can be found in Table 8.1.

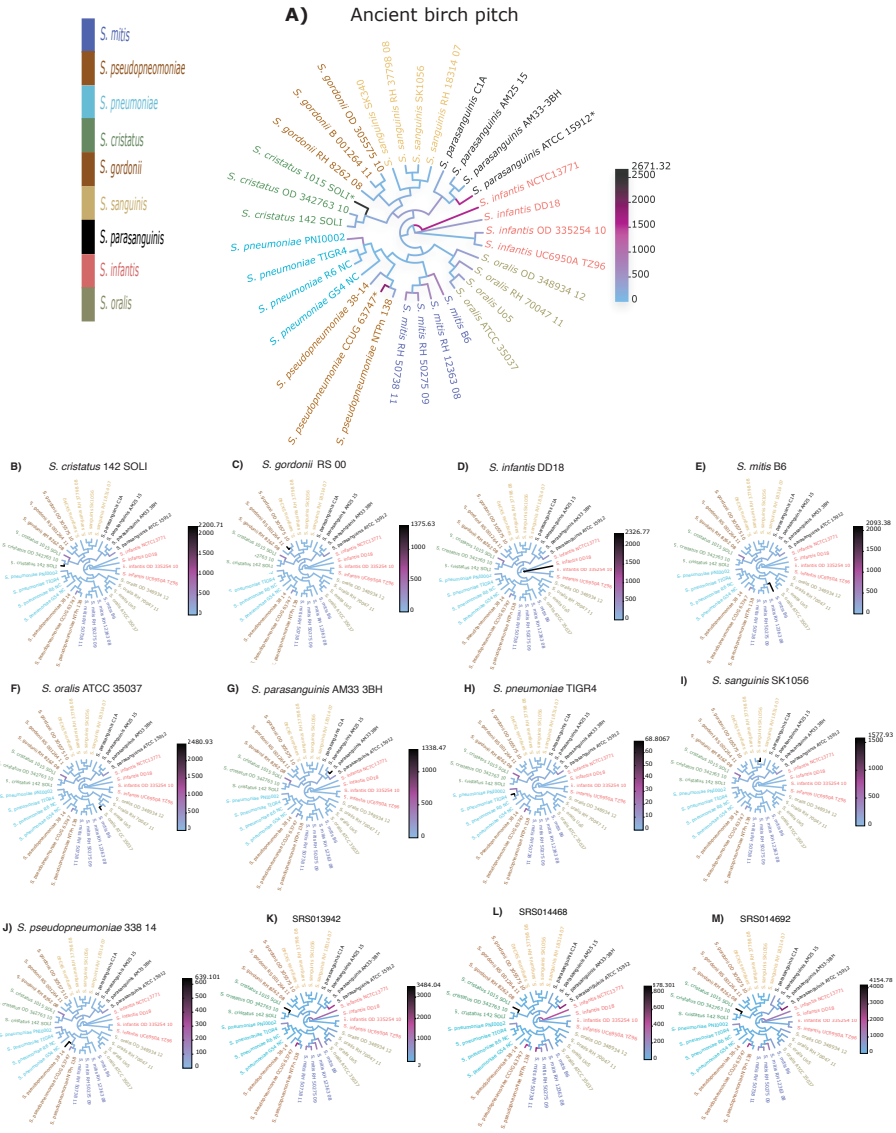

Figure 8.4. Phylogenetic placement of *Streptococcus* reads to a phylogenetic tree constructed from 312 *Streptococcus* core-genes. A) phylogenetic placement of the *Streptococcus* reads of the ancient pitch. B-J) phylogenetic placement of 100,000 simulated ancient DNA sequences from different 9 different MGS species. K-M) phylogenetic placement of *Streptococcus* reads from three different HOMS. **This figure is not a part of the accepted paper**.

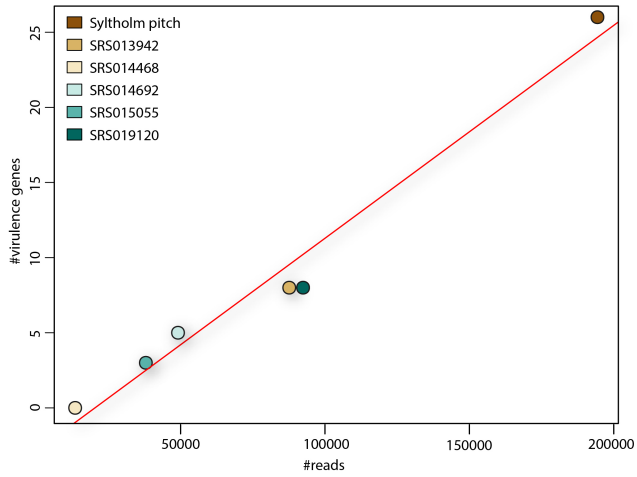

Figure 8.5. The number of virulence genes identified in the ancient pitch sample and five human oral microbiome samples from the human microbiome project [28]. The plot in **Supplementary Figure 16** of the acccepted paper.

Table 8.1. List of the 26 identified virulence genes in the ancient pitch. The table includes the virulence gene name, the contig name of the ancient pitch, the ID of the virulence gene in Virulence Factor DataBase (VFDB), the functional category, the function, and the VFDB reference genome. The table is part of the **Supplementary Data 6** of the accepted paper.

Virulence gene	qseqid	sseqid	Functional category	Function	VFDB reference genome
cppA	k79_290	VFG005512(gi:15903347)	protease	C3-degrading protease CVF147	*Streptococcus pneumoniae R6*
cps4C	k79_726	VFG001367(gb—NP_344881)	immune evation	capsular polysaccharide biosynthesis protein Cps4C	*Streptococcus pneumoniae TIGR4*
cps4D	k79_2768	VFG001368(gb—NP_344882)	immune evation	capsular polysaccharide biosynthesis protein Cps4D	*Streptococcus pneumoniae TIGR4*
cps4F	k79_2014	VFG001370(gb—NP_344884)	immune evation	capsular polysaccharide biosynthesis protein Cps4F	*Streptococcus pneumoniae TIGR4*
cps4K	k79_2043	VFG001375(gb—NP_344892)	immune evation	capsular polysaccharide biosynthesis protein Cps4K	*Streptococcus pneumoniae TIGR4*
eno	k79_2160	VFG005580(gi:15900994)	enzyme	phosphopyruvate hydratase, Streptococcal enolase CVF153	*Streptococcus pneumoniae TIGR4*
htrA/degP	k79_1192	VFG005532(gi:15904086)	protease	serine protease CVF148	*Streptococcus pneumoniae R6*
lmb	k79_1396	VFG019033(gi:116515739)	adherence	laminin-binding protein CVF114	*Streptococcus pneumoniae D39*
lytB	k79_4109	VFG001356(gb—NP_345446)	adherence	endo-beta-N-acetylglucosaminidase, choline binding protein VF0145	*Streptococcus pneumoniae TIGR4*
lytC	k79_5	VFG005339(gi:118090031)	adherence	lysozyme, choline binding protein CVF122	*Streptococcus pneumoniae TIGR4*
nanB	k79_3409	VFG001360(gb—NP_346126)	enzyme	neuraminidase B, neuraminidase VF0148	*Streptococcus pneumoniae TIGR4*
pavA	k79_3555	VFG005196(gi:15900843)	adherence	adherence and virulence protein A, fibronectin-binding protein CVF113	*Streptococcus pneumoniae TIGR4*
pce/cbpE	k79_370	VFG019069(gi:169832880)	adherence	choline binding protein E, choline-binding protein CVF122	*Streptococcus pneumoniae Hungary19A-6*
piuA	k79_2742	VFG005813(gi:15901700)	iron Uptake	iron-compound-binding protein CVF183, pneumococcal iron uptake	*Streptococcus pneumoniae TIGR4*
plr/gapA	k79_314	VFG005355(gi:22537907)	adherence	streptococcal plasmin receptor GAPDH CVF123	*Streptococcus agalactiae 2603V/R*
plr/gapA	k79_3756	VFG019078(gi:182684949)	adherence	streptococcal plasmin receptor GAPDH CVF123	*Streptococcus pneumoniae CGSP14*
psaA	k79_1969	VFG019102(gi:182684593)	manganse uptake	pneumococcal surface antigen A, metal binding protein SloC CVF181	*Streptococcus pneumoniae CGSP14*
rmlA	k79_1998	VFG006070(gi:55821251)	immune evation	glucose-1-phosphate thymidyl transferase, capsule CVF186	*Streptococcus thermophilus LMG 18311*
rmlD	k79_463	VFG005838(gi:125717701)	immune evation	DTDP-L-rhamnose synthase, putative, capsule CVF186	*Streptococcus sanguinis SK36*
slrA	k79_2400	VFG005372(gi:116516991)	adherence	peptidyl-prolyl cis-trans isomerase, cyclophilin-type, streptococcal lipoprotein rotamase A CVF129	*Streptococcus pneumoniae D39*
srtA	k79_3838	VFG019081(gi:169833213)	adherence	sortase A CVF130	*Streptococcus pneumoniae Hungary19A-6*
SSA_1516	k79_2551	VFG006027(gi:125718327)	immune evation	cell-wall biogenesis glycosyltransferase, putative, capsule CVF186	*Streptococcus sanguinis SK36*
SSU98_1513	k79_2119	VFG043469(gi:146321359)	adherence	phosphopyruvate hydratase, fibronectin-binding protein AI215	*Streptococcus suis 98HAH33*
tig/ropA	k79_2793	VFG005552(gi:125718786)	protease	trigger factor CVF149	*Streptococcus sanguinis SK36*
tig/ropA	k79_2648	VFG019091(gi:169834028)	protease	trigger factor CVF149	*Streptococcus pneumoniae Hungary19A-6*
wzd	k79_680	VFG019112(gi:182683318)	immune evation	capsular polysaccharide biosynthesis protein Cps14C, capsule CVF186	*Streptococcus pneumoniae CGSP14*

