Universitat
de les Illes Balears

DOCTORAL THESIS
2023

BACTERIAL WHOLE-GENOME SEQUENCING FOR ESTABLISHMENT OF REFERENCE SEQUENCES, COMPARATIVE GENOMICS, BIOMARKER DISCOVERY AND CHARACTERIZATION OF NOVEL TAXA

Francisco Salvà Serra

Universitat
de les Illes Balears

DOCTORAL THESIS
2023

Doctoral Programme in Environmental and Biomedical Microbiology

BACTERIAL WHOLE-GENOME SEQUENCING FOR ESTABLISHMENT OF REFERENCE SEQUENCES, COMPARATIVE GENOMICS, BIOMARKER DISCOVERY AND CHARACTERIZATION OF NOVEL TAXA

Francisco Salvà Serra

Thesis Supervisor: Antoni Bennasar Figueras
Thesis Supervisor: Edward R. B. Moore
Thesis Supervisor: Hedvig Engström Jakobsson
Thesis Tutor: Antoni Bennasar Figueras

Doctor by the Universitat de les Illes Balears

A la meva familia

Play hard!

Acknowledgements

I would like to start by thanking my PhD supervisors Dr. Antoni Bennasar Figueras (Toni), Dr. Edward R. B. Moore (Ed) and Dr. Hedvig E. Jakobsson. Thank you for all your support, guidance, knowledge, for all the inspiration provided, all your enthusiasm, and for the uncountable number of scientific and soft skills that I have learned from each of you.

Special thanks also to my friend Daniel Jaén Luchoro (Dani), lab colleague first in Mallorca and later in Sweden. Thank you, Dani, for being a great person, friend, and a great scientist. You have been an example to me, and a great source of inspiration. I have learned a lot from you during all these years and I have really enjoyed working with you and forming a super-Mallorcan team! I really wish you all the best in your very promising career, and I hope that we can continue working together for many more years.

I would also like to thank all my other colleagues and people that I have met over the years at the Microbiology Lab of the University of the Balearic Islands (UIB), starting with Jorge Lalucat (Jordi). You are a great reference, a source of inspiration, and the Professor that opened me the door of the Lab when I was a 4th year undergraduate student, thanks. Thanks also to Margarita Gomila (Marga), Balbina Nogales, Rafael Bosch and Elena García-Valdés, for all the teaching that I received from you when I was a degree and master student, and for being so welcoming every time I visit the lab when I go to Mallorca. Thanks also to Toni Busquets (and of course, again, to Marga), for all the moments together in Mallorca and Sweden; right before moving to Sweden, you told me that I would enjoy Sweden, and look, I love it and I am still here more than eight years later, in what I call my second home and where my beloved daughter is growing up.

I also would like to thank Francisco Aliaga Lozano (Xisco), a great person and a great reference in clinical microbiology. Thank you for being always open to help and collaborate. Thanks also to all the other people that I have met at UIB over the years and that, in some way, have enriched me as a person and as a scientist. I would like to mention Farid Beiki, Arantxa Peña, Magda Mulet, Cris, Yasmeiri Mena, Carolina Seguí and Víctor Fernández Juárez, but there are many others. Of course, I thank also all the technicians working in the lab and performing essential tasks, thank you for always being so nice and receiving me with a smile every time I come by the lab.

Special thanks also to my current and past colleagues at the Culture Collection University of Gothenburg (CCUG). I arrived there in December 2014 as a Leonardo da Vinci trainee, and I felt part of the team from day 1. Thank you for always being so welcoming to me and to all the other people that came after me. These include my dear "CCUG ladies" (Sofia Cardew, Christel Unosson, Susanne Jensie-Markopoulos, Maria Ohlén, Elisabeth Inganäs) (Queens of the Collection and clinical bacteriology; thank you for all the knowledge, for being always open to help, and for all the 'fikas' and moments lived together), Liselott Svensson-Stadler (thank you for being a great boss and a great colleague), Kent Molin (King of cell fatty acid analyses, among other things), Derek Cardew (IT King – Thank you for being one of these persons that I am always happy to meet), Beatriz Piñeiro Iglesias (Bea, thank you for being so cheerful and talkative), my friend Lucia Gonzales-Siles (thank you for everything that I have learned from you, for all the research, and for all the moments together), Shora Yazdanshenas (you are great in the lab and you were a great office-mate; thanks for all your smiles and conversations; hope life is going great in Trollhättan), my friend Leonarda Achá Alarcón (Leo, thank you for showing us La Paz, I am so glad to have you as a friend and colleague, and I am sure that your

PhD will be great! – I thank you also for being always so sweet with my beloved Laia (L)), Timur Tunovic (thanks for all your assistance during the year and a half that you were working with us), Anders Karlsson and Roger Karlsson (the Proteomics Kings and Mallorca lovers; the kind of colleagues that everybody would like to work with). Thank you to all of you for being so great and so good at what you do, for all the hundreds of "fika's" together, and for what you have taught to me about microbiology, proteomics, science, and Sweden. Special thanks also to Anders Malmborg (the King of the Substrate Department and of cultivation media), thank you for being such a great person and great head of the Substrate Department of the Sahlgrenska University Hospital. Thank you for being always so kind and open to help. I wish you a great retirement; I am sure that you will enjoy it very much. I also thank the entire Team at the Substrate Department and to all the other colleagues that maintain our labs and facilities, thank you for all your essential support during these years. Thanks also to all my other colleagues at the Sahlgrenska University Hospital, the University of Gothenburg and the Centre for Antibiotic Resistance Research (CARe): Hardis Rabe (my dear ex-office mate, thank you for always being there, ready to listen or have a nice conversation), Kaisa Thorell (Queen of genomics; thanks for all our collaborations and for being always super nice and open), Kristina Nyström (extremely helpful in the lab during my first stage in Sweden), Susann Skovbjerg, Erika Tång Hallbäck, Nahid Karami, Ia Adlerberth, the entire Larsson's group, Johan Bengtsson-Palme (thank you for being always open to talk, and to have constructive, positive and motivating discussions), Emil Burman (great lab mate and beer producer!). Not everyone has been mentioned but thank you to the entire Department of Infectious Diseases of the University of Gothenburg (including its head, Agnes Wold) and the Department of Clinical Microbiology of the Sahlgrenska University Hospital. Thanks also to Mira; thank you for your eagerness to learn microbiology, even though that was not the main goal of your master thesis project. Thank you for being so positive and cheerful. Thanks also to all the other people that I have not mentioned but I have met over the years or that I meet on an almost daily basis and that make my workdays great. I thank also the Sauna Team (especially Maria, Birgitta, Dan and Thomas), for welcoming me with open arms when I was just arrived in Sweden.

Thanks also to all my dear colleagues in Valparaíso (especially to Agustina Undabarrena, Andrés Cumsille, Valentina Méndez, Néstor Sena, Bea Cámara and Michael Seeger), thank you for being always so welcoming and for all the experiences and science done together. You make me feel at home every time I visit Chile. Thanks also to my dear collaborators Danilo Pérez Pantoja and Raúl Donoso, at Metropolitan Technical University (Santiago de Chile), your contributions to Paper VI were essential. Thanks also to my dear colleagues in Bolivia (Dra. Julia, Nataniel, Dra. Volga and all the students that I had), those two weeks in La Paz were unforgettable. Thanks, also, to the entire team and to all the students in Santo Domingo (Dominican Republic), for a such a great experience, course, and conference. Thanks especially to Yasmeiri Mena, ex-colleague and friend at UIB, who in 2017 invited me to visit Santo Domingo, organize a course and give a talk about our proteomics work at the International Conference of Scientific Research.

I would also like to thank my dear friends, colleagues, and collaborators in Bergen (Norway), with special emphasis on Nachiket Marathe and Didrik Grevskott. Our collaboration is being extremely productive, but more importantly, I really enjoy collaborating, working, and learning with you, so thanks a lot for everything, and I really hope that we can continue working together for many more years.

I also would like to thank Ana Carvalheira, friend and collaborator, from Porto (Portugal). Thank you for all the good moments, and for the fantastic and strong team that we built in 2019 – 2020 to lead and finish the description of two novel species of the genus *Acinetobacter*.

Before moving back to the past and to more personal aspects, I would also like to thank the two anonymous Reviewers of this Thesis. Thank you very much for the time and dedication that you have invested in reading, reviewing, and writing a report about this Thesis. Your feedback was very valuable and highly appreciated. I would also like to thank in advance the yet unknown opponents that will form the Evaluation Committee of the Defense of this PhD thesis, for the time and effort that you will invest in reading, assessing the Thesis and acting as opponents.

Going back to the past, I would also like to thank all the teachers that I had in Pinocho (because preschool education really matters) and Montesión. I thank you all for the education, inspiration, and knowledge that I got from you during three and twelve years, respectively. Thank you also so much to all my University teachers. Each of you marked me positively in some way and contributed to my development as a biologist and as a scientist. I also feel very thankful to all the colleagues that I had during my summer jobs (lifeguard, airport ramp agent, mystery shopper, etc.), from 2005 to 2013. None of those were related to science but they all provided me with invaluable experiences and skills. They provided me a broader perspective and helped me to develop and become a better and more resilient person. I would also like to thank all my colleagues at CBBA (Centre Balear de Biologia Aplicada), especially Gaby and Amal. Those three months were a great experience, and I learned a lot from all of you. I would also like to acknowledge CAEB (Confederation of Business Associations of the Balearic Islands) and the Leonardo da Vinci Program, of the European Commission, and again, Toni and Ed. Thanks to you I came to Sweden, which changed my life, and boosted my growth as a person and as a scientist. I also thank the National Doctoral Programme in Infections and Antibiotics (NDPIA, Sweden; www.ndpia.se), for supporting and funding my participation in top-level courses, workshops and conferences that have definitely helped me to grow as a scientist and as a person. I also thank all the other people that I have not mentioned in these lines but with who I have somehow interacted. All of you have somehow had a positive influence on this thesis.

Science apart, I would like to deeply thank all my dear Friends in Mallorca and in Sweden, especially the Son Roqueta City Crew, Biologueiros, “Reyes y Reinas” and the “Viking family”. Thank you to each and all of you for all the good moments, all the parties, all the beers and all the whatever crazy things that we have done and we will do together. You are all amazing.

I also deeply thank my entire family very much. I thank my parents, for all the love, all the values and education that you shared with me, for all your patience, all your efforts, and for doing all your best for me. You have always been there supporting me and motivating me to continue studying and to be the person I am nowadays. I also thank my brother Guillem (Es Pixador), my parents-in-law, my dear “Fem guy”, my sisters-in-law, cousins, uncles, Madrina, Padrino, Tia Bombón, my entire family and family in law. Thank you to all of you. Thank you also to all the beloved ones that are not here anymore (sobretot es pradins, el Tio Miquel i sa Tia Maria), for all the love you shared, and for always being in my heart.

To my dearest wife Nuri, the girl of my life and the mum of my beloved daughter Laia and of our coming second baby, thank you for all your patience and support during all these years, thank you for all the great moments, adventures, and trips that we continuously live together. Finally, I thank my dearest daughter Laia, “Miniyo hembra”, thank you for being as great, sweet, and intense as you are. Thank you for all the smiles, all the hugs, and all the love you share, thank you for being such a great boost of energy, power, and motivation during the last part of this PhD, and thank you for being one more reason to travel around the World and for bringing out the best of me.

List of papers

This Doctoral Thesis is formed by a compendium of the following published papers:

I. **Complete Genome Sequence of *Pseudomonas balearica* DSM 6083^{T}**
Antoni Bennasar-Figueras, Francisco Salvà-Serra, Daniel Jaén-Luchoro, Carolina Seguí, Francisco Aliaga, Antonio Busquets, Margarita Gomila, Edward R. Moore, Jorge Lalucat
Genome Announcements 4(2): e00217-16 (2016)
DOI: 10.1128/genomea.00217-16
2-years impact factor SJR* (2015): n/a
Q3 in Genetics in SJR (2015).

II. **Genome Sequences of Two Naphthalene-degrading Strains of *Pseudomonas balearica*, Isolated from Polluted Marine Sediment and from an Oil Refinery Site**
Francisco Salvà-Serra, Hedvig E. Jakobsson, Antonio Busquets, Margarita Gomila, Daniel Jaén-Luchoro, Carolina Seguí, Francisco Aliaga-Lozano, Elena García-Valdés, Jorge Lalucat, Edward R. B. Moore, Antoni Bennasar-Figueras
Genome Announcements 5(14): e00116-17 (2017)
DOI: 10.1128/genomeA.00116-17
2-years impact factor SJR (2016): n/a
Q3 in Genetics in SJR (2016).

III. **Draft Genome Sequence of *Streptococcus gordonii* Type Strain CCUG 33482^{T}**
Francisco Salvà-Serra, Hedvig E. Jakobsson, Kaisa Thorell, Lucia Gonzales-Siles, Erika T. Hallbäck, Daniel Jaén-Luchoro, Fredrik Boulund, Per Sikora, Roger Karlsson, Liselott Svensson, Antoni Bennasar, Lars Engstrand, Erik Kristiansson, Edward R. B. Moore
Genome Announcements 4(2): e00175-16 (2016)
DOI: 10.1128/genomeA.00175-16
2-years impact factor SJR (2015): n/a
Q3 in Genetics in SJR (2015).

IV. **Complete genome sequences of *Streptococcus pyogenes* type strain reveal 100%-match between PacBio-solo and Illumina-Oxford Nanopore hybrid assemblies**
Francisco Salvà-Serra, Daniel Jaén-Luchoro, Hedvig E. Jakobsson, Lucia Gonzales-Siles, Roger Karlsson, Antonio Busquets, Margarita Gomila, Antonio Bennasar-Figueras, Julie E. Russell, Mohammed Abbas Fazal, Sarah Alexander, Edward R. B. Moore
Scientific Reports 10:116 (2020)
DOI: 10.1038/s41598-020-68249-y
2-years impact factor SJR (2019): 4.149
Q1 in Multidisciplinary in SJR (2019).

V. **Beware of false "type strain" genome sequences**
Francisco Salvà-Serra, Daniel Jaén-Luchoro, Roger Karlsson, Antoni Bennasar-Figueras, Hedvig E. Jakobsson, Edward R. B. Moore
Microbiology Resource Announcements 8(22): e00369-19 (2019)
DOI: 10.1128/MRA.00369-19
2-years impact factor SJR (2018): 0.911
Q4 in Genetics in SJR (2018).

VI. **Comparative genomics of *Stutzerimonas balearica* (*Pseudomonas balearica*): diversity, habitats and biodegradation of aromatic compounds**
Francisco Salvà-Serra, Danilo Pérez-Pantoja, Raúl A. Donoso, Daniel Jaén-Luchoro, Víctor Fernández-Juárez, Hedvig Engström-Jakobsson, Edward R. B. Moore, Jorge Lalucat, Antoni Bennasar-Figueras
Frontiers in Microbiology 14: 1159176 (2023)
DOI: 10.3389/fmicb.2023.1159176
2-years impact factor SJR (2023): 6.064
Q1 in Microbiology in SJR (2023).

VII. **Detection of "Xisco" gene for identification of *Streptococcus pneumoniae* isolates**
Francisco Salvà-Serra, Gwendolyn Connolly, Edward R. B. Moore, Lucia Gonzales-Siles
Diagnostic Microbiology and Infectious Disease 90(4): 248-250 (2018)
DOI: 10.1016/j.diagmicrobio.2017.12.003
2-years impact factor SJR (2016): 2.686
Q1 in Medicine (miscellaneous) in SJR (2016).

VIII. ***Scandinavium goeteborgense* gen. nov., sp. nov., a new member of the family *Enterobacteriaceae* isolated from a wound infection, carries a novel quinolone-resistance gene variant**
Nachiket P. Marathe#, Francisco Salvà-Serra#, Roger Karlsson, D. G. Joakim Larsson, Edward R. B. Moore, Liselott Svensson-Stadler, Hedvig E. Jakobsson
Frontiers in Microbiology 10: 2511 (2019)
DOI: 10.3389/fmicb.2019.02511
# These authors contributed equally to this work.
2-years impact factor SJR (2018): 4.304
Q1 in Microbiology in SJR (2018).

*SJR, Scimago Journal & Country Ranking (www.scimagojr.com). Data source: Scopus®. Q, quartile; n/a: not available.

For additional publications by the author, see: https://orcid.org/0000-0003-0173-560X

List of abbreviations and acronyms

AAI: average amino acid identity

ANIb: average nucleotide identity based on BLAST

ATCC: American Type Culture Collection

ATP: adenosine triphosphate

BLAST: basic local alignment search tool

bp: base pair

CABOG: Celera Assembler with the Best Overlap Graph

CCUG: Culture Collection University of Gothenburg

CIP: Collection de l'Institut Pasteur

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

DDH: DNA-DNA hybridization

dDDH: digital DNA-DNA hybridization

ddNTP: dideoxyribonucleoside triphosphate

dNTP: deoxyribonucleoside triphosphate

DDBJ: DNA Data Bank of Japan

DFAST: DDBJ Fast Annotation and Submission Tool

DNA: deoxyribonucleic acid

DUF: domain of unknown function

ESBL: extended-spectrum β-lactamase

FDA-ARGOS: Food and Drug Administration dAtabase for Regulatory-Grade micrObial Sequences

FRC: feature response curves

G+C: guanine plus cytosine

GCM: Global Catalogue of Microorganisms

gcType: Global Catalogue of Type Strain

gDNA: genomic deoxyribonucleic acid

GEBA: Genomic Encyclopedia of *Bacteria* and *Archaea*

GGDC: Genome-to-Genome Distance Calculator

GOLD: Genomes OnLine Database

GTDB: Genome Taxonomy Database

HGAP: Hierarchical Genome Assembly Process

HGT: horizontal gene transfer

ICE: integrative and conjugative element

ICNP: International Code of Nomenclature of Prokaryotes

IJSEM: International Journal of Systematic and Evolutionary Microbiology

IMAGE: Iterative Mapping and Assembly for Gap Elimination

IME: integrative and mobilizable element

INSDC: International Nucleotide Sequence Database Collaboration

LPSN: List of Prokaryotic names with Standing in Nomenclature

MAG: metagenome-assembled genome

MALDI-TOF MS: matrix-assisted laser desorption ionization time-of-flight mass spectrometry

MiGA: Microbial Genomes Atlas

MLSA: multilocus sequence analysis

MLST: multilocus sequence typing

NCBI: National Center for Biotechnology Information

NCTC: National Collection of Type Cultures

NGS: next-generation sequencing

OGRI: overall genome relatedness index

PAGIT: Post-Assembly Genome-Improvement Toolkit

PCR: polymerase chain reaction

PGAP: Prokaryotic Genome Annotation Pipeline

QUAST: Quality Assessment Tool for Genome Assemblies

REAPR: Recognition of Errors in Assemblies using Paired Reads

RNA: ribonucleic acid

rRNA: ribosomal ribonucleic acid

SAG: single-amplified genome

SNP: single nucleotide polymorphism

SSPACE: SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension

TYGS: Type (Strain) Genome Server

WGS: whole-genome sequencing

Contents

Summary

Bacteria are the most ubiquitous and widely distributed organisms and play major roles in almost any environment. Therefore, studying and understanding their biology is essential to secure the well-being of the planet and humanity. This can be done by determining, analyzing, and characterizing their genome sequences, which has been facilitated thanks to the development of high-throughput DNA sequencing technologies.

In this thesis, whole-genome sequencing methodologies were used for establishing bacterial reference genome sequences, including those of the type strains of species within three selected taxa (Papers I, II, III, IV, VI and VIII): *Stutzerimonas balearica* (formerly, *Pseudomonas balearica*), a marine bacterium with capacities for degrading aromatic compounds; species of the genus *Streptococcus*, which encompasses well-known commensal species as well as major human pathogens; and of the family *Enterobacteriaceae*, an ecologically diverse and taxonomically complex group of bacteria, members of which can be found in many different environments and also can cause an extensive range of diseases in humans. The different methodologies utilized in the studies of this thesis reflect the marked evolution of high-throughput DNA sequencing technologies that has occurred in the last years; this includes the capacity for determining highly accurate complete genome sequences, using the latest long-read sequencing technologies.

These developments have led to vast amounts of publicly available genome sequence data, which are essential for downstream studies, such as those described in Papers VI – VIII. However, not everything is positive ("All that glitters is not gold") about the databases of whole-genome sequences, and Paper V warns users of publicly available genome sequences about the presence of "false" type strain genome sequences and the importance of performing quality controls on the sequence data used in research studies. Subsequently, in Paper VI, the genome sequences determined in Papers I and II were used in combination with publicly available genome sequences to perform a comparative genomic study for elucidating the genomic diversity of *S. balearica* and its potential for biodegradation of aromatic compounds. Genome sequence data also facilitated the establishment of a strategy for detecting additional strains of *S. balearica*, based on 16S rRNA gene signature nucleotide positions and sequence similarities for determining the habitats of the species. In Paper VII, hundreds of genome sequences of the Mitis-Group of the genus *Streptococcus* allowed the determination of a biomarker gene specific for the human pathogen *Streptococcus pneumoniae* and the establishment of a PCR-based species-specific assay for differentiating *S. pneumoniae* from closely-related species, which has often hindered accurate identification. In Paper VIII, whole-genome sequencing, in combination with publicly available type strain genome sequences, enabled the confirmation that a clinical isolate of the family *Enterobacteriaceae*, which was not able to be further identified at clinical laboratories, represents a novel genus and species (*Scandinavium goeteborgense*) within the family *Enterobacteriaceae* (*Scandinavium goeteborgense*) and to accurately determine its taxonomic position.

The specific contributions of this thesis exemplify and demonstrate that the latest developments of high-throughput DNA sequencing and whole-genome sequencing have certainly pushed the limits of microbiology and life sciences to a next level, in which we can establish solid grounds for down-stream research and applications and explore the genomic insights of bacteria with extremely high resolution.

Resum

Els bacteris són els organismes més ubics i àmpliament distribuïts, a més de desenvolupar rols importants en pràcticament qualsevol ambient que habiten. Per tant, l'estudi i la comprensió de la seva biologia són essencials per garantir el benestar del planeta i de la humanitat. El desenvolupament de les tecnologies de seqüenciació d'ADN d'alt rendiment ha permès dur a terme l'abordatge d'aquests aspectes mitjançant l'anàlisi i la caracterització de les seqüències genòmiques.

En aquesta tesi, es van utilitzar metodologies de seqüenciació de genomes per a establir seqüències de referència de genomes de diversos bacteris d'interès clínic o ambiental; incloent-hi soques tipus d'espècies de tres tàxons (articles I, II, III, IV, VI i VIII): *Stutzerimonas balearica* (anteriorment, *Pseudomonas balearica*), un bacteri marí amb capacitats per a degradar compostos aromàtics; espècies del gènere *Streptococcus*, que abasta espècies comensals i també importants patògens humans; i de la família *Enterobacteriaceae*, un grup de bacteris ecològicament divers i taxonòmicament complex, els membres del qual es poden trobar en ambients molt diversos i alguns dels quals poden causar una àmplia gamma de malalties en humans. Les diferents metodologies utilitzades en els estudis d'aquesta tesi reflecteixen la marcada evolució de les tecnologies de seqüenciació d'ADN d'alt rendiment que ha tingut lloc en els últims anys. En aquest sentit, cal destacar la capacitat per determinar amb gran exactitud seqüències completes de genomes, fent servir les últimes tecnologies de seqüenciació de lectura llarga.

Aquests avanços han donat lloc a la generació de grans quantitats de dades de seqüenciació de genomes, els quals estan disponibles en bases de dades públiques i són essencials per a estudis posteriors, com els descrits en els articles VI - VIII. No obstant això, no és or tot el que lluu i en aquest cas no tot és positiu sobre les bases de dades de seqüències de genomes complets. Així, l'article V adverteix els usuaris de bases de dades de seqüències públiques sobre la presència de seqüències de genomes assignades "falsament" a soques tipus; així com de la importància de realitzar controls de qualitat en les dades públiques, especialment en aquelles que s'utilitzin com a referències. Posteriorment, en l'article VI, les seqüències de genomes determinades en els articles I i II van ser utilitzades en combinació amb seqüències públiques de genomes per a realitzar un estudi genòmic comparatiu i elucidar la diversitat genòmica de *S. balearica* i el seu potencial per a degradar compostos aromàtics. Les seqüències de genomes també van facilitar l'establiment d'una estratègia per a detectar soques addicionals de *S. balearica*, basada en els nucleòtids signa del gen 16S rRNA i similitud de seqüència per a determinar els hàbitats de l'espècie. En l'article VII, l'ús de centenars de seqüències de genomes del Grup Mitis del gènere *Streptococcus* van permetre la determinació d'un gen biomarcador específic per al patogen humà *Streptococcus pneumoniae* i l'establiment d'un assaig PCR per a diferenciar *S. pneumoniae* d'espècies estretament relacionades, que sovint han obstaculitzat la seva identificació precisa. En l'article VIII, la seqüenciació de genomes, en combinació amb seqüències públiques de genomes de soques tipus, va permetre confirmar que un aïllat clínic de la família *Enterobacteriaceae*, el qual no havia pogut ser identificat als laboratoris clínics aplicant les metodologies clàssiques basades en el cultiu, representa una nova espècie i nou gènere (*Scandinavium goeteborgense*) de la família *Enterobacteriaceae*, i va permetre determinar amb precisió la seva posició taxonòmica.

Les contribucions específiques d'aquesta tesi exemplifiquen i demostren que els darrers avenços en la seqüenciació d'ADN d'alt rendiment i en la seqüenciació de genomes sencers han portat al camp de les ciències de la vida en general i de la microbiologia en particular a un nivell

superior, on podem establir bases sòlides per a investigacions i aplicacions posteriors i explorar les entranyes genòmiques dels bacteris amb una resolució extremadament alta.

Resumen

Las bacterias son los organismos más ubicuos y ampliamente distribuidos, además de desempeñar roles importantes en prácticamente cualquier ambiente que habitan. Por lo tanto, el estudio y la comprensión de su biología son esenciales para garantizar el bienestar del planeta y de la humanidad. El desarrollo de las tecnologías de secuenciación de ADN de alto rendimiento ha permitido llevar a cabo el abordaje de estos aspectos mediante el análisis y caracterización de sus secuencias genómicas.

En esta tesis, se utilizaron metodologías de secuenciación de genomas para establecer secuencias de referencia de genomas de varias bacterias de interés clínico o ambiental, incluyendo cepas tipo de especies de tres taxones (artículos I, II, III, IV, VI y VIII): *Stutzerimonas balearica* (anteriormente, *Pseudomonas balearica*), una bacteria marina con capacidades para degradar compuestos aromáticos; especies del género *Streptococcus*, que abarca especies comensales y también importantes patógenos humanos; y de la familia *Enterobacteriaceae*, un grupo de bacterias ecológicamente diverso y taxonómicamente complejo, cuyos miembros se pueden encontrar en ambientes muy diversos y algunos de los cuales pueden causar una amplia gama de enfermedades en humanos. Las diferentes metodologías utilizadas en los estudios de esta tesis reflejan la marcada evolución de las tecnologías de secuenciación de ADN de alto rendimiento que ha tenido lugar en los últimos años. En este sentido, cabe destacar la capacidad para determinar con gran exactitud secuencias completas de genomas, utilizando las últimas tecnologías de secuenciación de lectura larga.

Estos avances han dado lugar a la generación de grandes cantidades de datos de secuenciación de genomas, los cuales están disponibles en bases de datos públicas y son esenciales para estudios posteriores, como los descritos en los artículos VI - VIII. Sin embargo, no es oro todo lo que reluce y en este caso no todo es positivo acerca de las bases de datos de secuencias de genomas completos. Así, el artículo V advierte a los usuarios de bases de datos de secuencias públicas acerca de la presencia de secuencias de genomas asignadas “falsamente” a cepas tipo; así como de la importancia de realizar controles de calidad en los datos públicos, especialmente en aquellos que vayan a ser utilizados como referencias. Posteriormente, en el artículo VI, las secuencias de genomas determinadas en los artículos I y II fueron utilizadas en combinación con secuencias públicas de genomas para realizar un estudio genómico comparativo y elucidar la diversidad genómica de *S. balearica* y su potencial para biodegradar compuestos aromáticos. Las secuencias de genomas también facilitaron el establecimiento de una estrategia para detectar cepas adicionales de *S. balearica*, basada en los nucleótidos firma del gen 16S rRNA y similitud de secuencia para determinar los hábitats de la especie. En el artículo VII, el uso de cientos de secuencias de genomas del Grupo Mitis del género *Streptococcus* permitieron la determinación de un gen biomarcador específico para el patógeno humano *Streptococcus pneumoniae* y el establecimiento de un ensayo PCR para diferenciar *S. pneumoniae* de especies estrechamente relacionadas, que a menudo han obstaculizado su identificación precisa. En el artículo VIII, la secuenciación de genomas, en combinación con secuencias públicas de genomas de cepas tipo, permitió confirmar que un aislado clínico de la familia *Enterobacteriaceae*, el cual no había podido ser identificado en los laboratorios clínicos aplicando las metodologías clásicas, basadas en el cultivo, representa una nueva especie y nuevo género (*Scandinavium goeteborgense*) dentro de la familia *Enterobacteriaceae*, y permitió determinar con precisión su posición taxonómica.

Las contribuciones específicas de esta tesis ejemplifican y demuestran que los últimos avances en la secuenciación de ADN de alto rendimiento y en la secuenciación de genomas enteros, han

llevado al campo de las ciencias de la vida en general y de la microbiología en particular a un nivel superior, en el que podemos establecer bases sólidas para investigaciones y aplicaciones posteriores y explorar las entrañas genómicas de las bacterias con una resolución extremadamente alta.

Sammanfattning

Bakterier finns överallt och spelar en stor roll i många olika miljöer. Att studera och förstå deras biologi är viktig för att bättre förstå dess relation och påverkan på miljön och mänsklig hälsa. Detta kan göras genom att karakterisera och analysera deras genomsekvenser, något som har underlättats med den revolutionerande utvecklingen av DNA-sekvenseringsteknologier.