Paper 5

Genomic characterization, phylogenetic analysis, and identification of virulence factors in *Aerococcus sanguinicola* and *Aerococcus urinae* strains isolated from infection episodes

Derya Carkaci*, Katrine Højholt*, Xiaohui Chen Nielsen, Rimtas Dargis, Simon Rasmussen, Ole Skovgaard, Kurt Fuursted, Paal Skytt Andersen, Marc Stegger, and Jens Jørgen Christensen.

Aerococcus urinae and *Aerococcus sanguinicola* are gram-positive bacteria and are related to the non-hemolytic *Streptococcus* spp.. They grow in clusters and have a colony morphology similar to viridans streptococci [150]. Due to the similarities between aerococci, staphylococci, streptococci, and enterococci, these bacteria have probably been misidentified in many cases [150]. After the introduction of improved species identification using Matrix-Assisted Laser Desorption Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF MS) in routine laboratory assays, aerococci have increasingly been recognized as a human pathogen [150].

The knowledge about specific virulence genes and infectious mechanisms in *Aerococcus urinae* and *Aerococcus sanguinicola* is limited, but these species have been identified as infectious agents in IE, urinary tract infections, and sepsis. *A. urinae* and *A. sanguinicola* are observed to be able to form biofilms on plastic surfaces such as urinary catheters, and it has furthermore been observed that *A. urinae* can aggregate with human platelets [151, 152]. Both biofilm formation and platelet aggregation are believed to play a role in the pathogenesis of *A. urinae* and *A. sanguinicola* and their ability to cause IE

[151, 152]. To our knowledge, this study is the first to perform a genomic comparison of *A. urinae* and *A. sanguinicola* strains. We reconstructed the strain phylogeny, performed hierarchical clustering, and investigating genes known to be associated with virulence in other genera.

Contribution: I (Katrine Højholt Iversen) contributed to the design and development of the phylogenetic aspect of the paper. I performed the majority of the bioinformatics analysis and contributed to the interpretation of results. I furthermore constructed Figure 1 and contributed to writing and revision of the manuscript.

Link to paper:
https://doi.org/10.1016/j.micpath.2017.09.042

Discussion and Conclusion

Species-level identification of Mitis group streptococci (MGS) is challenging in a microbiological laboratory setting. Correct species-level identification is a crucial part of the diagnosis of infective endocarditis (IE), identification of treatment failure, and in some cases, infection relapse. It is mainly the close relationship between the MGS and their ability to exchange genetic material that complicates the correct species identification. As the majority of the MGS are harmless oral commensals, it is a paradox that the same species can cause severe infections as IE. With the five studies that were presented in this thesis, we have addressed some of these issues by using thorough phylogenetic analyses of 80 MGS species. Several genes associated with virulence were identified and investigated. Comprehensive genomics comparisons of strains isolated from healthy individuals and pathogenic strains isolated from patients with IE did not reveal any significant difference between the two isolation groups. Investigation of the ancient chewed birch pitch revealed that this type of artifact is an excellent source for ancient DNA, which can give novel insights into the human prehistory. Furthermore, the same level of virulence genes was detected in the ancient *Streptococcus* DNA as in modern oral samples. Lastly, two species of the genera *Aerococcus* was investigated. They contained several potential virulence genes that could explain their ability to cause infectious diseases.

In the work presented in this thesis, we have illustrated the complexity of MGS, especially regarding reconstructing their phylogeny. We showed that correct species identification of 80 MGS strains was possible when using a single marker-gene but that the robustness of the phylogenetic reconstruction was fairly low as indicated by the bootstrap values. Much higher robustness was observed in the studies using core-gene phylogeny (Paper 1, Paper 3, and Paper 5). Paper 2, Paper 3, and Paper 5 also illustrated the potential of applying hierarchical clustering on the pan-genome. Here we showed that even at subspecies level, the analyzed MGS strains contained specific differences in their

gene content. Core-genome phylogeny or pan-genome hierarchical clustering could, therefore, be the most suited method for correct and precise species identification. Core-genome phylogenetic reconstruction has been applied successfully in several other studies, confirming the potential of this approach [112, 153, 24]. The downside of these methods is that they are more time-consuming and, therefore, not (yet) suited for routine diagnostics in the clinic. MLSA and CIS phylogeny might be better alternatives for clinical work, as they both presented high phylogenetic robustness with high bootstrap values, and they are less time-consuming. A study by Bishop *et al.* [26] included 420 streptococcal strains for MLSA of seven house-keeping genes. The phylogenetic reconstruction showed that all MGS strains were clustering into their respective species clusters, except *S. pseudopneumoniae* and *S. perioditis*, which were not well resolved from the *S. mitis* cluster and the *S. infantis* cluster, respectively. Additionally, several strains were re-analyzed as they were outliers of the species clusters [26]. This could indicate that the robustness of MLSA phylogenetic reconstruction might decrease when more strains are included in the analysis. Another study by Hanage *et al.* [78], investigated the phylogenetics of closely related species of *Neisseria* spp., which, like the MGS, have high recombination rates and colonize the oral cavity. As inter-species recombination can impact the phylogenetic resolution when using a single gene, the authors used a multi-locus approach to reconstruct the phylogeny of 700 strains of 11 different *Neisseria* species. Using the multi-locus approach, they were able to make a clear resolution of the phylogeny and separate the species. However, it was suggested that the inter-species recombination resulted in intermediate species in the phylogenetic tree. The inter-species recombination was also observed in the study by Bishop *et al.*, and it was suggested that this mechanism could explain the great diversity that is observed in *S. mitis* and *S. oralis* clusters [26, 154]. For further studies of species identification of MGS, it would be interesting to reconstruct the phylogeny of 400-700 MGS strains using MLSA, whole-genome, or core-genome data and then investigate if the robustness of phylogenetic tree decreases when more strains are added. The inter-species recombination and the close relationship between the species might be an explanation for the failure of reconstruction of the phylogeny of ancient streptococci DNA in Paper 4. A study by Rasmussen 2015 *et al.* [145], was able to reconstruct the phylogeny of an ancient *Yersinia pestis* using the same pipeline as we applied in our study. However, their samples were likely to contain only one strain of *Y. pestis* and therefore experienced less reference bias in their phylogenetic reconstruction. Our ancient sample, on the other hand, most likely contained a mix of different MGS species. The combination

of the potential presence of several closely related species, low sequencing coverage, and DNA damage complicated the analysis. However, when we applied phylogenetic placement on a reference core-genome tree, we did find indications of the presence of *S. mitis*, *S. pneumoniae*, and *S. infantis* in the ancient birch pitch. The false positives and biases in the phylogenetic reconstruction did, however, introduce a certain amount of uncertainty in our results.

The close relationship between *S. pneumoniae* and *S. mitis* and their very different virulence potential have to lead to the discussion: who came first - the pathogen or the commensal? The study by Kilian *et al.* 2014 [153], investigated the phylogenetic linkage between *S. pneumoniae* and *S. mitis*. Their study showed that the two species have evolved in parallel from a common ancestor [153]. As our sample was 5,700 years old, likely, the divergence of the two species from a common ancestor started long before that. We, therefore, need to investigate even older samples to be able to prove the parallel evolution hypothesis, but nothing in our study suggests that it should not be true. Future studies involving metagenomic binning using deep learning might utilize separation of the MGS species in ancient samples. This would give a unique possibility to investigate the genomic evolution of these species.

In Paper 2, Paper 3, and Paper 5, we showed that by accessing the whole-genome sequences of bacterial strains, it is possible to analyze the virulence gene content. Identification of potential virulence genes in the oral MGS and *A. urinae* and *A. sanguinicola* can provide a more in-depth insight into the pathogenesis of IE. It was furthermore interesting to investigate the possible difference in virulence potential of strains isolated from IE patients and strains isolated from healthy individuals, and to elucidate a potential difference in virulence potential in ancient oral and modern oral *Streptococcus* DNA. We showed that several genes that were known to contribute to the virulence potential of *S. pneumoniae* were encoded in strains investigated in all three studies. As the virulence potential of *A. urinae* and *A. sanguinicola* is not fully covered; the identification of potential virulence genes in those strains would give a novel insight into their virulence mechanisms. Several virulence genes important for adhesion were present in all three studies. This indicates that these virulence genes are important for bacterial life as commensals in the oral cavity as well as pathogens causing infectious diseases. Especially bacterial adhesion to fibronectin may contribute to the development of IE [7]. Fibronectin is a protein secreted by a variety of cells, and it is present in saliva as well as blood. The gene *pavA* has been shown to facilitate adherence to human epithelial and endothelial cells in *S. pneumoniae* [61, 60]. As *pavA* was identified in all analyzed *S. mitis*, *S. oralis*, *S. sanguinis* genomes and in the ancient *Streptococcus* DNA,

this gene could very well be an important virulence factor. *lmb* was identified in the core-genome of all analyzed strains in Paper 2-5, and is, therefore, potentially an important virulence gene. *lmb* encodes for the lipoprotein, and the study by Spellerberg *et al.* [62] showed that *lmb* in *Streptococcus agalactiae* meditates the attachment to the human laminin. It is, furthermore, promoting the transfer of bacteria to the bloodstream and the colonization of damaged epithelium [62]. In Paper 2-5, we furthermore identified putative virulence genes, which are important for the immune modulation and colonization of the host. However, not all of the analyzed strains seemed to include the full capsular polysaccharide (CPS) locus, including Cps4A-D, which are indispensable for the virulence of *S. pneumoniae* [64, 65]. This indicates that the species might have a different encapsulation mechanism or can be encapsulated without the full locus. It is also possible that these species can invade the immune system and cause an infective disease without forming the inner shield that prevents phagocytosis, facilitated by the CPS. The lack of difference between the IE and oral isolates suggests that the oral strains had the same virulence potential as the IE isolates. This was also supported as both types of isolates contain very similar virulence gene content. The gene expression was, however, not investigated. Previous studies have shown that transcription of specific genes in *S. pneumoniae* are down-or up-regulated, when the bacteria are harvested in lung, blood, heart, or nasopharynx of mice [155, 156]. The study by Kilian *et al.* 2019 [157] investigated the virulence gene content as well as the regulatory mechanisms in *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, *S. infantis*, and three subspecies of *S. oralis*. They found a significant difference between the commensal oral streptococci and *S. pneumonoaie* in the matter of the number of regulatory mechanisms [157]. These commensal oral streptococci did, however, contain similar virulence genes, as we identified in Paper 2-5. Future studies in regulation profiles of strains isolated from IE patients and isolates of the oral cavity of healthy individuals could, therefore, be relevant, as the two isolation groups might present different regulation profiles. *In vivo* studies could furthermore help to understand the effect of the specific virulence genes of the oral commensal streptococci. That is, however, a challenging task, as many of the adherence mechanisms, as well as the immune modulation mechanisms, are specific to human cells [158, 159]. The host-specific virulence proteins might not facilitate binding or enzymatic activity in, e.g., a mouse model as they would in human tissue. Interpretations of the results of such *in vivo* studies should, therefore, be taken with care, and *in vitro* models with human cells might be a more suited setting for future studies.

Diet has a great impact on our oral health, and the diversity of species

colonizing the oral cavity [10, 19]. It has furthermore been suggested, that the carbohydrate-rich diet, introduced in the Neolithic period resulted in a less diverse and more pathogenic microbiome in humans [19]. The ancient chewed birch pitch examined in this paper is dated to be $\approx 5,700$ years old, which places it at the onset of the Neolithic period. The diet and the human ancestral traits did, however, suggest that the woman chewing the pitch was hunting and gathering food, rather than farming. Looking at the order-level microbial composition, we could see fewer bacteria belonging to the Bacteroidales order in the ancient sample than in the modern oral sample. Both *Porphyomanes gnigivalis* and *Tannerella forsythia* belong to the Baceroidales order and are associated with periodontal disease [12]. The lower level of this order could, therefore, support the study of Adler *et al.* [19]. However, a much more detailed investigation of the abundances of the individual species has to be performed to establish lower oral pathogenicity in the ancient sample.

Future Perspectives

The genomic contents of a pathogenic bacteria should not be the only parameter that should be considered in the quest to understand IE pathogenesis. Differences in host individual immune response and their overall health, diet, and hygiene habits might also provide important knowledge of the development of the disease. Furthermore, knowledge of the oral biofilm compositions of patients suffering from IE could provide new insight into the importance of oral microbiota regarding infectious diseases. Many body-sites do contain pathogenic species as a part of the natural microbiota without the development of an infection. Studies have shown that a shift in species abundance in the specific body-site can introduce infections such as community-acquired pneumonia [160], inflammatory bowel disease [161], and topic dermatitis [162]. This might also be the case of IE caused by MGS and needs to be enlightened further.

The studies of this thesis have provided essential knowledge about the phylogenetic relationship of MGS. This knowledge is applicable in the clinical setting, which might help detection of the causative pathogen using next-generation sequencing directly from blood, a sequencing method that has been identified to be a promising diagnostic platform for critically ill patients suffering from bloodstream infections [49, 47, 48]. Whole-genome sequencing would further-

more provide information regarding antibiotic resistance, which could enable a more precise and effective treatment of the patient.

References

[1] Holland, T. L. *et al.* Infective endocarditis. *Nature Reviews Disease Primers* **2**, 16059 (2016). URL http://www.nature.com/articles/nrdp201659.

[2] Que, Y.-A. & Moreillon, P. Infective endocarditis. *Nature Reviews Cardiology* **8**, 322–336 (2011). URL http://www.nature.com/articles/nrcardio.2011.43.