I denna avhandling användes metoder för sekvensering av hela genom för att fastställa bakteriella referenser för helgenomsekvenser, inklusive de typstammarna av arter inom tre utvalda taxa (artikel I, II, III, IV, VI och VIII): *Stutzerimonas balearica* (tidigare, *Pseudomonas balearica*), en marin bakterie med förmåga att bryta ned aromatiska föreningar; arter av släktet *Streptococcus*, som omfattar välkända kommensala arter såväl som viktiga mänskliga patogener; och av familjen *Enterobacteriaceae*, en ekologiskt mångfaldig och taxonomiskt komplex grupp av bakterier, vars medlemmar kan hittas i många olika miljöer och också kan orsaka ett stort antal sjukdomar hos människor. De olika sekvenseringsmetoder som användes i denna avhandling återspeglar den markanta utvecklingen av DNA-sekvenseringsteknologier som har skett de senaste åren; detta inkluderar även möjligheten till att bestämma kompletta genomsekvenser.

Den ökade mängden av sekvenserade bakteriegenom har bidragit till mängder med offentligt tillgängliga genomsekvensdata, såsom de som beskrivs i artiklar VI – VIII. Allt är inte positivt med detta ("Allt som glittrar är inte guld"), och artikel V varnar användare för användandet av allmänt tillgängliga genomsekvenser och om förekomsten av "falska" typstamgenomsekvenser och vikten av att utföra kvalitetskontroller av sekvensdata. I artikel VI, användes genomsekvenserna som karakteriserades i artiklarna I och II i kombination med allmänt tillgängliga genomsekvenser för att utföra en jämförande genomisk studie för att belysa den genomiska mångfalden av *S. balearica* och dess potential för biologisk nedbrytning av aromatiska föreningar. Genomsekvensdata underlättade också upprättandet av en strategi för att detektera ytterligare stammar av *S. balearica*, baserad på 16S rRNA-gensignaturnukleotidpositioner och sekvenslikheter för att bestämma artens livsmiljöer. I artikel VII möjliggjorde hundratals genomsekvenser av Mitis-gruppen av släktet *Streptococcus* bestämning av en biomarkörgen specifik för den mänskliga patogenen *Streptococcus pneumoniae* och upprättandet av en PCR-baserad artspecifik analys för att skilja *S. pneumoniae* från närabesläktade arter, vilket ofta har hindrat korrekt identifiering. I artikel VIII möjliggjorde helgenomsekvensering, i kombination med allmänt tillgängliga genomsekvenser av typstam, bekräftelsen av ett kliniskt isolat inom familjen *Enterobacteriaceae*, som inte tidigare identifierats vid något kliniskt laboratorium, och dessutom ett nytt släkte och art (*Scandinavium goeteborgense*) inom familjen *Enterobacteriaceae*.

De specifika bidragen från denna avhandling exemplifierar och visar att den senaste utvecklingen av DNA-sekvensering verkligen har flyttat gränserna för mikrobiologi och biovetenskap till nästa nivå, där vi kan etablera solida grunder för nedströms forskning och tillämpningar och utforska genomiska insikter av bakterier med extremt hög upplösning.

1 Introduction

1.1 Domain *Bacteria*

Bacteria are single-celled, prokaryotic (*pro*, 'before'; *karyon*, 'kernel', referring to a membrane-bound nucleus) microorganisms that constitute one of the three domains of life proposed by Carl Woese *et al*. (Woese and Fox, 1977; Woese et al., 1990). Since their emergence, nearly four billion years ago (Schopf, 1993), they have managed to adapt and colonize nearly all environments on Earth, including those with the most extreme conditions for life, thus becoming the most ubiquitous and widespread group of organisms and playing major roles in almost every ecosystem. Thus, bacteria are involved in a wide range of biogeochemical cycles and are essential for the stability of the existing Earth conditions and environments. They virtually affect all aspects of life, generally causing innumerable beneficial effects, but regularly also harms. Furthermore, bacteria represent a large fraction of the microbiome of animals and plants. For instance, the estimated bacteria-to-human cells ratio is about 1:1 (Sender et al., 2016) and thus, bacteria play key roles in human health, on one side providing numerous benefits that are vital for the human body to function properly and on the other side sometimes causing a wide range of infectious diseases. Consequently, studying and understanding the biology of bacteria is crucial for us humans to have optimal and sustainable interactions with our planet that warrant the prosperity and well-being of all parts.

1.2 DNA and whole-genome sequencing

Deoxyribonucleic acid (DNA) is a polymer composed by nucleotides that contains the genetic information of living organisms, encoding the basis of their biology. The entire DNA set of an organism forms its genome. Thus, determining, analyzing and understanding the DNA composition and nucleotide order of the genome of a particular organism can provide crucial information about its lifestyle and characteristics, so much so that whole-genome sequencing has become an essential component and often requirement for numerous downstream applications of most basic and applied fields of biological sciences.

The molecular structure of DNA was first unveiled by Watson and Crick in 1953, after several years of work, which was heavily based on the work of several other researchers (Watson and Crick, 1953). However, it was not until two decades later that the first main methodologies for determining sequences of DNA emerged and initiated one of the major revolutions in life sciences.

1.2.1 First DNA sequencing methods

In 1975, Frederick Sanger and Alan Coulson presented the first enzymatic sequencing method, the "plus and minus" method (Sanger and Coulson, 1975), which depended on the primer-based incorporation of nucleotides (one of them radiolabeled) by a DNA polymerase, to obtain radiolabeled oligonucleotides of different sizes, followed by the four "plus" reactions and the four "minus" reactions. On the one hand, the "plus" reactions incorporated a single type of nucleotide (thus, the last incorporated nucleotide was known). On the other hand, in the "minus" reactions, three of the four nucleotides were used, so that the missing nucleotide after the last incorporated one could be deduced. Subsequently, an eight-lane acrylamide gel electrophoresis was run, to separate the oligonucleotides and determine the sequence by radioautography.

During those years, another successful DNA sequencing method was developed by Allan Maxam and Walter Gilbert (Maxam and Gilbert, 1977), which was based on chemical cleavages. Briefly, DNA radiolabeled on one end was cleaved using four chemical reactions with different base-specificities (adenine, cytosine, adenine/guanine, and cytosine/thymine); the cleaving reactions were followed by a four-lane polyacrylamide gel electrophoresis, which ordered the cleaved fragments by size, which was subsequently autoradiographed to reveal the sequence.

Two years after the publication of the "plus and minus" method, a second method was published by Frederick Sanger, similar to the first method, but based on the use of terminal nucleotides. This was the so-called "chain-termination", later termed, the "Sanger sequencing" method (Sanger et al., 1977b), which would become one of the most important and widely-used DNA sequencing methods in history. Briefly, this approach was based on using radiolabeled dideoxyribonucleoside triphosphates (ddNTP), i.e., analogues of deoxyribonucleoside triphosphates (dNTP) that lack the 3'-hydroxyl group required for binding the 5'-phosphate of the following dNTP to extend the DNA strand. Four different reactions were performed (one per nucleotide) using a mixture of dNTPs and radiolabeled ddNTPs. Thus, DNA chains were extended by DNA polymerase until a ddNTP was randomly incorporated, terminating the extension and yielding radiolabeled DNA fragments of different sizes that were subsequently separated on polyacrylamide gel electrophoreses and autoradiographed.

Not surprisingly, Frederick Sanger and Walter Gilbert were jointly awarded one half of the Nobel Prize in Chemistry in 1980 "*for their contributions concerning the determination of base sequences in nucleic acids*". However, because of its ease of use, robustness and accuracy, the Sanger sequencing method would soon become the more widely used of the two and numerous improvements and optimizations would be made to it in the following years. A major development was the replacement of radiolabel-based detection by detection of fluorophores, initially labelling primers (Smith et al., 1985), and the development of the first semi-automatic DNA sequencing machine (Smith et al., 1986), introduced into the market by Applied Biosystems Inc. One year later, the first fluorescent chain-terminating dideoxynucleotides were developed, which allowed the four Sanger sequencing reactions to be performed jointly and analyzed in a single-lane polyacrylamide gel electrophoresis, based on laser excitation and fluorescence detection (Prober et al., 1987). Later, improved yield and resolution were achieved by using capillary-based electrophoresis (Swerdlow and Gesteland, 1990).

1.2.2 Additional developments and whole-genome sequencing milestones

The development of the first effective and practical DNA sequencing methods, opened the door to an endless number of possibilities, and consequently, contributed to motivate the development of additional methods and resources, and brought with it the achievement of numerous sequencing milestones. A major landmark in DNA sequencing history was the publication of the first sequenced genome, the genome of the bacteriophage, φX174, which was sequenced provisionally, using the "plus and minus" method (Sanger et al., 1977a) and completed afterwards, using the "chain-termination" method (Sanger et al., 1978). Soon after the publication of the φX174 whole-genome sequence, shotgun sequencing (i.e., random fragmentation, cloning and sequencing of DNA fragments) plus computerized assembly strategies was proposed (Staden, 1979) and started to be used, for instance in the sequencing of the genome of bacteriophage lambda (Sanger et al., 1982). Another milestone that was reached during those years was the determination of the genome sequence of the human mitochondrion (Anderson et al., 1981).

All these advances and successes had a direct consequence: large amounts of sequence data were being generated. This led to the creation of GenBank in 1982 (Sayers et al., 2022), with 680,338 bases distributed among 606 sequences in December that year. Less than five years later, in February 1987, GenBank already contained 1.1 M bases distributed among 10,913 sequences (GenBank and Whole Genome Shotgun Statistics: https://www.ncbi.nlm.nih.gov/genbank/statistics/). Later, Kary Mullis and collaborators developed the polymerase chain reaction (PCR) technique, with a thermostable polymerase, as a method for DNA amplification (Saiki et al., 1988). This crucial and revolutionary technique allowed *in vitro* amplification of specific DNA fragments, to generate pure and large amounts of template DNA required for sequencing.

All of these and a myriad of other developments that were nicely reviewed on the 40th anniversary of DNA sequencing (Heather and Chain, 2016; Shendure et al., 2017) contributed to an increased throughput and capacity for sequencing DNA and genomes. Altogether, this allowed numerous laboratories to access DNA sequencing techniques and stimulated the foundation of large sequencing projects, such as the renowned Human Genome Project, which started in 1990, aiming to determine the nucleotide sequence of the human genome and map all the genes. However, despite the successful establishment of numerous DNA sequencing landmarks, huge efforts still had to be done to obtain large amounts of sequence reads. For instance, it took 11 years and the efforts of hundreds of scientists, to present the first draft sequence of the human genome (International Human Genome Sequencing Consortium, 2001). The Project was finally completed in 2003, with a "nearly-complete" sequence (International Human Genome Sequencing Consortium, 2004) and a total cost of USD 2.9 billion.

While the Human Genome Project was running, numerous other historical sequencing milestones were accomplished. A major one was the determination of the first complete genome sequence of a bacterium and free-living organism: *Haemophilus influenzae* (Fleischmann et al., 1995). Immediately after, in 1996, an international consortium presented the first complete genome sequence of a eukaryote, the yeast, *Saccharomyces cerevisiae* (Goffeau et al., 1996), and researchers from the US presented the first complete genome sequence of a representative of the domain *Archaea* (Bult et al., 1996). Two years later, the first genome sequence of a multicellular organism (*Caenorhabditis elegans*) was presented (The *C. elegans* sequencing consortium, 1998), and, later the first genome sequence of a plant (*Arabidopsis thaliana*) (The Arabidopsis Genome Initiative, 2000).

During the years that followed the determination of the complete genome sequence of *H. influenzae*, numerous other sequencing landmarks were accomplished, especially in the bacteriology field. For instance, in 1996, *Mycoplasma* became the first genus to have complete genome sequences of two species (*Mycoplasma genitalium* and *Mycoplasma pneumoniae*) (Fraser et al., 1995; Himmelreich et al., 1996). Soon after, the first complete genome sequence of *Escherichia coli* (strain K-12) was determined, after almost six years of work (Blattner et al., 1997). Two years later, *Helicobacter pylori* became the first bacterial species to have genome sequences of two different strains (Alm et al., 1999), opening the door to intra-species comparative genomics. In year 2000, Stover *et al*. presented the first genome sequence of a member of the genus *Pseudomonas*, i.e., the complete genome sequence of *Pseudomonas aeruginosa* PAO1 (Stover et al., 2000); and in 2001, the complete genome sequence of *Streptococcus pyogenes* strain SF370, the first of the genus *Streptococcus* (Ferretti et al., 2001).

All these sequencing projects were extremely laborious and required huge amounts of funding, workforce, and time. Nonetheless, during those years, novel and high-throughput DNA sequencing approaches were already being developed, and in 1996, Ronaghi and collaborators

presented “pyrosequencing”, a novel real-time DNA sequencing approach, based on the detection of the light produced by a cascade reaction triggered by the incorporation of nucleotides (Ronaghi et al., 1996). The pyrosequencing approach would later become the basis of the first “next-generation sequencing” technology to be introduced into the market: Roche 454.

1.2.3 Next-generation DNA sequencing technologies

The Sanger sequencing method allowed the achievement of numerous whole-genome sequencing milestones, including the determination of genome sequences of viruses, bacteria, archaea, plants and animals. However, despite the advancements and automatizations, the method continued to be laborious and costly, thus being a bottleneck in most sequencing projects and restricting the access to whole-genome sequencing to a few specialized laboratories. This situation encouraged the development of alternative sequencing methods that could provide higher-throughput. Thus, three decades after the publication of the first DNA sequencing methods, the first so-called “second generation” or “next-generation sequencing” (NGS) methodologies came into the market, enabling high-throughput DNA sequencing by massive parallelization, a cost decrease, and therefore a democratization of genome sequencing. These technologies allowed the determination of thousands and millions of DNA sequences (so-called “reads”) in a single sequencing run. In the following lines, three of the most successful and widely used NGS technologies will be briefly introduced, with special focus on whole-genome sequencing: Roche 454 (currently discontinued), Illumina and Ion Torrent,

1.2.3.1 Roche 454

This was the first NGS technology to be introduced into the market, in 2005, and initially one of the most successful ones, although Roche announced its discontinuation in 2013. This technology was based on the pyrosequencing approach (Ronaghi et al., 1996), which was licensed for commercialization to 454 Life Sciences Corp. (Branford, CT, USA; acquired by Roche in 2007) (Margulies et al., 2005).

During library preparation of genomic DNA (gDNA), gDNA is fragmented[1], size selected, ligated to common adapters and immobilized on magnetic beads (optimally one fragment per bead), which are subsequently emulsified in a water-oil solution. Thus, each bead is enclosed in a microreactor that contains all the necessary reagents to perform an emulsion PCR (Nakano et al., 2003; Margulies et al., 2005), which will result in numerous copies of the same fragment on each bead. After PCR amplification, the emulsions are broken, the spheres containing the DNA fragments are selected, the sequencing primer is added, and each bead is introduced in a picolitre sequencing well, on a fiber-optic slide, where the pyrosequencing takes place (Margulies et al., 2005).