[3] Mitchell, J. Streptococcus mitis: walking the line between commensalism and pathogenesis: Review of S. mitis biology and pathogenesis. *Molecular Oral Microbiology* **26**, 89–98 (2011). URL http://doi.wiley.com/10.1111/j.2041-1014.2010.00601.x.

[4] Cabell, C. H. *et al.* Changing Patient Characteristics and the Effect on Mortality in Endocarditis. *ARCH INTERN MED* **162**, 5 (2002).

[5] Murdoch, D. R. *et al.* Clinical Presentation, Etiology, and Outcome of Infective Endocarditis in the 21st Century: The International Collaboration on Endocarditis–Prospective Cohort Study. *Archives of Internal Medicine* **169**, 463 (2009). URL http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinternmed.2008.603.

[6] Prendergast, B. D. The changing face of infective endocarditis. *Heart* **92**, 879–885 (2006). URL http://heart.bmj.com/cgi/doi/10.1136/hrt.2005.067256.

[7] Moreillon, P., Que, Y. A. & Bayer, A. S. Pathogenesis of streptococcal and staphylococcal endocarditis. *Infectious Disease Clinics of North America* **16**, 297–318 (2002). URL http://linkinghub.elsevier.com/retrieve/pii/S0891552001000095.

[8] Becker, M. R. *et al.* Molecular Analysis of Bacterial Species Associated with Childhood Caries. *Journal of Clinical Microbiology* **40**, 1001–1009 (2002). URL http://jcm.asm.org/cgi/doi/10.1128/JCM.40.3.1001-1009.2002.

[9] Stingu, C.-S., Eschrich, K., Rodloff, A. C., Schaumann, R. & Jentsch, H. Periodontitis is associated with a loss of colonization by Streptococcus sanguinis. *Journal of Medical Microbiology* **57**, 495–499 (2008). URL https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.47649-0.

[10] Wade, W. G. The oral microbiome in health and disease. *Pharmacological Research* **69**, 137–143 (2013). URL https://linkinghub.elsevier.com/retrieve/pii/S1043661812002277.

[11] Roberts, A. P. & Kreth, J. The impact of horizontal gene transfer on the adaptive ability of the human oral microbiome. *Frontiers in Cellular and Infection Microbiology* **4** (2014). URL http://journal.frontiersin.org/article/10.3389/fcimb.2014.00124/abstract.

[12] Gao, L. *et al.* Oral microbiomes: more and more importance in oral cavity and whole body. *Protein & Cell* **9**, 488–500 (2018). URL http://link.springer.com/10.1007/s13238-018-0548-1.

[13] Kreth, J., Merritt, J. & Qi, F. Bacterial and Host Interactions of Oral Streptococci. *DNA and Cell Biology* **28**, 397–403 (2009). URL http://www.liebertpub.com/doi/10.1089/dna.2009.0868.

[14] Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the Normal Bacterial Flora of the Oral Cavity. *Journal of Clinical Microbiology* **43**, 5721–5732 (2005). URL http://jcm.asm.org/cgi/doi/10.1128/JCM.43.11.5721-5732.2005.

[15] Priya Nimish Deo & Revati Deshmukh. Oral microbiome: Unveiling the fundamentals. *Journal of Oral Maxillofacial Pathology* **23**, 12 (2019).

[16] Nelson-Filho, P. *et al.* Dynamics of Microbial Colonization of the Oral Cavity in Newborns. *Brazilian Dental Journal* **24**, 415–419 (2013). URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-64402013000400415&lng=en&tlng=en.

[17] Rotimi, V. O. & Duerden, B. THE DEVELOPMENT OF THE BACTERIAL FLORA IN NORMAL NEONATES. *J. Med. Microbiol* **14**, 51–62 (1980).

[18] Caufield, P. W. *et al.* Natural History of Streptococcus sanguinis in the Oral Cavity of Infants: Evidence for a Discrete Window of Infectivity. *Infection and Immunity* **68**, 4018–4023 (2000). URL http://iai.asm.org/cgi/doi/10.1128/IAI.68.7.4018-4023.2000.

[19] Adler, C. J. *et al.* Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics* **45**, 450–455 (2013). URL http://www.nature.com/articles/ng.2536.

[20] Pitts, N. B. *et al.* Dental caries. *Nature Reviews Disease Primers* **3**, 17030 (2017). URL http://www.nature.com/articles/nrdp201730.

[21] Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211–1259 (2017). URL https://linkinghub.elsevier.com/retrieve/pii/S0140673617321542.

[22] Jensen, M. E. Diet and Dental Caries. *The Dental Clinics of North America* **43** (1999).

[23] Gao, X.-Y., Zhi, X.-Y., Li, H.-W., Klenk, H.-P. & Li, W.-J. Comparative Genomics of the Bacterial Genus Streptococcus Illuminates Evolutionary Implications of Species Groups. *PLoS ONE* **9**, e101229 (2014). URL https://dx.plos.org/10.1371/journal.pone.0101229.

[24] Lefébure, T. & Stanhope, M. J. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biology* **8**, R71 (2007). URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-5-r71.

[25] Nobbs, A. H., Lamont, R. J. & Jenkinson, H. F. Streptococcus Adherence and Colonization. *Microbiology and Molecular Biology Reviews* **73**, 407–450 (2009). URL http://mmbr.asm.org/cgi/doi/10.1128/MMBR.00014-09.

[26] Bishop, C. J. *et al.* Assigning strains to bacterial species via the internet. *BMC Biology* **7**, 3 (2009). URL http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-7-3.

[27] Kawamura, Y., Hou, X.-G., Sultana, F., Miura, H. & Ezaki, T. Determination of 16s rRNA Sequences of Streptococcus mitis and Streptococcus gordonii and Phylogenetic Relationships among Members of the Genus Streptococcus. *Internatiional Journal of Systematic Bacteriology* 3 (1995).

[28] Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**, baq013–baq013 (2010). URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/baq013.

[29] Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **44**, D7–D19 (2016). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1290.

[30] Zheng, W. *et al.* StreptoBase: An Oral Streptococcus mitis Group Genomic Resource and Analysis Platform. *PLOS ONE* **11**, e0151908 (2016). URL http://dx.plos.org/10.1371/journal.pone.0151908.

[31] Jakubovics, N. & Kolenbrander, P. The road to ruin: the formation of disease-associated oral biofilms: Formation of oral biofilms. *Oral Diseases* **16**, 729–739 (2010). URL http://doi.wiley.com/10.1111/j.1601-0825.2010.01701.x.

[32] Kreth, J., Merritt, J., Shi, W. & Qi, F. Competition and Coexistence between Streptococcus mutans and Streptococcus sanguinis in the Dental Biofilm. *Journal of Bacteriology* **187**, 7193–7203 (2005). URL http://jb.asm.org/cgi/doi/10.1128/JB.187.21.7193-7203.2005.

[33] Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention. *Frontiers in Immunology* **9**, 1366 (2018). URL https://www.frontiersin.org/article/10.3389/fimmu.2018.01366/full.

[34] O'Brien, K. L. *et al.* Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. *The Lancet* **374**, 893–902 (2009). URL https://linkinghub.elsevier.com/retrieve/pii/S0140673609612046.

[35] Wahl, B. *et al.* Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. *The Lancet Global Health* **6**, e744–e757 (2018). URL https://linkinghub.elsevier.com/retrieve/pii/S2214109X1830247X.

[36] Cabell, C. H., Abrutyn, E. & Karchmer, A. W. Bacterial Endocarditis: The Disease, Treatment, and Prevention. *Circulation* **107** (2003). URL https://www.ahajournals.org/doi/10.1161/01.CIR.0000071082.36561.F1.

[37] Cahill, T. J. & Prendergast, B. D. Infective endocarditis. *The Lancet* **387**, 882–893 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S0140673615000677.

[38] Moser, C. *et al.* 7. Infektiøs endocarditis (2017). Https://www.cardio.dk/endocarditis.