In each sequencing cycle, one of the four deoxynucleoside triphosphates (dNTPs) is added (sequencing by synthesis). If the nucleotide is incorporated, a pyrophosphate group is released, which triggers an enzymatic cascade. When released, the pyrophosphate group is converted into adenosine triphosphate (ATP) by ATP sulfurylase. Subsequently, ATP serves as substrate for luciferase to convert luciferin to oxyluciferin, a reaction that generates a measurable light peak. The light intensity is proportional to the number of nucleotides incorporated, which allows the sequence determination. Thus, thousands of sequencing reactions are run in parallel on a

[1]Note that other sequencing approaches (e.g., targeted amplicon sequencing) do not require gDNA as starting material and therefore may have slightly different library preparation approaches that for instance do not require fragmentation. Nevertheless, gDNA is presented as the starting material for library preparation because this is the most commonly used starting material for whole-genome sequencing, which is the main focus of this thesis.

picotiter plate, yielding thousands of reads of lengths of as many as 600 base pairs (bp). Thus, Roche 454 was able to provide relatively long sequence reads; however, it had difficulties determining the sequence of homopolymers, because of the difficulty to distinguish between large peaks of light (Goodwin et al., 2016).

1.2.3.2 Illumina

This technology was initially developed by Solexa Inc. (Saffron Walden, UK) and later acquired by Illumina, Inc. (San Diego, CA, USA; www.illumina.com). During library preparation, gDNA is fragmented, size-selected, and two different adapters are ligated. Subsequently, the DNA fragments are clonally amplified (i.e., amplified from a single DNA molecule) by "bridge PCR" (also called "aka cluster PCR") (Fedurco et al., 2006; Bentley et al., 2008), on a solid surface that contains the PCR primers attached, which are complementary to the adapters. Briefly, the DNA fragments are, first, denaturalized, separating the double strands, and one of the adapters is attached to a primer on the solid support. After attachment, a polymerase synthesizes the complementary strand. The resulting double stranded DNA is denatured, and the original strand eliminated. Afterwards, the free adapter of each fragment is attached to the solid surface containing complementary primers, thus forming a bridge of single-stranded DNA. At this point, the DNA is amplified and as many as 1,000 copies of each fragment are obtained, fixed on the surface, forming clonal DNA clusters.

After amplification, the sequencing is done also by synthesis, using DNA polymerase, a sequencing primer, and nucleotides with reversible terminators (Bentley et al., 2008). In each sequencing cycle, a mixture of the four nucleotides is added, each one labelled with a different fluorochrome; although, in each cycle, only one of them will be incorporated. After incorporation, the fluorochromes of each cluster are excited by a laser beam, which results in the emission of detectable fluorescence. After imaging, the fluorochromes are cleaved and the 3'-OH groups restored, before starting a new cycle. Illumina sequencing can provide millions of reads of lengths of as many as 300 bp, with high accuracy and good confidence (Goodwin et al., 2016). So far, Illumina is the most successful and, currently, the most widely used NGS platform.

1.2.3.3 Ion Torrent

Developed by Ion Torrent Systems, Inc. (Gilford, CT, USA; acquired by Life Technologies Corporation, acquired later by Thermo Fisher Scientific; https://www.thermofisher.com/se/en/home/brands/ion-torrent.html) (Rothberg et al., 2011) and, also, based on sequencing by synthesis. Similarly to Roche 454, during library preparation, gDNA is fragmented, size-selected and adapters ligated. Subsequently, clonal DNA libraries are created, using emulsion PCR. Beads with the amplified DNA fragments are then introduced in sequencing wells. The sequencing takes place by adding nucleotides in turns. The incorporation of each nucleotide results in the release of an H^+ ion. The release of H^+ ions causes a pH decrease in the well, which is proportional to the number of nucleotides incorporated. The pH decrease is measured by a sensor, which allows sequence determination (Rothberg et al., 2011). Ion Torrent can provide millions of reads to an average length of 400 bp and can have running times as short as two hours, which makes it convenient for clinical applications. However, it has problems for determining homopolymers, due to the difficulty to distinguish between large pH drops (Goodwin et al., 2016).

1.2.4 Third-generation DNA sequencing technologies

Soon after the success of the NGS technologies, which enabled high-throughput DNA sequencing, alternative sequencing technologies were developed and introduced into the field

of DNA sequencing, giving birth to the so-called "third generation" DNA sequencing technologies. While the major NGS technologies rely on PCR to produce clonal DNA clusters for sequencing (i.e., template amplification), third-generation sequencing technologies are able to determine the sequence of single DNA molecules, which facilitates faster turn-around times, avoids PCR bias and enables direct detection of base modifications such as methylation (Flusberg et al., 2010). Additionally, while the NGS technologies produce sequence reads of only a few hundred base pairs, third-generation sequencing technologies can routinely produce reads of tens of kilobases length. Thus, these technologies have enabled single molecule, real-time sequencing. In the following sections, the two most successful third-generation sequencing technologies (PacBio and Oxford Nanopore) will be briefly presented.

1.2.4.1 PacBio

Developed by Pacific Biosciences of California, Inc. (Menlo Park, CA, USA; www.pacb.com) (Eid et al., 2009). PacBio sequencing (defined by the company as "Single Molecule, Real-Time Sequencing") is based on the optical observation of DNA synthesis, in real time. Briefly, after gDNA size-selection, circular single-stranded DNA templates are generated using hairpin adapters that cap the ends of the fragments. Subsequently, single template molecules are deposited in wells (zero-mode waveguides) (Levene et al., 2003) that contain a single highly-processive polymerase fixed at the bottom. During synthesis, the polymerase incorporates dNTPs labeled with nucleotide-specific fluorophores that are excited by a laser beam in order to emit detectable fluorescence, which allows sequence determination. Additionally, the polymerase performs multiple laps to the circular template, which results in multiple reads of the same template and enables the determination of a consensus sequence of higher quality for each template (Eid et al., 2009). PacBio raw reads have a relatively high error rate. However, the nearly-random nature of the errors allows the generation of composite genome sequences of extremely high-quality (Koren et al., 2013; Goodwin et al., 2016).

1.2.4.2 Oxford Nanopore

This sequencing technology was developed by Oxford Nanopore Technologies Ltd. (Oxford, UK; www.nanoporetech.com), aiming "*to enable the analysis of any living thing, by any person, in any environment*". This technology offers simple and rapid library preparation kits, as well as the possibility of sequencing via a small, portable, USB-powered device (i.e., MinION). Additionally, it offers the possibility of skipping the size selection step, which enables the possibility of sequencing DNA fragments of any length. With the Oxford Nanopore Rapid Sequencing kits, DNA fragments of any size are attached to adapters, one of which contains a motor enzyme. Subsequently, the library is loaded on to a Flow Cell containing nanopores, through which an ion current is passing. Afterwards, the motor enzyme starts to push single stranded DNA through the pores, thus causing disruptions of the ion flow. Different nucleotides cause different disruptions, thus allowing sequence determination (Goodwin et al., 2016).

This technology has started a revolution by itself in the field of DNA sequencing, enabling rapid and relatively cheap on-site high-throughput DNA sequencing, reducing the dependence of sequencing at specialized core facilities or companies, and providing from short to ultra-long sequence reads (up to >2 Mb). Oxford Nanopore reads typically also have a relatively high error rate and, even though high-quality genome sequences can be obtained, the non-random distribution of the errors (e.g., systematic problems with homopolymers), make it recommended to combine it with a technology, such as Illumina, which is able to compensate for those errors. However, later developments have enabled the obtention of near-finished bacterial genome sequences using only Oxford Nanopore sequencing (Sereika et al., 2022).

1.2.5 From DNA sequence reads to genome sequence

A major challenge that came with advances in whole-genome sequencing and the generation of large amounts of DNA sequence reads was the necessity to assemble the data into accurate draft or complete genome sequences (Pop, 2009). This process is especially challenging if one is using only short reads, because genomes usually contain multicopy regions, such as the ribosomal RNA operons and large repetitive regions (Kingsford et al., 2010; Schmid et al., 2018). Many parts of genomes can be assembled easily when using an adequate genome coverage (e.g., 30x) and, indeed, increasing the coverage can provide less fragmented assemblies. However, at a certain point, a saturation level will be reached and increased coverage will result in no improvement in the genome assembly, only higher computational costs (Salvà Serra, 2014). This is simply because current technology-based short reads are not long enough to resolve these complex regions; although this problem was partially alleviated by the development of paired-end (Edwards et al., 1990) and mate pair sequencing strategies (Wetzel et al., 2011).

In any case, the ultimate and most effective solution to beat these assembly challenges, is to use ultra-long reads, such as those provided by third-generation sequencing technologies, particularly Oxford Nanopore sequencing, the technology that holds the current sequence read length record of 2.3 Mb (Payne et al., 2018) and that is able to resolve repetitive regions of several tens of kilobases (Schmid et al., 2018). Thus, numerous pieces of software have been developed, most open access but also proprietary, to deal with the different aspects of whole-genome sequencing data analysis.

1.2.5.1 Quality control and filtering of DNA sequence reads

The first and key step, when receiving high-throughput DNA sequencing data is to perform a quality control analysis. Several tools have been developed for this purpose, such as FastQC (Andrews, 2010), which is widely used for NGS reads, or NanoPlot, specialized in quality assessment of long-read datasets (De Coster et al., 2018). Subsequently, reads can be trimmed and filtered, according to their quality scores, in order to reduce the number of erroneously called bases that are used in the assembly process. For short reads, this can be done with tools such as Sickle (Joshi and Fass, 2011), Trimmomatic (very widely-used) (Bolger et al., 2014), or the more recently presented, FastP (Chen et al., 2018). For long reads, this can be done using tools such as NanoFilt (De Coster et al., 2018) or Filtlong (Wick, 2017).

1.2.5.2 Assembly of genome sequences

After quality control, trimming and filtering, the reads need to be assembled into contigs (i.e., contiguous sequences), which can then be joined into scaffolds (i.e., an ordered and oriented set of contigs), to obtain draft or complete genome sequences. To achieve that, the most commonly used strategy is *de novo* assembly, which consists of an assembly of reads from scratch, without any reference.

Throughout the years, multiple algorithms have been developed for *de novo* assembly of short and long sequence reads. Most of the assembly algorithms can be classified into three main classes (Li et al., 2011; Liao et al., 2019): overlap-layout-consensus; de Bruijn graphs (Idury and Waterman, 1995); and string graphs (Simpson and Durbin, 2010). These algorithms have been implemented in numerous software tools for genome assembly (so-called “assemblers”), each with different strengths and weaknesses and that can differ significantly in performance (Earl et al., 2011; Salzberg et al., 2012; Bradnam et al., 2013; Magoc et al., 2013; Wick and

Holt, 2019). Additionally, the continuous software and algorithm developments make genome assembly a dynamic and rapidly evolving field, with new methods and tools being presented every year and with many of the already existing tools evolving rapidly, in order to improve performance, offer new possibilities and to become adapted to the latest advances in DNA sequencing technologies. In addition to the three main algorithms, genome assemblers can also be classified according to their capacity to assemble short reads, long reads and to perform hybrid assemblies.

It is important to note that short reads might be single or paired (i.e., paired-end or mate pair sequencing) (Edwards et al., 1990; Wetzel et al., 2011). Both paired-end and mate pair facilitate the assembly task by providing an estimated distance between paired reads, which depends on the size of the inserts generated during the library preparation. Paired-end reads are typically generated using short insert sizes (e.g., several hundreds of base pairs) (Edwards et al., 1990), while mate pairs are typically generated using longer insert sizes (e.g., multiple kilobases; https://www.illumina.com/science/technology/next-generation-sequencing/mate-pair-sequencing.html), which further facilitates the assembly of complex regions of the genomes (Wetzel et al., 2011).

Thus, together with the development of NGS, numerous bioinformatics tools were developed, specialized in the assembly of relatively short read sequences, such as CABOG (Celera Assembler with the Best Overlap Graph; designed to perform hybrid assemblies of reads from Roche 454 and Sanger sequencing) (Miller et al., 2008) and Newbler (which was developed and maintained by 454 Life Sciences), that are based on overlap-layout-consensus, or Velvet (Zerbino and Birney, 2008), SPAdes (Bankevich et al., 2012), and the assembler of the proprietary platform CLC Genomics Workbench (Qiagen Aarhus A/S, Aarhus, Denmark), based on de Bruijn graphs.

More recently, many other tools have been developed to assemble long-read sequences, most of them based on overlap-layout-consensus, in order to take advantage of the length of the reads (Sohn and Nam, 2018), such as HGAP (Hierarchical Genome Assembly Process) (Chin et al., 2013), Miniasm (Li, 2016), Canu (Koren et al., 2017) or Flye (Kolmogorov et al., 2019), which was recently shown to be one of the most reliable long-read assemblers, especially for strains carrying plasmids (Wick and Holt, 2019).

Additionally, several tools have been developed for performing hybrid *de novo* assemblies of short and long reads, such as HybridSPAdes (Antipov et al., 2016) or the pipeline Unicycler (Wick et al., 2017), designed to take advantage of both kinds of reads. An alternative to *de novo* assemblies are reference-guided assemblies, which rely on an already assembled reference sequence to assemble the genome sequences of closely-related strains. Several pieces of software can perform reference-guided assemblies, such as Newbler (from 454 Life Sciences Corp.) or CLC Genomics Workbench.

1.2.5.3 Improvement of genome sequence assembly

Genome assemblies often result in draft assemblies that are formed by multiple contigs or scaffolds. This is the most common outcome when using only short reads (Pop, 2009). The development of third-generation sequencing technologies, with capacity to provide long reads, has facilitated the assembly process, but still, they are not always capable of completing the sequence of the genome. Additionally, assemblies may contain errors such as miss-assemblies

(Salzberg and Yorke, 2005) or errors at the sequence level that can seriously alter protein prediction (Watson and Warr, 2019).

Hence, numerous strategies and tools have been developed to improve different aspects of genome assemblies. Multiple tools have been developed to provide less fragmented assemblies, for instance by scaffolding contigs using short, paired reads or long-read data, such as SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) (Boetzer et al., 2011; Boetzer and Pirovano, 2014). Other tools have been developed to combine assemblies and thus make use of the different advantages of different assemblers, such as Metassembler (Wences and Schatz, 2015). Other tools are available for closing gaps present in scaffolds, correcting sequence errors, and/or correcting misassemblies, such as for instance PAGIT (Post-Assembly Genome-Improvement Toolkit) (Swain et al., 2012), IMAGE (Iterative Mapping and Assembly for Gap Elimination) (Tsai et al., 2010), GapFiller (Nadalin et al., 2012), Racon (Vaser et al., 2017) or Pilon (Walker et al., 2014). These tools are sometimes useful to obtain completely closed and gap-free genome assemblies, i.e., finished assemblies, and they can also be used to polish assemblies with low sequence accuracy, such as those obtained when assembling only Oxford Nanopore reads.

1.2.5.4 Evaluation of genome sequence assembly

Once the reads have been assembled into contigs or scaffolds, assemblies can be evaluated by using different strategies and tools. An important such tool is QUAST (Quality Assessment Tool for Genome Assemblies) (Gurevich et al., 2013), which provides general assembly statistics and plots, and optionally can use a reference genome to detect misassemblies and mismatches. Other tools have also been developed to estimate different kinds of errors in the assemblies, such as feature response curves (FRC) (Narzisi and Mishra, 2011), which can be applied, using FRCbam to detect various kinds of abnormalities and possible errors (Vezzi et al., 2012) or REAPR (Recognition of Errors in Assemblies using Paired Reads) (Hunt et al., 2013), which reports errors and provides an updated assembly, broken where errors are detected.

Such tools are especially important when evaluating the performance of new assembly strategies or when comparing different strategies. Furthermore, several competitions and studies have taken place in order to benchmark and compare different assemblers (Earl et al., 2011; Zhang et al., 2011; Salzberg et al., 2012; Bradnam et al., 2013; Magoc et al., 2013; Wick and Holt, 2019; Chen et al., 2020b).

In addition, other tools have been developed to detect contamination in genome assemblies. A well-known tool is VecScreen, which is designed to detect vector contamination and indeed is implemented in GenBank, where it screens the submitted sequences (Schäffer et al., 2017). Another widely used tool is CheckM, which infers the level completeness and contamination using lineage-specific gene markers (Parks et al., 2015).

1.2.5.5 Complete versus draft genome sequences

The assembly of DNA sequence reads of a whole-genome sequencing experiment, can result in draft or complete genome sequences. Draft genome sequences are fragmented into contigs or scaffolds, while complete genome sequences are assembled into one circular contig per replicon. The outcome of an assembly depends on multiple variables, such as the complexity of the genome, the sequencing technologies being used, the sample preparation process, the quality of the sequencing and the assembly strategy used.

The ideal outcome of any whole-genome sequencing project would be to obtain complete and very high-quality genome sequences, as that would be as close to reality as possible. However, determining complete genome sequences is typically more expensive and requires much more effort than determining draft genome sequences, although the recent development of third-generation sequencing technologies has hugely facilitated this task (Koren et al., 2013; Sereika et al., 2022).

Additionally, draft genome sequences generally provide the absolute majority of the genes of a genome and, therefore, they are useful enough for a wide range of applications, such as strain typing, outbreak investigations or comparative genomic analyses. Indeed, until 2015, most microbial genome large-scale studies had produced only short-read sequence data (Koren and Phillippy, 2015), and as of February 2023, only 11.3% of the bacterial genome assemblies in the RefSeq were complete genome sequences (O'Leary et al., 2016).

Despite their usefulness for a wide range of applications, draft genome sequences also present several limitations, especially if they are of low quality (e.g., low coverage) or highly fragmented. They may present lower completeness levels, more misassemblies, and are more susceptible to contain foreign contaminant sequences (e.g., vector sequences, human genome sequences or sequences from other microorganisms) than complete genome sequences (Fraser et al., 2002). Moreover, they can also cause mapping artifacts, errors in gene annotation and very often inability to reconstruct repeats (Ricker et al., 2012). Therefore, despite their usefulness, draft genome sequences may be a source of errors for downstream applications, such as mislabeling in meta-omic studies (e.g., metagenomics, metatranscriptomics, metaproteomics), or wrong interpretations about horizontal gene transfer.

Additionally, because of the lack of contiguity, they often do not allow distinguishing structural variations or contextualizing certain genomic elements which have important roles of adaptation to the environment (Ricker et al., 2012), such as for instance, if certain antibiotic resistance genes are located on plasmids and other mobile genetic elements (Grevskott et al., 2020; Jaén-Luchoro et al., 2020). Therefore, the value of draft genome sequences is clearly limited (Fraser et al., 2002).

Meanwhile, complete high-quality genome sequences offer an accurate and more realistic image of the genome of a particular strain in a moment of the evolutionary history. They are the best possible foundation for accurate genome annotation (Omasits et al., 2017) and, therefore, the most reliable basis for any downstream application and future research relying on genome sequences, such as functional genomic studies, genome organization analyses, accurate comparative genomic studies or microbial forensics (Fraser et al., 2002).

Thus, there is no doubt about the higher value of complete genome sequences, although unfortunately, obtaining them is usually more expensive and often requires manual curation. The good news is that the recent development of third-generation sequencing technologies and pipelines for automatic assembly and finishing of genome sequences has enormously facilitated this task and thus, determining complete genome sequences is becoming simpler and cheaper than ever (Wick et al., 2017).

1.2.5.6 Annotation of genome sequences

Once assembled, the genome sequences can be easily annotated, using several software tools, such as Prokka (Seemann, 2014), which is very useful for homogenizing the annotation of a set of multiple genome sequences; DFAST (DDBJ Fast Annotation and Submission Tool), which is a rapid annotation pipeline from the DNA Data Bank of Japan (DDBJ) (Tanizawa et al., 2017; Ogasawara et al., 2020); or the NCBI (National Center for Biotechnology Information)

Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016), which is really convenient for getting an "official" (and afterwards publicly available) annotation when submitting a genome sequence to GenBank (Sayers et al., 2022).

1.2.5.7 Integrative platforms

Most of the pieces of software described above are open source and free to use, having been created by developers from around the world. This means that anyone can use them and access and edit their codes, thus, enabling a "democratization" of bioinformatics and empowering software development collaborations. However, in most cases, this implies that the support is limited and that users need to have, at least, basic bioinformatic knowledge, know how to deal with software installations, how to work on command line, and sometimes basic knowledge in programming languages such as Perl or Python, which has definitely been a cumbersome and discouraging obstacle for many potential users.

To overcome this barrier and to facilitate access to genomic bioinformatics for users without prior experience, several user-friendly platforms integrating many of the above-mentioned processes have been developed in recent years. One of the most widely-used ones is the CLC Genomics Workbench (Qiagen Aarhus A/S, Aarhus, Denmark), a comprehensive and user-friendly NGS data analysis proprietary platform, which despite being relatively expensive, has facilitated the implementation of NGS bioinformatic workflows in clinical routine settings and has enabled access to genomic analyses to users with little or no bioinformatic background.

Worth mentioning is also Galaxy, an open source, free, and community centered on-line platform which offers high computational capacity, high reproducibility and aims to facilitate access to bioinformatics and data analysis to users without much informatics expertise (Afgan et al., 2018), and KBase (The Department of Energy Systems Biology Knowledgebase), which is also an on-line platform that offers user-friendly and highly-reproducible bioinformatics (Arkin et al., 2018).

1.2.6 Rise in the number of sequencing projects: a myriad of possibilities

During the first decades of DNA sequencing, multiple developments and improvements gradually increased the data sequencing throughput. However, obtaining large amounts of sequence reads remained challenging and expensive during those years. This situation persisted until the development and commercialization of the first NGS technologies, which opened the doors to high-throughput DNA sequencing.

A direct consequence of the introduction of NGS technologies into the market, was an abrupt decline in the costs of DNA sequencing. For years, the cost of DNA sequencing followed a pattern similar to the Moore's law (which describes a long-term trend in computer science in which the number of transistors in a dense integrated circuit doubles every two years). However, the commercialization of NGS technologies caused an abrupt decline in the cost that broke the trend (Figure 2). Indeed, in January 2007, the costs of a raw megabase of DNA sequence and of sequencing a human genome were estimated to be 523 USD and 9,400,000 USD, respectively, while in April 2008, they were estimated to be 15 USD and 1,500,000 USD. The latest estimations of these costs (August 2021) are 0.006 USD and 562 USD, respectively (source: National Human Genome Research Institute, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost; accessed 12th February 2023).

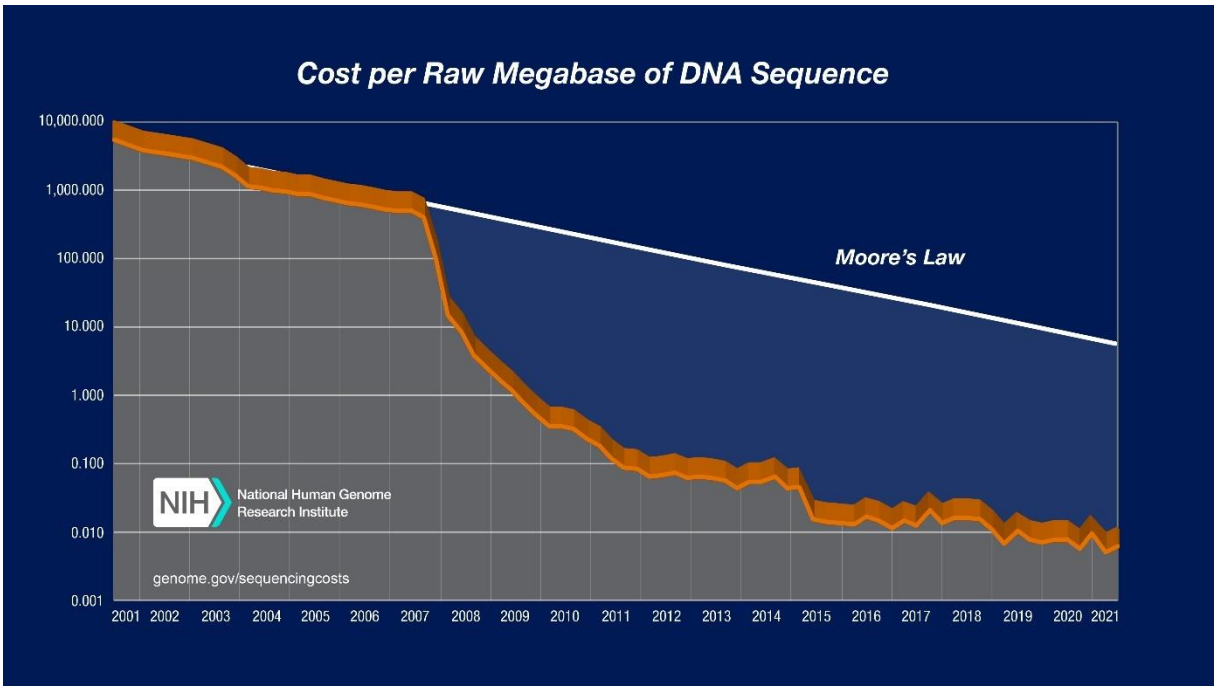

Figure 2. Evolution of the cost (in US dollars) of determining a raw megabase of DNA sequence (source: National Human Genome Research Institute, https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost; accessed 22th February 2023).

This steady decrease in the costs of sequencing allowed many laboratories to access high-throughput DNA sequencing technologies for basic and applied research. This development opened the door to an endless number of possibilities and triggered a "genomics revolution". As a consequence, the number of sequencing projects has grown almost exponentially during the last fifteen years, with the domain *Bacteria* being the group with the highest number of sequencing projects (Figure 3), which is having now and will continue having a longstanding impact in bacteriology (Loman and Pallen, 2015). For instance, by the year of starting this thesis project (i.e., 2015), the number of bacterial sequencing projects in GOLD (https://gold.jgi.doe.gov/) was 45,915 (Mukherjee et al., 2021). As of February 2023, the number was already 217,149 (i.e., +473%). Moreover, in February 2023, GenBank (Sayers et al., 2022) and RefSeq (O'Leary et al., 2016), two major sequence databases that are part of the International Nucleotide Sequence Database Collaboration (INSDC) (Arita et al., 2020), contained 1,467,686 and 275,284 bacterial and 13,744 and 1,405 archaeal genome assemblies, respectively.

The massive increase in the number of sequencing projects that followed the development of high-throughput DNA sequencing technologies, reflects the enormous interest of the scientific community in genome sequencing as a powerful approach for understanding basic biological processes and as a tool that opens the door to a myriad of otherwise unimaginable possibilities and applications.

A good example is the increasingly important role of whole-genome sequencing in clinical microbiology and public health, where, for instance, NGS and whole-genome sequencing have been successfully applied for pathogen surveillance and inference of virulence and antibiotic resistance (Aanensen et al., 2016), as well as for molecular outbreak investigations, where it offers higher reproducibility and resolution than the traditional "gold-standard" epidemiological method of pulsed-field gel electrophoresis for comparing closely-related isolates (Salipante et al., 2015). Despite some important bottlenecks, this is gradually leading

towards the implementation of bacterial whole-genome sequencing in routine clinical microbiology (Balloux et al., 2018; Rossen et al., 2018).

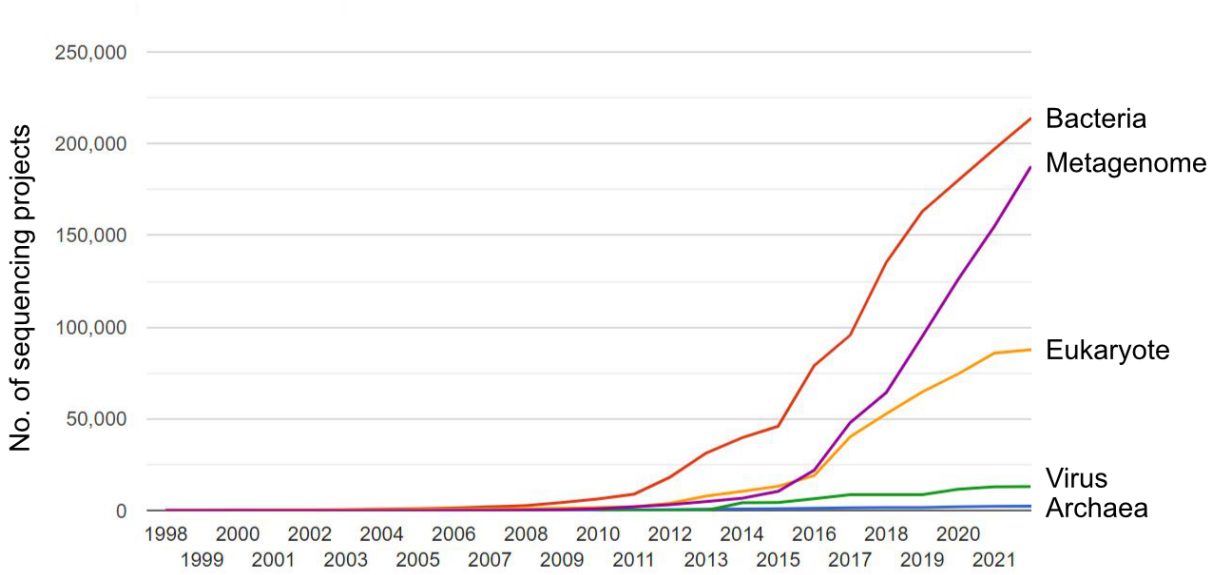

Figure 3. Number of sequencing projects over time in the Genomes OnLine Database (GOLD), classified as archaea, bacteria, eukaryote, virus and metagenome (source: GOLD, https://gold.jgi.doe.gov/statistics; accessed 12th February 2023).

1.2.7 Quality crisis and misidentified genome sequences

Before the development of high-throughput DNA sequencing technologies, an important bottleneck in any whole-genome sequencing project was to obtain enough sequence reads to cover the entire genome. At that time, relatively few genome sequences were determined each year, and the sequencing projects required months or even years of effort from dozens of scientists. Nonetheless, it was possible to carry out most sequencing projects to yield manually curated high-quality complete genome sequences.

However, the trend of finishing genomes started to shift at the beginning of the 2000s, when quantity began to be prioritized over quality (Fraser et al., 2002). This new trend was enormously reinforced by the introduction of NGS technologies, which enabled the sequencing of hundreds of genomes and the production of millions of sequences in just a few days. This was certainly advantageous, because it enabled sequence determination of more inter- and intra-species diversity, which opened the door to new studies and applications.

Indeed, the development of NGS eliminated the sequencing bottleneck, but immediately created a tremendous bioinformatics bottleneck, including data processing and interpretation. Additionally, the short lengths of the NGS reads difficulted even more the completion of the sequenced genomes. For instance, as of 4th October 2020, 243,598 of the 263,097 bacterial genome sequences that were publicly available in GenBank were in draft status, i.e., only 7.4% were complete, a percentage that decreased to 2.5% as of 12th February 2023 (only 36,813 of the 1,467,686 bacterial genome sequences available in GenBank were complete), partially due to the presence of numerous metagenome-assembled genomes (MAGs).