[39] Yew, H. S. & Murdoch, D. R. Global Trends in Infective Endocarditis Epidemiology. *Current Infectious Disease Reports* **14**, 367–372 (2012). URL http://link.springer.com/10.1007/s11908-012-0265-5.

[40] de Sa, D. D. C. *et al.* Epidemiological Trends of Infective Endocarditis: A Population-Based Study in Olmsted County, Minnesota. *Mayo Clinic Proceedings* **85**, 422–426 (2010). URL https://linkinghub.elsevier.com/retrieve/pii/S0025619611603273.

[41] Cahill, T. J., Dayer, M., Prendergast, B. & Thornhill, M. Do patients at risk of infective endocarditis need antibiotics before dental procedures? *BMJ* j3942 (2017). URL http://www.bmj.com/lookup/doi/10.1136/bmj.j3942.

[42] Osler, W. The Gulstonian Lectures, on Malignant Endocarditis. *BMJ* **1**, 577–579 (1885). URL http://www.bmj.com/cgi/doi/10.1136/bmj.1.1264.577.

[43] Beynon, R. P. & Prendergast, V. K. B., Bernard D. Infective endocarditis. *BMJ* **333**, 334–339 (2006).

[44] Sambola, A. *et al.* Sex Differences in Native-Valve Infective Endocarditis in a Single Tertiary-Care Hospital. *The American Journal of Cardiology* **106**, 92–98 (2010). URL https://linkinghub.elsevier.com/retrieve/pii/S0002914910006144.

[45] Alizzi, A., Tantiongco, J.-P. & Shepard, S. Chapter 10 - Infective Endocarditis. In *Microbiology for Surgical Infections*, 169–184 (Elsevier, 2014). URL https://linkinghub.elsevier.com/retrieve/pii/B9780124116290000106.

[46] Fricke, W. F. & Rasko, D. A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics* **15**, 49–55 (2014). URL http://www.nature.com/articles/nrg3624.

[47] Brenner, T. *et al.* Next-generation sequencing diagnostics of bacteremia in sepsis (Next GeneSiS-Trial): Study protocol of a prospective, observational, noninterventional, multicenter, clinical trial. *Medicine* **97**, e9868 (2018). URL http://Insights.ovid.com/crossref?an=00005792-201802090-00039.

[48] Grumaz, S. *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Medicine* **8**, 73 (2016). URL http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-016-0326-8.

[49] Decker, S. *et al.* Immune-Response Patterns and Next Generation Sequencing Diagnostics for the Detection of Mycoses in Patients with Septic Shock—Results of a Combined Clinical and Experimental Investigation. *International Journal of Molecular Sciences* **18**, 1796 (2017). URL http://www.mdpi.com/1422-0067/18/8/1796.

[50] Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms. *PLoS ONE* **9**, e104984 (2014). URL http://dx.plos.org/10.1371/journal.pone.0104984.

[51] QIAGEN. CLC Genomics Workbench (QIAGEN). URL https://www.qiagenbioinformatics.com.

[52] Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**, 75 (2008). URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-75.

[53] Jenkinson, H. F. Adherence and accumulation of oral streptococci. *Trends in Microbiology* **2**, 209–212 (1994). URL https://linkinghub.elsevier.com/retrieve/pii/0966842X9490114K.

[54] Rosan, B. & Lamont, R. J. Dental plaque formation. *Microbes and Infection* **2**, 1599–1607 (2000). URL https://linkinghub.elsevier.com/retrieve/pii/S1286457900013162.

[55] Moschioni, M., Pansegrau, W. & Barocchi, M. A. Adhesion determinants of the Streptococcus species: Adhesion determinants of Streptococcus species. *Microbial Biotechnology* **3**, 370–388 (2009). URL http://doi.wiley.com/10.1111/j.1751-7915.2009.00138.x.

[56] Scheld, W. M., Valone, J. A. & Sande, M. A. Bacterial adherence in the pathogenesis of endocarditis. Interaction of bacterial dextran, platelets, and fibrin. *Journal of Clinical Investigation* **61**, 1394–1404 (1978). URL http://www.jci.org/articles/view/109057.

[57] Plaut, A. G. The IgA1 Proteases of Pathogenic Bacteria. *Annual Review of Microbiology* **37**, 603–622 (1983). URL http://www.annualreviews.org/doi/10.1146/annurev.mi.37.100183.003131.

[58] Bek-Thomsen, M., Poulsen, K. & Kilian, M. Occurrence and Evolution of the Paralogous Zinc Metalloproteases IgA1 Protease, ZmpB, ZmpC, and ZmpD in Streptococcus pneumoniae and Related Commensal Species. *mBio* **3**, e00303–12 (2012). URL https://mbio.asm.org/lookup/doi/10.1128/mBio.00303-12.

[59] Berry, A. M. & Paton, J. C. Sequence Heterogeneity of PsaA, a 37-Kilodalton Putative Adhesin Essential for Virulence of Streptococcus pneumoniae. *INFECT. IMMUN.* **64**, 8 (1996).

[60] Pracht, D. *et al.* PavA of Streptococcus pneumoniae Modulates Adherence, Invasion, and Meningeal Inflammation. *Infection and Immunity* **73**, 2680–2689 (2005). URL http://iai.asm.org/cgi/doi/10.1128/IAI.73.5.2680-2689.2005.

[61] Holmes, A. R. *et al.* The pavA gene of Streptococcus pneumoniae encodes a fibronectin-binding protein that is essential for virulence. *Molecular Microbiology* **41**, 1395–1408 (2001). URL http://doi.wiley.com/10.1046/j.1365-2958.2001.02610.x.

[62] Spellerberg, B. *et al.* Lmb, a Protein with Similarities to the LraI Adhesin Family, Mediates Attachment of Streptococcus agalactiae to Human Laminin. *INFECT. IMMUN.* **67**, 8 (1999).

[63] Kerrigan, S. W. *et al.* Role of Streptococcus gordonii Surface Proteins SspA/SspB and Hsa in Platelet Function. *Infection and Immunity* **75**, 5740–5747 (2007). URL http://iai.asm.org/cgi/doi/10.1128/IAI.00909-07.

[64] Rukke, H. V. *et al.* Protective Role of the Capsule and Impact of Serotype 4 Switching on Streptococcus mitis. *Infection and Immunity* **82**, 3790–3801 (2014). URL http://iai.asm.org/lookup/doi/10.1128/IAI.01840-14.

[65] Paton, J. C. & Trappetti, C. Streptococcus pneumoniae Capsular Polysaccharide. *Microbiology Spectrum* **7** (2019). URL http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.GPP3-0019-2018.

[66] Kerrigan, S. W. & Cox, D. Platelet–bacterial interactions. *Cellular and Molecular Life Sciences* **67**, 513–523 (2010). URL http://link.springer.com/10.1007/s00018-009-0207-z.

[67] Love, R. M., Mcmillan, M. D. & Jenkinson, H. F. Invasion of Dentinal Tubules by Oral Streptococci Is Associated with Collagen Recognition Mediated by the Antigen I/II Family of Polypeptides. *INFECT. IMMUN.* **65**, 8 (1997).

[68] Ghoshal, K. & Bhattacharyya, M. Overview of Platelet Physiology: Its Hemostatic and Nonhemostatic Role in Disease Pathogenesis. *The Scientific World Journal* **2014**, 1–16 (2014). URL http://www.hindawi.com/journals/tswj/2014/781857/.

[69] Janoff, E. N. *et al.* Pneumococcal IgA1 protease subverts specific protection by human IgA1. *Mucosal Immunology* **7**, 249–256 (2014). URL http://www.nature.com/articles/mi201341.

[70] Geno, K. A., Saad, J. S. & Nahm, M. H. Discovery of Novel Pneumococcal Serotype 35d, a Natural WciG-Deficient Variant of Serotype 35b. *Journal of Clinical Microbiology* **55**, 1416–1425 (2017). URL http://jcm.asm.org/lookup/doi/10.1128/JCM.00054-17.