A consequence of the “trivialization” of whole-genome sequencing was that the average level of manual curation decreased drastically, evidently because it was not feasible anymore to

spend months completing and checking every single genome sequence that was being produced. Thus, despite thousands of genome sequences being determined every year, most of them were left in draft status of variable quality levels.

Additionally, with the decrease in sequencing costs, most microbiology laboratories gained direct or indirect access to whole-genome sequencing. This situation had many positive effects, although it also emphasized an increasing problem, which is the low quality of some of the sequences that have been deposited in the public databases, even though this problem already existed before the bloom of the NGS technologies, most probably because of the lack of careful evaluation by the authors, but also because of the lack of quality controls in primary sequence databases (Ashelford et al., 2005; Salzberg and Yorke, 2005).

For instance, despite the implementation of tools such as VecScreen in public sequence repositories (Schäffer et al., 2017), vector contamination of microbial genome sequences was shown to be abundant some years ago (Mukherjee et al., 2015). Also, more than 2,000 publicly available complete and draft prokaryotic genome sequences were shown to contain human contamination, leading to more than 3,000 "false" bacterial proteins of human origin, which can have serious consequences for downstream analyses and data interpretation (Breitwieser et al., 2019). A more recent study reported contamination in numerous draft and complete genome sequences in GenBank and RefSeq databases (Steinegger and Salzberg, 2020). On top of that, during the last years, several studies have reported the presence of high rates of misidentified genome sequences in public databases (Beaz-Hidalgo et al., 2015; Gomila et al., 2015; Jensen et al., 2016), a situation that we have personally encountered constantly during the last years and that had already been reported for public 16S rRNA gene sequences (Yarza et al., 2008).

Unfortunately, all of these problems can have major consequences in downstream applications and analyses, such as metagenomics for infectious disease (Breitwieser et al., 2019), comparative genomics (Smits, 2019) or shotgun proteotyping (Karlsson et al., 2018). Hence, the quality of public sequence data needs to be improved, in order to increase its usefulness and to prevent drawing incorrect biological conclusions (Bengtsson-Palme et al., 2016).

Evidently, the primary responsible for these situations are the authors who are depositing the sequence data, who are accountable for the data that they produce, archive and release. In some cases, these errors might be due to carelessness and negligence by the authors, but, in some cases, they might just be *bona fide* mistakes, often derived from the unawareness of the authors about these problems and their consequences.

Therefore, public databases should also assume responsibility and establish measures and mechanisms to deal with these errors and quality problems. For instance, as mentioned above, GenBank implemented VecScreen (Schäffer et al., 2017) to detect vector contamination, and ANI calculated against type or proxitype reference genome sequences, to verify the taxonomic assignments of submitted genome sequence deposits (Federhen, 2015; Federhen et al., 2016; Ciufo et al., 2018; Kannan et al., 2023). Additionally, RefSeq, provides a basis for curated (with fewer, but not free of errors) and regularly reannotated sequences, based on the original deposits performed over time to the INSDC databases (O'Leary et al., 2016).

However, despite the implementation of controls and measures to combat the quality issues, public sequence databases still contain numerous errors, and error-free databases probably will never be assured. Therefore, it is paramount to give visibility to the quality issues in order to increase the awareness among all stakeholders and to encourage the performance of quality

controls at all levels by all parts, from the authors producing the primary data to the end users of this data.

1.2.8 High-quality reference genome sequences

In any case, despite the presence of errors and low-quality sequences, the development of high-throughput DNA sequencing technologies and bioinformatic tools for data analysis has facilitated the establishment of high-quality genome sequences, which serve as important references for a wide range of applications.

In this thesis, the term "reference genome sequences", refers to genome sequences that serve as a reference at any level, from sequences of "prominent" type strains (i.e., the nomenclatural type of a species or subspecies) (Parker et al., 2019) or well-established and widely-used laboratory model strains (e.g., *P. aeruginosa* PAO1 or *S. pneumoniae* TIGR4 and R6), to sequences of strains having particular phenotypic features that might be used for validation purposes (e.g., antimicrobial resistance factors or toxins) and sequences of local strains, such as those isolated during outbreak investigations; e.g., *E. coli* CCUG 73778, an early isolate from an outbreak of extended-spectrum β-lactamase (ESBL)-producing *E. coli* at a neonatal surgical ward, at the Sahlgrenska University Hospital, Sweden (Jaén-Luchoro et al., 2020).

The genome sequence of each bacterial strain essentially provides a solid grounding and a rich source of information about its biology and about the species to which it belongs to. For this reason, obtaining and establishing good quality reference sequences is paramount for downstream applications that rely on them, such as, for instance metagenomic analyses of clinical samples (Li et al., 2021a).

Thus, numerous projects and scientific consortia have been established throughout the years, aiming to deliver high-quality reference sequences of selected strains. A prominent example was the Human Microbiome Project (Turnbaugh et al., 2007), which, in its first phase was aiming to determine the genome sequences of numerous selected microbial taxa associated with the human body (The Human Microbiome Jumpstart Reference Strains Consortium, 2010). Another well-known initiative that includes multiple projects was the Genomic Encyclopedia of *Bacteria* and *Archaea* (GEBA; https://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/), which was established with the aim of sequencing diverse prokaryotic genomes across diverse branches of the Tree of Life (Wu et al., 2009), including hundreds of type strains (Mukherjee et al., 2017). Other projects have also been devoted to sequence genomes of type strains, such as the NCTC 3000 project (https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx), which aimed to determine the complete genome sequences of 3,000 type and reference bacterial strains, or the Global Catalogue of Microorganisms (GCM) 2.0 sequencing project, which aims to sequence the genomes of more than 10,000 prokaryotic and fungal type strains (Wu et al., 2018).

Worth mentioning is also FDA-ARGOS (Food and Drug Administration dAtabase for Regulatory-Grade micrObial Sequences), a public database of high-quality manually curated reference complete genome sequences, which is collecting and sequencing thousands of microbes such as pathogens, related species or biothreat microorganisms. The database aims to serve as a genomic reference for diagnostic and regulatory science and, additionally, it aims to include high-quality sequences of at least five strains per species (Sichtig et al., 2019).

All of these and many other large- and small-scale scientific efforts have, altogether, contributed numerous high-quality genome sequences for important reference strains, type strains and additional strains of numerous species that are currently available in public databases (Ogasawara et al., 2020; Harrison et al., 2021; Mukherjee et al., 2021; Sayers et al., 2022). Thus, public databases currently contain hundreds of thousands of complete and draft genome sequences of varying quality levels.

However, despite all these efforts, to date (December 2022), numerous bacterial species with validly published names do not have any public genome sequence representing them yet (e.g., *Streptococcus orisuis*). This is partly due to the bias towards species that have attracted more scientific attention for instance because of their clinical and epidemiological significance, such as *E. coli* or *S. pneumoniae*, which currently comprise 33,463 and 9,074 genome sequences in GenBank (Sayers et al., 2022), respectively.

Thus, even a situation that might have seemed simple, such as the creation of a comprehensive genome sequence catalogue of the type strains of each bacterial species, similar to that of 16S rRNA gene sequences (Yarza et al., 2013), has not been reached yet. In fact, the number of validly published species names is currently around 20,000 (data obtained from the List of Prokaryotic names with Standing in Nomenclature, LPSN) (Parte, 2014), and many of them do not have determined and publicly available genome sequences for the type strain yet. For instance, in May 2019, we determined that 812 of the 3,725 bacterial type strains deposited in the Culture Collection University of Gothenburg (CCUG, Gothenburg, Sweden), i.e., a 21.7%, did not have a publicly available genome sequence.

Type strains serve as fundamental taxonomic reference points; thus, they are critical for the correct identification of other genome sequences (Tindall, 2008; Tindall and Garrity, 2008; Tindall et al., 2010) and hence for any other downstream application relying on public genome sequence databases. Moreover, type strains are typically well-characterized and have rich associated metadata, which increases their value (Kyrpides et al., 2014). Additionally, having diverse genome sequences across the Tree of Life has numerous potential benefits, such as better identification of orthologous genes and proteins across different taxa, improved taxonomic binning of metagenomic data, discovery of new genes and protein families and prediction of novel biological functions, better understanding of the mechanisms driving the evolution of microbes and of the evolutionary history, and improved correlations between genotype and phenotype (Wu et al., 2009).

Thus, further efforts are definitely needed to, 1) cover the taxonomic and genomic gaps that still exist among bacterial species, and 2) increase the number of sequenced strains per species. Furthermore, since type strain genome sequences serve as references for classifying other genome sequences and for numerous downstream applications, it is paramount that they are correct, as any error present in them will be perpetuated to any application making use of that sequence.

1.3 Comparative genomics

Having a single genome sequence provides large amounts of information, and indeed thousands of scripts and pieces of software have been developed for analyzing and mining data from individual genome sequences. However, the evolution and the composition of prokaryotic genomes is driven, not only by vertical transfer between generations, but also by mechanisms of horizontal gene transfer (HGT) and other events, such as gene loss or gene duplication, which

implies that intra-species genetic variation can be huge (Ochman et al., 2000; Koonin et al., 2001).

These observations led to the concept of a "pan-genome", i.e., the whole set of genes of a bacterial species, formed by the "core genome" (the set of genes that are present in all strains of a designated subset) and the "accessory genome" (the set of genes that are not present in all strains of a designated subset, also known as "adaptive genome", which encodes features that allow those strains having them to be adapted to a specific niche, or to perform a particular function). The definition of "pan-genome" was presented by Tettelin *et al.*, in a pioneering study in which they analyzed eight genome sequences of *Streptococcus agalactiae* and observed that after the addition of an eighth genome, new genes continued to be found, and hypothesized that new genes would emerge even after the addition of hundreds of genome sequences (Tettelin et al., 2005).

A pan-genome can be classified as "open" or "closed", depending on the probability of finding new genes after the addition of new genomes. In an "open" pan-genome, new genes are expected to be found when adding new genomes, whereas in a "closed" pan-genome, few new genes are expected to be found (Mira et al., 2010). The size of the genetic repertoire of a bacterial species is a good indicator of the degree of versatility and adaptability of a species, and typically reflects the lifestyle of the species, with sympatric species (those living in large communities) typically having larger pan-genomes than allopatric species (those living more isolated, in narrow communities) (Rouli et al., 2015). For instance, *S. agalactiae*, which can be found in large communities, as a commensal of numerous animals and sometimes causing disease, has an "open" pan-genome. Alternatively, *Bacillus anthracis*, which typically remains sporulated in the soil until it can multiply in a host, has a "closed" pan-genome that can be defined after compiling just four genome sequences (Medini et al., 2005; Tettelin et al., 2005).

Thus, the study of pan-genomes allows the determination of which genes are specific for a species or a subgroup of strains of a species. The study of such genes and their potential functions can be related to the ecological implications and the lifestyle of the particular species or group of strains and can provide clues about the selective forces that the environment has exerted on the bacteria and how they have adapted to their environmental niches.

These kinds of studies are more feasible now than ever before, since the development of high-throughput DNA sequencing technologies has facilitated the determination of genome sequences from multiple strains of each species, hence enabling the possibility of performing intraspecific comparative genomic studies. In response, numerous pieces of software have been developed for analyzing the pan-genome of a group of strains, such as, for instance, GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013), Roary (Page et al., 2015) or Panaroo (Tonkin-Hill et al., 2020), a recently developed and promising pipeline that aims to deal with errors derived from annotations, draft assemblies and contamination, in order to infer more accurate pan-genome sizes.

Consequently, numerous pan-genome studies of prokaryotes have been carried out, showing that different species may have different pan-genome structures, with some having "closed" pan-genomes of limited size, such as *B. anthracis*, and some having extremely "open" and versatile pan-genomes, such *P. aeruginosa*, for which the analysis of 1,311 genomes revealed a core genome of 665 genes, which represented only 1% of the pan-genome (Freschi et al., 2019). Altogether, comparative genomics has definitely revolutionized our understanding of the evolution, composition, structure and biology of prokaryotic species, and these approaches

serve as basis for numerous downstream applications, such as identification of potential diagnostic and vaccine targets or epidemiological and evolutionary studies (Costa et al., 2020).

1.4 Bacterial genomics for biomarker discovery

"Biomarker" (or "biological marker") is a term with a broad meaning that can be defined as a substance, characteristic, molecule, gene, etc. that indicates a biological state or condition. Hence, biomarkers are widely-used in medicine and clinical research to determine conditions, for instance, in patients (Strimbu and Tavel, 2010).

In clinical microbiology, identifying the causal agent of infection is the first step towards taking effective countermeasures. For bacterial infections, diagnoses traditionally rely on the isolation and characterization of bacteria causing disease, to detect differential features that alone or in combination allow their identifications. Additionally, numerous methods have been developed along the years for detection of pathogens directly in clinical samples, without prior cultivation, such as specific PCR assays, immunoassays or, even, direct MALDI-TOF MS (matrix-assisted laser desorption ionization time-of-flight mass spectrometry) (Welker and Moore, 2011; Briggs et al., 2021) (Briggs et al., 2021). Thus, identification methods are based on the detection of specific phenotypic features and/or molecular markers which typically allow the identification of bacterial isolates to the species level.

However, intra-species strain variability, which can be extremely large in sympatric species with an open pan-genome, such as *S. pneumoniae* or *P. aeruginosa*, and close relationships with other bacteria (which in many cases have very different pathogenic potential), often challenge the reliability of the used markers and compromise their sensitivity and specificity. This sometimes hinders identification and leads to errors, which potentially can affect patient treatments, increase the use of antibiotics and thus promote the development of antibiotic resistance. Thus, it is important that markers for identification are tested and validated on large numbers of strains and that they are continuously reevaluated (Strimbu and Tavel, 2010).

Whole-genome sequencing allows for the comparison of the entire genetic repertoire of a set of strains. Thus, by looking at the whole picture, the comparison of multiple genome sequences clearly can reveal genes and potential proteins and functions that are specific for a particular group of bacteria. However, to reliably assess the sensitivity and specificity, it is critical to include numerous genomes in the analyses.

Fortunately, thanks to the development of high-throughput DNA sequencing technologies and to the contributions from thousands of scientists, there are numerous publicly available genome sequences for the most clinically-relevant groups of bacteria. These genome sequences are a "gold mine" for finding genes or proteins that are exclusive to a species or a subset of genomes. These genes can potentially be used for differentiating that species and to detect its presence in a sample. But obviously, it is important to consider the quality and the reliability of these genome sequences and, if they are going to be used for identification purposes, they must be taxonomically verified to avoid errors and miss potential specific markers.

Since whole-genome-based screenings include all the genes and sequences of the analyzed strains, they evidently maximize the chances of finding the best possible biomarkers, based on DNA sequence. Thus, whole-genome sequencing has definitely become a basic and essential tool for discovery and validation of bacterial biomarkers for identification and for detection of particular features such as antimicrobial resistance factors.