[71] Jakubovics, N. S., Yassin, S. A. & Rickard, A. H. Community Interactions of Oral Streptococci. In *Advances in Applied Microbiology*, vol. 87, 43–110 (Elsevier, 2014). URL https://linkinghub.elsevier.com/retrieve/pii/B9780128002612000025.

[72] Steinmoen, H., Knutsen, E. & Havarstein, L. S. Induction of natural competence in Streptococcus pneumoniae triggers lysis and DNA release from a subfraction of the cell population. *Proceedings of the National Academy of Sciences* **99**, 7681–7686 (2002). URL http://www.pnas.org/cgi/doi/10.1073/pnas.112464599.

[73] Andam, C. P. & Hanage, W. P. Mechanisms of genome evolution of Streptococcus. *Infection, Genetics and Evolution* **33**, 334–342 (2015). URL https://linkinghub.elsevier.com/retrieve/pii/S1567134814004109.

[74] Donati, C. *et al.* Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology* **11**, R107 (2010). URL https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r107.

[75] Feil, E. J. *et al.* Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences* **98**, 182–187 (2001). URL http://www.pnas.org/cgi/doi/10.1073/pnas.98.1.182.

[76] Hao, W. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research* **16**, 636–643 (2006). URL http://www.genome.org/cgi/doi/10.1101/gr.4746406.

[77] Spratt, B. The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Current Opinion in Microbiology* **4**, 602–606 (2001). URL http://linkinghub.elsevier.com/retrieve/pii/S1369527400002575.

[78] Hanage, W. P., Fraser, C. & Spratt, B. G. Fuzzy species among recombinogenic bacteria. *BMC Biology* **3**, 6 (2005). URL http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-3-6.

[79] Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016). URL https://linkinghub.elsevier.com/retrieve/pii/S0888754315300410.

[80] Hiyoshi, A., Miyahara, K., Kato, C. & Ohshima, Y. Does a DNA-less cellular organism exist on Earth?: Search for a DNA-less microbe. *Genes to Cells* **16**, 1146–1158 (2011). URL http://doi.wiley.com/10.1111/j.1365-2443.2011.01558.x.

[81] Betts, J. G. *et al.* DNA Nucleotides (2016).

[82] Dabney, J., Meyer, M. & Paabo, S. Ancient DNA Damage. *Cold Spring Harbor Perspectives in Biology* **5**, a012567–a012567 (2013). URL http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a012567.

[83] Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry* **55**, 641–658 (2009). URL http://www.clinchem.org/cgi/doi/10.1373/clinchem.2008.112789.

[84] Philippidis, A. Top 10 Sequencing Companies. *Genetic Engineering & Biotechnology News* (2018). Https://www.genengnews.com/a-lists/top-10-sequencing-companies-2/.

[85] Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nature Reviews Genetics* **20**, 341–355 (2019). URL http://www.nature.com/articles/s41576-019-0113-7.

[86] Garrido-Cardenas, J. A. & Manzano-Agugliaro, F. The metagenomics worldwide research. *Current Genetics* **63**, 819–829 (2017). URL http://link.springer.com/10.1007/s00294-017-0693-8.

[87] Rappé, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority. *Annual Review of Microbiology* **57**, 369–394 (2003). URL http://www.annualreviews.org/doi/10.1146/annurev.micro.57.030502.090759.

[88] Nielsen, H Bjørn *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* **32**, 822–828 (2014). URL http://www.nature.com/articles/nbt.2939.

[89] Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019). URL https://peerj.com/articles/7359.

[90] Alneberg, J. *et al.* CONCOCT: Clustering cONtigs on COverage and Composition. *arXiv:1312.4038 [q-bio]* (2013). URL http://arxiv.org/abs/1312.4038. ArXiv: 1312.4038.

[91] Nissen, J. N. *et al.* Binning microbial genomes using deep learning. preprint, Bioinformatics (2018). URL http://biorxiv.org/lookup/doi/10.1101/490078.

[92] Orlando, L. & Cooper, A. Using Ancient DNA to Understand Evolutionary and Ecological Processes. *Annual Review of Ecology, Evolution, and Systematics* **45**, 573–598 (2014). URL http://www.annualreviews.org/doi/10.1146/annurev-ecolsys-120213-091712.

[93] Pääbo, S. *et al.* Genetic Analyses from Ancient DNA. *Annual Review of Genetics* **38**, 645–679 (2004). URL http://www.annualreviews.org/doi/10.1146/annurev.genet.37.110801.143214.

[94] Hagelberg, E., Hofreiter, M. & Keyser, C. Ancient DNA: the first three decades. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20130371 (2015). URL https://royalsocietypublishing.org/doi/10.1098/rstb.2013.0371.

[95] Hofreiter, M., Serre, D., Poinar, H. N., Kuck, M. & Pääbo, S. Ancient DNA. *Nature Reviews Genetics* **2**, 352–359 (2001).

[96] Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15**, 121–132 (2014). URL http://www.nature.com/articles/nrg3642.

[97] Illumina, I. Quality Scores for Next-Generation Sequencing. *Technical Note: Sequencing, Illumina, Inc* 2 (2011).

[98] Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E. & Orlando, L. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153–2155 (2011). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr347.

[99] Sohn, J.-i. & Nam, J.-W. The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics* bbw096 (2016). URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw096.

[100] Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29**, 987–991 (2011). URL http://www.nature.com/articles/nbt.2023.

[101] Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012). URL http://www.liebertpub.com/doi/10.1089/cmb.2012.0021.

[102] Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **13**, 303–314 (2012). URL http://www.nature.com/articles/nrg3186.

[103] Katherine, S. J. Review Paper: The Shape of Phylogenetic Treespace. *Systematic Biology* syw025 (2016). URL https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syw025.

[104] Kinene, T., Wainaina, J., Maina, S. & Boykin, L. Rooting Trees, Methods for. In *Encyclopedia of Evolutionary Biology*, 489–493 (Elsevier, 2016). URL https://linkinghub.elsevier.com/retrieve/pii/B9780128000496002158.

[105] Yang, Z. *Computational Molecular Evolution* (Oxford University Press, 2006). URL http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198567028.001.0001/acprof-9780198567028.

[106] Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu033.

[107] Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307–321 (2010). URL https://academic.oup.com/sysbio/article/59/3/307/1702850.

[108] Bazinet, A. L., Zwickl, D. J. & Cummings, M. P. A Gateway for Phylogenetic Analysis Powered by Grid Computing Featuring GARLI 2.0. *Systematic Biology* **63**, 812–818 (2014). URL https://academic.oup.com/sysbio/article/63/5/812/2847779.

[109] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh340.

[110] Chung, M., Munro, J. B., Tettelin, H. & Hotopp, J. C. D. Using Core Genome Alignments To Assign Bacterial Species **3**, 21 (2018).

[111] Janda, J. M. & Abbott, S. L. 16s rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology* **45**, 2761–2764 (2007). URL http://jcm.asm.org/cgi/doi/10.1128/JCM.01228-07.

[112] Zheng, W. *et al.* Distinct Biological Potential of Streptococcus gordonii and Streptococcus sanguinis Revealed by Comparative Genome Analysis. *Scientific Reports* **7** (2017). URL http://www.nature.com/articles/s41598-017-02399-4.

[113] Lees, J. *et al.* Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res* (2018).

[114] Li, L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13**, 2178–2189 (2003). URL http://www.genome.org/cgi/doi/10.1101/gr.1224503.

[115] Lukjancenko, O., Thomsen, M. C., Voldby Larsen, M. & Ussery, D. W. PanFunPro: PAN-genome analysis based on FUNctional PROfiles. *F1000Research* (2013). URL http://f1000research.com/articles/2-265/v1.

[116] Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). URL http://www.biomedcentral.com/1471-2105/10/421.

[117] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts565.

[118] Eddy, S. R. Hidden Markov Models. *Current Opinion in Structural Biology* **6**, 361–365 (1996).

[119] Pearson, W. R. An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics* **42**, 3.1.1–3.1.8 (2013). URL http://doi.wiley.com/10.1002/0471250953.bi0301s42.

[120] Pearson, W. R. Selecting the Right Similarity-Scoring Matrix. *Current Protocols in Bioinformatics* **43** (2013). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi0305s43.

[121] Attwood, T. The quest to deduce protein function from sequence: the role of pattern databases. *The International Journal of Biochemistry & Cell Biology* **32**, 139–155 (2000). URL http://linkinghub.elsevier.com/retrieve/pii/S1357272599001065.

[122] Williamson, M. P. *How proteins work, Chapter 2* (Garland Science, New York, 2012).

[123] Kitts, P. A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Research* **44**, D73–D80 (2016). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1226.

[124] Henikoff, S. Gene Families: The Taxonomy of Protein Paralogs and Chimeras. *Science* **278**, 609–614 (1997). URL http://www.sciencemag.org/cgi/doi/10.1126/science.278.5338.609.

[125] Enright, A. J. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584 (2002). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/30.7.1575.

[126] Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Research* **40**, D290–D301 (2012). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr1065.

[127] Haft, D. H. The TIGRFAMs database of protein families. *Nucleic Acids Research* **31**, 371–373 (2003). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg128.

[128] Wilson, D. *et al.* SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research* **37**, D380–D386 (2009). URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn762.

[129] Zdobnov, E. M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/17.9.847.

[130] Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5**, 83 (2019). URL http://www.nature.com/articles/s41524-019-0221-0.

[131] Mullainathan, S. & Spiess, J. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* **31**, 87–106 (2017). URL http://pubs.aeaweb.org/doi/10.1257/jep.31.2.87.

[132] Baharudin, B., Lee, L. H. & Khan, K. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology* **1**, 4–20 (2010). URL http://www.jait.us/index.php?m=content&c=index&a=show&catid=160&id=859.

[133] Dada, E. G. *et al.* Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* **5**, e01802 (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S2405844018353404.

[134] Angelov, P. P. & Gu, X. Toward Anthropomorphic Machine Learning. *Computer* **51**, 18–27 (2018). URL https://ieeexplore.ieee.org/document/8481253/.

[135] Larrañaga, P. *et al.* Machine learning in bioinformatics. *Briefings in Bioinformatics* **7**, 86–112 (2006). URL https://academic.oup.com/bib/article/7/1/86/264025.

[136] Dittman, D. J., Khoshgoftaar, T. M. & Napolitano, A. The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data. In *2015 IEEE International Conference on Information Reuse and Integration*, 457–463 (IEEE, San Francisco, CA, USA, 2015). URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7301012.

[137] Dey, A. Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies* **7**, 6 (2016).

[138] Li, H. *et al.* Correlation and redundancy on machine learning performance for chemical databases: Correlation and Redundancy on Machine Learning Regressions. *Journal of Chemometrics* **32**, e3023 (2018). URL http://doi.wiley.com/10.1002/cem.3023.

[139] Bishop, C. M. *Pattern recognition and machine learning.* Information science and statistics (Springer, New York, 2006).

[140] Goodfellow, I., Bengio, Y. & Courville, A. Chaper 5. Machine Learning Basics. In *Deep Learning* (MIT Press, 2017).

[141] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intellligence* 7 (1995).

[142] Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* S2210832718301546 (2018). URL `https://linkinghub.elsevier.com/retrieve/pii/S2210832718301546`.

[143] Bewick, V., Cheek, L. & Ball, J. Statistics review 13: Receiver operating characteristic curves. *Critical Care* **8**, 508 (2004). URL `http://ccforum.biomedcentral.com/articles/10.1186/cc3000`.

[144] Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE* **12**, e0177678 (2017). URL `https://dx.plos.org/10.1371/journal.pone.0177678`.

[145] Rasmussen, S. *et al.* Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* **163**, 571–582 (2015). URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867415013227`.

[146] Czech, L., Barbera, P. & Stamatakis, A. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics* **35**, 1151–1158 (2019). URL `https://academic.oup.com/bioinformatics/article/35/7/1151/5088318`.

[147] Rasmussen, L. H. *et al.* Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of Mitis group streptococci. *European Journal of Clinical Microbiology & Infectious Diseases* **35**, 1615–1625 (2016). URL `http://link.springer.com/10.1007/s10096-016-2700-2`.

[148] Iversen, K. *et al.* Similar genomic patterns of clinical infective endocarditis and oral isolates of Streptococcus sanguinis and Streptococcus gordonii. *Manuscript in review* .

[149] Renaud, G., Hanghøj, K., Willerslev, E. & Orlando, L. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* btw670 (2016). URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw670`.

[150] Rasmussen, M. Aerococcus : an increasingly acknowledged human pathogen. *Clinical Microbiology and Infection* **22**, 22–27 (2016). URL `https://linkinghub.elsevier.com/retrieve/pii/S1198743X15008964`.

[151] Shannon, O., Morgelin, M. & Rasmussen, M. Platelet Activation and Biofilm Formation by Aerococcus urinae, an Endocarditis-Causing Pathogen. *Infection and Immunity* **78**, 4268–4275 (2010). URL `http://iai.asm.org/cgi/doi/10.1128/IAI.00469-10`.

[152] Rasmussen, M. Aerococci and aerococcal infections. *Journal of Infection* **66**, 467–474 (2013). URL `https://linkinghub.elsevier.com/retrieve/pii/S0163445312003878`.

[153] Kilian, M., Riley, D. R., Jensen, A., Bruggemann, H. & Tettelin, H. Parallel Evolution of Streptococcus pneumoniae and Streptococcus mitis to Pathogenic and Mutualistic Lifestyles. *mBio* **5** (2014). URL `http://mbio.asm.org/cgi/doi/10.1128/mBio.01490-14`.

[154] Hanage, W. P., Fraser, C. & Spratt, B. G. Sequences, sequence clusters and bacterial species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1917–1927 (2006). URL https://royalsocietypublishing.org/doi/10.1098/rstb.2006.1917.

[155] Ogunniyi, A. D. *et al.* Identification of Genes That Contribute to the Pathogenesis of Invasive Pneumococcal Disease by *In Vivo* Transcriptomic Analysis. *Infection and Immunity* **80**, 3268–3278 (2012). URL http://iai.asm.org/lookup/doi/10.1128/IAI.00295-12.

[156] Shenoy, A. T. *et al.* Streptococcus pneumoniae in the heart subvert the host response through biofilm-mediated resident macrophage killing. *PLOS Pathogens* **13**, e1006582 (2017). URL https://dx.plos.org/10.1371/journal.ppat.1006582.

[157] Kilian, M. & Tettelin, H. 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**, e01985–19, /mbio/10/5/mBio.01985–19.atom (2019). URL http://mbio.asm.org/lookup/doi/10.1128/mBio.01985-19.

[158] Lu, L. *et al.* Species-Specific Interaction of *Streptococcus pneumoniae* with Human Complement Factor H. *The Journal of Immunology* **181**, 7138–7146 (2008). URL http://www.jimmunol.org/lookup/doi/10.4049/jimmunol.181.10.7138.

[159] Kilian, M., Reinholdt, J., Lomholt, H., Poulsen, K. & Frandsen, E. V. G. Biological significance of IgA1 proteases in bacterial colonization and pathogenesis: critical evaluation of experimental evidence. *APMIS* **104**, 321–338 (1996). URL http://doi.wiley.com/10.1111/j.1699-0463.1996.tb00724.x.

[160] Weiser, J. N., Ferreira, D. M. & Paton, J. C. Streptococcus pneumoniae: transmission, colonization and invasion. *Nature Reviews Microbiology* **16**, 355–367 (2018). URL http://www.nature.com/articles/s41579-018-0001-8.