1.5 Impact of DNA sequencing in prokaryotic taxonomy

Prokaryotic taxonomy, historically, was based on the isolation and phenotypic characterization of strains (morphology, metabolic features, etc.) (Rosselló-Mora and Amann, 2001). However, progress in molecular biology techniques and increasing knowledge about the nature of DNA, led early in the 1960s to the idea that bacteria could probably be better classified by comparing their genomes (Schildkraut et al., 1961). Initial comparisons were based on G+C content (mol%) differences; that was a good start, but a method with greater resolution was needed. Subsequently, various DNA-DNA hybridization (DDH) techniques were developed, which enabled more precise determinations of taxonomic relationships (Schildkraut et al., 1961). DDH experiments, usually using type strains of species, as reference strains, were used for many decades, as the "gold standard", for determining whether strains belong to the same species. Strains of the same species generally should present a DNA-DNA hybridization value of, at least, 70% genomic similarity and a difference in melting temperature (ΔT_m) $\leq$ 5°C (Wayne et al., 1987). Although DDH methods were an essential and long-standing tool for prokaryotic taxonomy studies, a limitation was that they could only define strain-to-strain species relationships and could not determine degrees of relatedness at higher taxonomical levels. In addition, DDH methods were laborious, exhibited high error rates, often required to collaborate with specialized laboratories, and the data was not cumulative; hence, they became a bottleneck for many taxonomic studies (Stackebrandt, 2006).

Soon after, the analysis of ribosomal ribonucleic acids (rRNA), especially of the 16S rRNA gene, enabled the inference of close (i.e., species-level), as well as distant phylogenetic relationships and became a major breakthrough in the understanding of the phylogeny of prokaryotes (Woese and Fox, 1977). Studies based on rRNA comparisons led to the proposal of "domain", a taxonomic rank above the kingdom level, which would be constituted by *Archaea*, *Bacteria* and *Eukarya* (Woese et al., 1990). Thus, analyses of the rRNAs or of the corresponding genes, allowed the determination of major phylogenetic relationships, although it soon became clear that it sometimes possessed relatively low resolution and could not elucidate species-level identities (Fox et al., 1992). For instance, 16S rRNA gene sequence analyses do not have enough resolution to discriminate many species within genera such us *Pseudomonas* or *Streptococcus* (Janda and Abbott, 2007) and, sometimes, even higher taxonomic levels such as genera within the order *Enterobacterales* (Adeolu et al., 2016).

Due to their different strengths and weaknesses, DDH experiments and 16S rRNA gene sequence analyses complemented each other very well and became two DNA-based "gold standards" in prokaryotic taxonomy. On one side, 16S rRNA gene sequence analyses were regarded as being extremely useful for determining high taxonomic rank phylogenetic relationships and establishing the framework of prokaryotic phylogeny, while, on the other side, DNA-DNA hybridization experiments were considered to have the highest resolution for circumscribing species and were deemed particularly useful for 16S rRNA gene complexes, i.e., groups of closely-related species with high 16S rRNA gene sequence identity. Indeed, to reduce the DDH bottleneck, a 16S rRNA gene sequence threshold of 97% (later updated to 98.7%) was proposed, below which, DDH experiments were not considered necessary because the strains could be recognized as different species (Stackebrandt and Goebel, 1994; Stackebrandt, 2006). Currently, an important advantage in using 16S rRNA gene sequences for establishing taxonomic frameworks is that nearly all type strains of species with validly published names also have publicly available sequences (Yarza et al., 2013). Indeed, a highly-curated database, alignment, and phylogenic tree of 16S rRNA gene sequences of type strains

of almost all prokaryotic species is maintained and publicly available at the "All-Species Living Tree" project (Yarza et al., 2008; Ludwig et al., 2021).

DDH experiments and 16S rRNA gene sequence analyses were extraordinarily useful and became essential in taxonomic studies. However, because of their limitations, additional methods were employed, in combination with DDH and 16S rRNA gene sequencing, to establish a, so-called, "polyphasic" approach for taxonomy (Vandamme et al., 1996), using multiple methods to try to fill "gaps" in the information on the relationships between microorganisms. But, novel and better approaches were needed, and soon it became evident that whole-genome sequencing would be the ultimate basis for the most accurate determination of microbial phylogenetic relationships. Indeed, already in the 1980's, an *ad hoc* committee of the International Committee for Systematic Bacteriology recognized that, "*the complete deoxyribonucleic acid (DNA) sequence would be the reference standard to determine phylogeny and that phylogeny should determine taxonomy*" (Wayne et al., 1987). However, by that time, DNA sequencing was still at its early stages and far from being high-throughput, which necessitated sequence comparisons to be limited to one or only a few genes.

Later, a groundbreaking sequence-based method came in to provide higher resolution than 16S rRNA gene-based analyses for inferring phylogeny but without need for whole-genome sequences: multilocus sequence analysis (MLSA), a modification of the multilocus sequence typing (MLST) approach proposed in 1998 (Maiden et al., 1998). MLSA uses several housekeeping genes for inferring phylogenetic relationships and, therefore, the sequences can be determined performing a relatively manageable set of Sanger sequencing reactions. The method has had a major and long-standing impact in prokaryotic taxonomy and has facilitated the inference of phylogenetic relationships that 16S rRNA gene sequence analyses could not resolve (Glaeser and Kämpfer, 2015). However, there is no standardized one set of target genes for performing MLSA studies, and many variables may influence the outcome, such as how many and which genes are selected, and which algorithms are used for calculating distances and constructing phylogenetic trees. Additionally, since only relatively few genes are analyzed, HGT events may cause noise and affect MLSA studies. In any case, MLSA has been and continues to be extremely useful for inferring more accurate phylogenies, even though it can be problematic with some groups of very closely-related species (Jensen et al., 2021b).

New developments in DNA sequencing enabled the possibility of using entire genome sequences to study taxonomic relationships. For instance, they opened the possibility of performing *in silico* pairwise whole-genome comparisons that could replace the tedious and non-accumulative DDH experiments. This motivated the development of several, so-called, overall genome relatedness indices (OGRIs), i.e., methods to determine how related two genome sequences are (Chun and Rainey, 2014). The most widely-used of these methods is the average nucleotide identity (ANI), for which a threshold of 95-96% corresponds well to the DDH species threshold of 70% (Konstantinidis and Tiedje, 2005a; Goris et al., 2007). Indeed, ANI of 90,000 prokaryotic genome sequences confirmed the existence of a clear species boundary (Jain et al., 2018). Another widely-used method is the digital DNA-DNA hybridization (dDDH), which was developed to mimic the classical DDH experiments (Meier-Kolthoff et al., 2013). Worth mentioning is also the average amino acid identity (AAI), which, unlike ANI and dDDH, uses amino acid sequences derived from shared genes to determine genome similarities (Konstantinidis and Tiedje, 2005b). These methods have been implemented in several tools and webservers, such as JSpecies (Richter and Rosselló-Móra, 2009) and JSpeciesWS (Richter et al., 2016), for calculating ANI values, Genome-to-Genome Distance Calculator (GGDC) (Meier-Kolthoff et al., 2013), for determining dDDH values, or AAI calculator and EzAAI, for determining AAI values (Rodriguez-R and Konstantinidis, 2016;

Kim et al., 2021). Using these tools, genome-to-genome pairwise comparisons can be performed within minutes or seconds, enabling the performance of thousands of pairwise comparisons within a few hours.

However, similarly to DDH, an inconvenience of OGRIs is that, despite their simplicity, one typically has to know in advance with what to compare a particular strain. Thus, a recent recommendation was to continue with the widely-used two-steps approach, i.e., first use the complete 16S rRNA gene sequence to determine the "approximate" phylogenetic relationships, and, second, calculate an OGRI between the strain in question and the type strains of species with 16S rRNA gene sequence identity values ≥98.7% (Chun et al., 2018). To facilitate this task, several tools have been developed in recent years to screen all type strains automatically with the available genome sequences, by finding the most closely-related phylogenetic neighbors and determining OGRIs with those. Some of the most relevant ones are Type (Strain) Genome Server (TYGS) (Meier-Kolthoff and Göker, 2019), Microbial Genomes Atlas (MiGA) (Rodriguez et al., 2018), TrueBac ID (Ha et al., 2019), Global Catalogue of Type Strain (gcType) (Shi et al., 2020). These tools allow automatic species identification and detection of possible novel taxa and hence their potential and utility are huge.

To fully flourish in the "age of genomics", these tools require a complete catalogue of genome sequences of type strains. Many efforts are currently on-going in order to complete this "catalogue", such as the sequencing projects mentioned above. Additionally, from January 2018, the International Journal of Systematic and Evolutionary Microbiology (IJSEM) started to require genome sequence data for proposing novel taxa; in fact, the GCM 2.0 sequencing project offers free type strain sequencing services for authors proposing novel taxa (Wu and Ma, 2019). Yet, despite these efforts, many type strains do not have publicly available genome sequences yet, which means that the power of these tools for rapid species identification and detection of potential novel species is still relatively limited and that 16S rRNA gene sequences should usually be analyzed in parallel, using for instance EzBiocloud, to determine which species have a sequence identity equal to or higher than 98.7% (Yoon et al., 2017).

In general, OGRIs are extremely useful for circumscribing species and for determining the degree of relatedness between closely-related taxa, although a drawback (especially of those based on nucleotide sequence alignments), is that they cannot measure degrees of relatedness between distantly related taxa. However, similar to what 16S rRNA gene sequence analyses did with DDH, core genome-based phylogenomic treeing can help to accurately define phylogenetic relationships at higher taxonomic levels. In addition, the advantage of phylogenomic treeing over 16S rRNA gene is its high resolution power, which allows for resolving phylogenetic relationships at virtually any taxonomic level with extremely high accuracy, as already predicted by Wayne *et al*. in the 80's (Wayne et al., 1987). For this reason, while OGRIs are useful for species circumscriptions, phylogenomic treeing, including at least 30 genes, has been proposed for classifying genera or higher taxa and is expected to help in reorganizing the phylogenies of poorly classified taxa (Chun et al., 2018). In fact, a few years ago, Parks *et al*. presented the Genome Taxonomy Database (GTDB), which maps all publicly available bacterial genome sequences on a tree of life, based on 120 universal single-copy genes. This tree serves to maintain a normalized (not official) taxonomy, based on monophyletic lineages of similar phylogenetic depths, which has revealed hundreds of polyphyletic taxa and not yet described species (Parks et al., 2018; Parks et al., 2020).

One of the basic principles of the current system for taxonomy of prokaryotes is that at least the type strain of a proposed novel species has to be isolated in pure culture in order to be characterized and deposited in at least two publicly accessible culture collections in two

different countries (Parker et al., 2019). However, the vast majority of prokaryotic diversity remains uncultivated (Steen et al., 2019). To deal with this limitation, the category "*Candidatus*" was proposed and implemented almost three decades ago, as a provisional status for prokaryotes that did not fulfill the requirements of the International Code of Nomenclature of Prokaryotes (ICNP), such as those identified through 16S rRNA gene sequences obtained from environmental samples (Murray and Schleifer, 1994; Murray and Stackebrandt, 1995). However, despite its implementation, this provisional rank never became widely used (Konstantinidis et al., 2017). Additionally, the advent of high-throughput DNA sequencing confirmed the prediction made by Wayne *et al*. (Wayne et al., 1987), by showing the usefulness of genome sequences for circumscribing species and for determining phylogenetic relationships. Moreover, it facilitated the recovery of draft and, in some cases, even complete genome sequences from uncultivated prokaryotes, by single cell sequencing (i.e., single-amplified genomes, SAGs) (Marcy et al., 2007; Woyke et al., 2010; Woyke et al., 2017) or assembling shotgun metagenomic data (metagenome-assembled genomes, MAGs) (Tyson et al., 2004; Chivian et al., 2008; Chen et al., 2020a; Moss et al., 2020), which can then be included in genome-based taxonomic analyses. For instance, the GTDB incorporates numerous SAGs and MAGs that are publicly available in GenBank (Parks et al., 2018).

This situation led to the idea that it should also be possible to describe new taxa based on whole-genome sequence data, and to several proposals to amend the ICNP (Parker et al., 2019). These included a proposal to allow the use of genome sequences as nomenclatural type material to enable validly published nomenclature for difficult-to-culture (e.g., extremely fastidious and endosymbionts) and uncultivated prokaryotes (Whitman, 2016), and to grant priority to names of previously published *Candidatus* taxa (Whitman et al., 2019). These proposals were discussed and voted upon by the International Committee on Systematics of Prokaryotes, although the result of the vote was to reject them (Sutcliffe et al., 2020). The rejection of these proposals triggered the development of SeqCode (The Code of Nomenclature of Prokaryotes Described from Sequence Data), a parallel code of nomenclature which aims to provide a solution for naming difficult-to-culture and uncultivated prokaryotes (Hedlund et al., 2022).

Regardless of all the on-going discussions and debates, what is clear and all agree on is that the latest advancements in DNA sequencing and whole-genome sequencing have provided a new level of understanding of prokaryotic evolution and diversity, and a basis to better capture and understand the natural relationships between prokaryotes; these advances are revolutionizing bacterial taxonomy, and have helped to clarify the phylogenetic relationships of complex groups of bacteria with a high level of resolution and to solve problems that were not obvious with previously available approaches.

1.6 Taxonomic groups covered

This thesis is formed by a compendium of studies in which several taxonomic groups of bacteria have been involved. For this reason and to put the overall study into context, this subsection has been further subdivided into several sections that present the main bacterial taxa included in the analyses presented in this thesis.

The first subsection presents *Stutzerimonas balearica* (*Pseudomonas balearica*), a bacterium that has been associated most often with marine and polluted environments, with capabilities for degrading aromatic compounds and, which therefore, is a microorganism of interest for bioremediation applications. *S. balearica* is the focus in Papers I, II and VI. The second

subsection presents the genus *Streptococcus*, most of which members are considered commensal species of warm-blooded animals but that also includes major human pathogens such as *S. pneumoniae* and *S. pyogenes*. The genus *Streptococcus* is the focus in Papers III, IV and VII. Additionally, this subsection has been further subdivided to present *S. pyogenes* and *S. pneumoniae*, two major human pathogens that are the focus in Papers IV and VII, respectively. The third and last subsection presents the family *Enterobacteriaceae*, an ecologically very diverse and taxonomically complex group of bacteria, some of which members have major human pathogenic potential. The family *Enterobacteriaceae* is the focuss of Paper VIII.

1.6.1 *Stutzerimonas balearica* (*Pseudomonas balearica*)

The species *Stutzerimonas balearica* (formerly, *Pseudomonas balearica*) was proposed as a novel species in 1996, after the study and characterization of two denitrifying and salt-tolerant (8.5% NaCl) strains of the genomovar 6 of *Pseudomonas stutzeri* (currently, *Stutzerimonas stutzeri*) that had been isolated as degraders of 2-methylnaphthalene, from wastewater and polluted marine sediment samples (Rossello et al., 1991; Rosselló-Mora et al., 1994; Bennasar et al., 1996).