[161] Walk, S. T., Blum, A. M., Ewing, S. A.-S., Weinstock, J. V. & Young, V. B. Alteration of the murine gut microbiota during infection with the parasitic helminth Heligmosomoides polygyrus:. *Inflammatory Bowel Diseases* **16**, 1841–1849 (2010). URL https://academic.oup.com/ibdjournal/article/16/11/1841-1849/4628179.

[162] Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research* **22**, 850–859 (2012). URL http://genome.cshlp.org/cgi/doi/10.1101/gr.131029.111.

Appendix

Additional Material for Streptococcal Phylogeny

Table S1. List of species and strain names that was used to generate the phylogenetic core-tree in Figure 1.2. All genomes were downloaded from NCBI August 30th 2019 [29]

Species	**Strain**	**BioSample**	**BioProject**
Streptococcus acidominimus	NCTC11291	SAMEA4504048	PRJEB6403
Streptococcus agalactiae	2603V/R	SAMN02604013	PRJNA330
Streptococcus anginosus	C238	SAMN02603663	PRJNA50425
Streptococcus australis	NCTC13166	SAMEA103899862	PRJEB6403
Streptococcus azizii	Dec-91	SAMN06174762	PRJNA224116
Streptococcus bovimastitidis	NZ1587	SAMN05199910	PRJNA324185
Streptococcus caballi	DSM 19004	SAMN02256438	PRJNA224116
Streptococcus canis	NCTC12191	SAMEA3662933	PRJNA224116
Streptococcus castoreus	DSM 17536	SAMN02441486	PRJNA188329
Streptococcus constellatus	subsp. pharyngis C232	SAMN02603657	PRJNA224116
Streptococcus criceti	HS-6	SAMN02436557	PRJNA224116
Streptococcus cristatus	AS 1.3089	SAMN02603400	PRJNA224116
Streptococcus cuniculi	CCUG 65085	SAMN06174763	PRJNA224116
Streptococcus devriesei	DSM 19639	SAMN02441502	PRJNA185665
Streptococcus didelphis	DSM 15616	SAMN02256413	PRJNA224116
Streptococcus downei	F0415	SAMN00115114	PRJNA224116
Streptococcus dysgalactiae	subsp. equisimilis AC-2713	SAMEA3138428	PRJNA224116
Streptococcus entericus	DSM 14446	SAMN02256412	PRJNA224116
Streptococcus equi	subsp. zooepidemicus H70	SAMEA2272197	PRJNA224116
Streptococcus equinus	AG46	SAMN02841210	PRJNA224116
Streptococcus ferus	DSM 20646 A3GYDRAFT	SAMN02256504	PRJNA224116
Streptococcus gallolyticus	subsp. gallolyticus DSM 16831	SAMN06140979	PRJNA224116
Streptococcus gordonii	Challis substr. CH1	SAMN02603977	PRJNA224116
Streptococcus halitosis	VT-4	SAMN08929985	PRJNA224116
Streptococcus halotolerans	HTS9	SAMN04562598	PRJNA224116
Streptococcus henryi	DSM 19005 F601	SAMN02441723	PRJNA224116
Streptococcus himalayensis	HTS2	SAMN05560386	PRJNA224116
Streptococcus hongkongensis	CAIM 1895	SAMN03093231	PRJNA263025
Streptococcus hyointestinalis	NCTC12224	SAMEA3594353	PRJEB6403
Streptococcus hyovaginalis	DSM 12219	SAMN02440576	PRJNA185666
Streptococcus ictaluri	707-05	SAMN02436329	PRJNA224116
Streptococcus infantarius	ICDDRB-NRC-S5	SAMN04348603	PRJNA224116
Streptococcus infantis	700779	SAMN00216972	PRJNA224116
Streptococcus intermedius	B196	SAMN02603659	PRJNA224116
Streptococcus lutetiensis	NCTC8738	SAMEA3643305	PRJNA224116
Streptococcus macacae	NCTC 11558	SAMN02436328	PRJNA224116
Streptococcus macedonicus	ACA-DC	SAMEA2272145	PRJNA224116

Table S1 continued

Species	Strain	BioSample	BioProject
Streptococcus marimammalium	DSM 18627	SAMN02256437	PRJNA224116
Streptococcus marmotae	HTS5	SAMN04592763	PRJNA224116
Streptococcus massiliensis	DSM 18628	SAMN02256435	PRJNA224116
Streptococcus merionis	DSM 19192	SAMN02256436	PRJNA224116
Streptococcus milleri	NCTC11169	SAMEA3649034	PRJNA224116
Streptococcus minor	DSM 17118	SAMN02441724	PRJNA224116
Streptococcus mitis	B6		PRJNA46097
Streptococcus mutans	UA159		PRJNA57947
Streptococcus oralis	Uo5	SAMEA2272261	PRJNA224116
Streptococcus orisasini	SH06	SAMD00042493	PRJDB4302
Streptococcus orisratti	DSM 15617	SAMN02256410	PRJNA224116
Streptococcus ovis	DSM 16829	SAMN02256411	PRJNA224116
Streptococcus pantholopis	TA 26	SAMN04534843	PRJNA224116
Streptococcus parasanguinis	ATCC 15912	SAMN00113608	PRJNA224116
Streptococcus parasuis	4253	SAMN10939614	PRJNA522436
Streptococcus parauberis	KCTC 11537	SAMN02603301	PRJNA224116
Streptococcus pasteurianus	ATCC 43144	SAMD00060984	PRJNA224116
Streptococcus penaeicida	CAIM 1838	SAMN04311937	PRJNA304970
Streptococcus peroris	ATCC 700780	SAMN00253298	PRJNA53059
Streptococcus pharyngis	CCUG 66496	SAMN12393120	PRJNA557276
Streptococcus phocae	subsp. salmonis strain C-4	SAMN03114893	PRJNA224116
Streptococcus pluranimalium	TH11417	SAMN08224108	PRJNA224116
Streptococcus plurextorum	DSM 22810	SAMN02441159	PRJNA224116
Streptococcus pneumoniae	TIGR4	SAMN02604002	PRJNA224116
Streptococcus porci	DSM 23759	SAMN02440826	PRJNA224116
Streptococcus porcinus	NCTC10999	SAMEA3632063	PRJNA224116
Streptococcus pseudopneumoniae	IS7493	SAMN02603728	PRJNA224116
Streptococcus pseudoporcinus	LQ 940-04	SAMN02436558	PRJNA224116
Streptococcus pyogenes	M1 GAS	SAMN02604089	PRJNA57845
Streptococcus ratti	FA-1	SAMN02428953	PRJNA224116
Streptococcus respiraculi	HTS25	SAMN07374896	PRJNA224116
Streptococcus rubneri	DSM 26920	SAMN11349107	PRJNA531058
Streptococcus ruminantium	GUT187T	SAMD00097183	PRJDB6417
Streptococcus salivarius	NCTC 8618	SAMN03174835	PRJNA224116
Streptococcus sanguinis	SK36		PRJNA58381
Streptococcus sinensis	HKU4	SAMN02848449	PRJNA251999
Streptococcus sobrinus	DSM 20742	SAMN02743354	PRJNA224116
Streptococcus suis	BM407		PRJNA59321
Streptococcus thermophilus	JIM 8232	SAMEA2272807	PRJNA224116
Streptococcus thoraltensis	DSM 12221	SAMN02256409	PRJNA165451
Streptococcus timonensis	Marseille-P2915	SAMEA4415361	PRJNA224116
Streptococcus troglodytae	TKU31	SAMD00017730	PRJDB2913
Streptococcus uberis	0140J	SAMEA3138207	PRJNA224116
Streptococcus urinalis	2285-97	SAMN02436559	PRJNA224116
Streptococcus varani	FF10	SAMEA3312815	PRJEB8936
Streptococcus vestibularis	NCTC12167	SAMEA3594358	PRJNA224116