S. balearica is a member of the recently proposed genus *Stutzerimonas* (family *Pseudomonadaceae*, order *Pseudomonadales*, class *Gammaproteobacteria*), which includes species of the former *P. stutzeri* phylogenetic group of the *Pseudomonas aeruginosa* lineage of the genus *Pseudomonas*, which among Gram-negative bacteria is the genus with the highest number of species (Lalucat et al., 2020; Gomila et al., 2022; Lalucat et al., 2022). *S. balearica* is closely related to *S. stutzeri*, which is a highly diverse species (Palleroni et al., 1970; Rossello et al., 1991) that was first described in 1896 (Burri, 1895), as *Bacillus denitrificans II*, later as *P. stutzeri* (van Niel and Allen, 1952) and recently reclassified as *S. stutzeri* (Lalucat et al., 2022). *S. stutzeri* is a Gram-stain negative, non-fluorescent, denitrifying bacterium with a particularly high intra-species phenotypic and genotypic diversity. Strains of *S. stutzeri* are strictly respiratory, and can often be identified by their vigorous denitrification capacity, the dry and wrinkled morphology of their colonies (although they may become smooth after repeated cultivation steps), and their ability to use starch as sole carbon and energy source (Palleroni, 2015). However, there are exceptions in well-documented strains. Additionally, multiple reported strains of *S. stutzeri* are capable of fixing nitrogen and others have natural transformation capacity (Lalucat et al., 2006).

S. stutzeri has a wide ecological distribution and has been isolated from many different environments, such as soil, rhizosphere, water, industrial and clinical samples, which is reflected by its extremely broad metabolic diversity. Indeed, numerous strains of the species have been recovered from anthropogenic and contaminated environments, such as wastewaters or oil spills, often with capacity for degrading a wide range of compounds, including aerobic degradation of numerous aromatic compounds. These, together with other properties such as denitrification, nitrogen fixation capacity or its natural transformation properties, have attracted much attention to the species (Lalucat et al., 2006).

From the genomic perspective, *S. stutzeri* was formed by multiple described genomic groups (Scotta et al., 2013), termed “genomovars” (Rossello et al., 1991). According to their relatively low genomic relatedness (i.e., DNA-DNA hybridization <70%), they could potentially be

described as novel species of the genus *Stutzerimonas*. However, recognition as novel, validly published species has been in most cases hindered by the inability to phenotypically differentiate them. The first exception was *P. balearica* (currently, *S. balearica*), that initially formed the genomovar 6 of *P. stutzeri* until it was described and proposed as a novel species of the genus *Pseudomonas*, distinct from *P. stutzeri* (Bennasar et al., 1996). Until recently, *S. stutzeri* was formed by 21 described genomovars (Scotta et al., 2013). However, recent publications proposed that multiple of these genomovars represent distinct species within the genus *Stutzerimonas* and that the remaining ones should be treated as phylogenomic species of *Stutzerimonas* (Gomila et al., 2022; Lalucat et al., 2022). Thus, *S. stutzeri* is limited to the strains of the former genomovar 1.

Cells of *S. balearica* are Gram-stain negative, rod-shaped, with an approximate size of 0.3 – 0.5 x 1.5 – 3.0 µm, non-pigmented, strictly oxidative, and motile by a single polar flagellum. Strong denitrification capacity. The optimum growth temperature is 30°C and the colonies have wrinkled dry morphology. Strains of *S. balearica* share many phenotypic characteristics with its most closely-related species (i.e., *S. stutzeri*). For instance, both species are vigorous denitrifiers; amilase, malate and maltose positive; arginine and dihydrolase negative; able to degrade starch, but not gelatine. However, according to the original description, *S. balearica* can be differentiated from *S. stutzeri* by its ability to use xylose as sole carbon and energy source, inability to degrade mannitol, ethylene glycol, 4-aminobutirate and suberate. Additionally, strains of *S. balearica* can grow at 46°C and with an 8.5% of NaCl concentration (Bennasar et al., 1996).

Multiple strains of *S. balearica* have been reported along the years (Dutta, 2001; Mulet et al., 2008; Ahmadi et al., 2017; Salgar-Chaparro et al., 2020; Bravakos et al., 2021). These reports suggest that *S. balearica* is a species with diverse properties (e.g., biodegradation of aromatic compounds) which is principally found in marine and polluted environments. However, little is known about the diversity of *S. balearica*. This is due to the heterogeneity of *Stutzerimonas* and to the limited discriminatory power of the 16S rRNA gene, which has often resulted in misclassified strains and in strains not being identified at the species level (Lalucat et al., 2006; Gomila et al., 2022; Li et al., 2022; Uddin et al., 2022).

The genomes of *Stutzerimonas* have a size ranging approximately from 3.7 to 6.3 Mb, and a guanine plus cytosine (G+C) content (mol%) ranging from 59.6 to 66.9%. Meanwhile, the genomes of *S. balearica* have a size ranging from 4.3 to 4.6 Mb, and a G+C content (mol%) ranging from 64.4 to 65.0%.

1.6.2 The genus *Streptococcus*

Streptococcus is a diverse genus of the family *Streptococcaceae*, order *Lactobacillales*, class *Bacilli*, that accommodates Gram-stain positive, facultatively anaerobic, non-spore forming, catalase-negative, chemo-organotrophic bacteria with fermentative metabolism and with complex and variable requirements. Cells are coccoid, arranged in pairs or chains, and non-motile. The DNA G+C content (mol%) ranges from 33 to 46%, and the genome sizes range approximately from 1.7 to 2.2 Mb (Whiley and Hardie, 2015).

The genus *Streptococcus* currently (January 2021) encompasses 107 species with validly-published and correct name (Parte, 2014). Most members of the genus are considered commensal species of warm-blooded animals and birds, and are typically located in the oral

cavity, upper respiratory tract and gastrointestinal tract. However, several streptococci can cause localized and systemic infections under certain circumstances, and besides, some are outstanding human pathogens such as *S. pyogenes* or *S. pneumoniae* (Whiley and Hardie, 2015).

1.6.2.1 *Streptococcus pyogenes*

Streptococcus pyogenes (Rosenbach, 1884), a β-haemolytic bacterium that forms the Lancefield's Group A *Streptococcus* (Lancefield, 1933), is a major and strictly human pathogen that can cause a wide range of diseases, such as invasive (e.g., meningitis) and non-invasive infections (e.g., throat and skin infections), toxin-mediated diseases (e.g., necrotizing fasciitis, scarlet fever, streptococcal toxic shock syndrome) and immune mediated diseases (e.g., rheumatic fever, rheumatic heart disease, post-streptococcal glomerulonephritis) (Ralph and Carapetis, 2013). *S. pyogenes* ranks in the top 15 bacterial infectious causes of mortality in humans (Barnett et al., 2019; Ikuta et al., 2022).

Additionally, *S. pyogenes* is the type species of the genus *Streptococcus*, the type genus of the family *Streptococcaceae*, and, therefore, represents a taxonomic hallmark among these bacteria and within the order *Lactobacillales*. The complete genome sequences of *S. pyogenes* that are publicly available in NCBI present a G+C content (mol%) ranging from 38.2 to 38.6%, and a genome size ranging from 1.70 to 1.92 Mb.

1.6.2.2 *Streptococcus pneumoniae*

Streptococcus pneumoniae (Klein, 1884; Chester, 1901), also known as "pneumococcus", is an α-haemolytic *Streptococcus* and a major human pathogen that can cause local and invasive infections, such as otitis media, pneumonia, septicaemia or meningitis. Pneumococcal infections have higher morbidity among children and elderly people, and result in numerous deaths worldwide. For instance, a study estimated that pneumococcal pneumonia caused 1.5 million deaths in all ages, worldwide, in 2015, of which nearly 400,000 were children under the age of five (Troeger et al., 2017). A more recent study placed *S. pneumoniae* among the top 3 bacterial infectious causes of mortality in humans in 2019 (Ikuta et al., 2022). Another study estimated that *S. pneumoniae* caused more than 300,000 deaths globally among children under the age of five, in 2015 (Wahl et al., 2018). The same study estimated that pneumococcal deaths in children under the age of five decreased a 51% from 2000 to 2015, following the implementation of pneumococcal conjugate vaccines targeting various of the already more than one hundred described capsular serotypes (Ganaie et al., 2020; Pimenta et al., 2021).

These data are encouraging and demonstrate the effectiveness of the vaccination programs. However, the selective pressure exerted by the vaccines often results in serotype replacement effects, i.e., the progressive replacement of serotypes covered by the conjugate vaccines by other serotypes not covered by the vaccines (Hicks et al., 2007; Weinberger et al., 2011). Additionally, resistance of *S. pneumoniae* to several classes of antibiotics is progressively increasing, which further complicates the treatment of infections (Whitney et al., 2000; Cherazard et al., 2017). Therefore, *S. pneumoniae* continues to represent a major human health problem that is far from being resolved. Indeed, the World Health Organization included penicillin-non-susceptible *S. pneumoniae* in the list of priority pathogens for which research and development of new antibiotics is urgently needed (Tacconelli et al., 2018).

On top of all that, identification of *S. pneumoniae* is often problematic. *S. pneumoniae* is a member of the Mitis-Group of the genus *Streptococcus* (Kawamura et al., 1995), a group that

contains multiple closely-related species: *S. pneumoniae, Streptococcus australis, Streptococcus cristatus, Streptococcus dentisani* (also proposed as subspecies of *S. oralis*) (Jensen et al., 2016), *Streptococcus gordonii, Streptococcus infantis, Streptococcus lactarius, Streptococcus massiliensis, Streptococcus mitis, Streptococcus oligofermentans, Streptococcus oralis, Streptococcus oricebi, Streptococcus panodentis, Streptococcus parasanguinis, Streptococcus peroris, Streptococcus pseudopneumoniae, Streptococcus rubneri, Streptococcus sanguinis, Streptococcus sinensis* and *Streptococcus tigurinus* (also proposed as subspecies of *S. oralis*) (Jensen et al., 2016). Most species of this group are typically regarded as commensals of the oropharyngeal tract, although some can also cause opportunistic infections under certain conditions (Whiley and Hardie, 2015).

Within this group, *S. pneumoniae* forms a well-defined cluster, together with *S. pseudopneumoniae* (its most closely-related species) and *S. mitis* (Jensen et al., 2016; Jensen et al., 2021b). These two other species are often considered commensal organisms of the human oropharyngeal tract, although there is increasing evidence for the pathogenic potential of *S. pseudopneumoniae* (Garriss et al., 2019), and, under certain circumstances, they can both cause local and invasive infections (Rolo et al., 2013; Shelburne et al., 2014).

Classical identification of *S. pneumoniae* typically relies on its optochin susceptibility and bile solubility (Satzke et al., 2013). However, correct identification is often challenged by atypical pneumococci and by bile soluble and/or optochin susceptible isolates of other species of the Mitis-Group, which may be erroneously identified as *S. pneumoniae* (Richter et al., 2008; Rolo et al., 2013; Sadowy et al., 2020). Other phenotypic methods of identification such as API® rapid ID 32 or VITEK® 2, and MALDI-TOF MS, are also problematic (Teles et al., 2011; Jensen et al., 2021a). This situation is both due to the close relationship of these species and to the extensive horizontal gene transfer occurring between streptococci sharing the same ecological niche (Donati et al., 2010; Sanguinetti et al., 2012).

The genomes of *S. pneumoniae* are relatively small and have a medium-low G+C-content. As of 12th January 2023, 146 publicly available complete genome sequences of *S. pneumoniae* were available in NCBI, with a G+C content (mol%) ranging from 39.4 to 39.9%, and a genome size ranging from 1.98 to 2.31 Mb.

1.6.3 The family *Enterobacteriaceae*

The family *Enterobacteriaceae* (Rahn, 1937) (within the order *Enterobacterales* and class *Gammaproteobacteria*) is a relatively large group of genetically and phenotypically-related bacteria that currently (January 2021) encompasses 34 genera and 134 species with validly published and correct name (Parte, 2014). Bacteria within the family are Gram-stain negative, facultatively anaerobic, non-spore forming, chemoorganotrophic, and have both respiratory and fermentative metabolism. When fermenting D-glucose, other carbohydrates and polyhydroxyl alcohols, they often produce acid and observable gas. Most members are catalase positive and all, except *Plesiomonas*, are oxidase negative (Brenner and Farmer III, 2015). Most can reduce nitrate to nitrite and are flagellated and motile. Cells are rod-shaped, with an approximate size of 0.3 – 1.0 x 1.0 – 6.0 µm (Brenner and Farmer III, 2015).

Members of the family are ecologically very diverse, and while some of them can be found in environmental samples such as water or soil, others are common members of the normal microbiota of animals and plants, of which some can cause disease. Indeed, several members

of the family have a well-known pathogenic potential for humans (e.g., *E. coli*, *Klebsiella pneumoniae*, *Shigella* spp. and *Salmonella* spp.) (Ikuta et al., 2022), while many others can cause opportunistic infections. Members of the family can cause an extensive range of diseases in humans, such as diarrheal disease (typically food or water-borne), infections of the urinary tract, the respiratory system and wounds, as well as septicaemia and meningitis (Brenner and Farmer III, 2015).

As for many other clinically-relevant groups of bacteria, resistance to the most commonly used antibiotics is an increasing problem among various members of the family *Enterobacteriaceae* (World Health Organization, 2014). Indeed, the World Health Organization included carbapenem-resistant, ESBL-producing *Enterobacteriaceae* members in the list of priority pathogens for which new antibiotics are urgently needed (Tacconelli et al., 2018).

The genomes of the members of the family *Enterobacteriaceae* have a variable size and G+C content (mol%), ranging from 3.3 to 5.7 Mb and from 47.5 to 58.6%, respectively, among the type strains of the type species of the genera.

1.7 Outline and contents

In this thesis, bacterial whole-genome sequencing was applied for establishing reference genome sequences of selected strains of the three different taxonomic groups of bacteria presented above (i.e., *Stutzerimonas balearica*, formerly *Pseudomonas balearica*, the genus *Streptococcus* and the family *Enterobacteriaceae*), with the purpose of serving as a solid ground for downstream studies and applications.

In Paper I, the first genome sequence of *S. balearica*, a marine bacterium with potential to degrade aromatic compounds and to thrive in polluted environments, was determined. This was the complete genome sequence of the type strain, and for that reason, and for being the first genome sequence of the species, represents an important hallmark within the genus *Stutzerimonas* and the family *Pseudomonadaceae*.

Subsequently, in Paper II, the draft genome sequences of two additional strains of *S. balearica*, the second and third strains of the species to be made publicly available, were determined. In this paper, the high-quality complete genome sequence of *S. balearica* DSM 6083^T was used as a reference to obtain reference-based assembled contigs of these two additional strains of the species, which were later combined with contigs assembled *de novo* in order to obtain high-quality draft genome sequences. These second and third genome sequences of *S. balearica* enabled the possibility of performing intra-species comparative genomic analyses to assess the diversity of the species from a genomic perspective.

In Paper III, the draft genome sequence of the type strain of *S. gordonii* was determined and presented. This whole-genome sequence, although not complete, established a taxonomic hallmark for one of the members of the Mitis-Group of the genus *Streptococcus* and contributed to increase the coverage of the genomic diversity of this group of closely-related species that has diverse direct implications for human health.

In Paper IV, the first complete genome sequences of the type strain of *S. pyogenes* (i.e., Group A *Streptococcus*) were presented. *S. pyogenes* is the type species of the genus *Streptococcus*, which in turn is the type genus of the family *Streptococcaceae*. Therefore, these two genome

sequences represent a major taxonomic hallmark not only for a species which is a major human pathogen, but also for higher taxonomic levels within the phylum *Firmicutes*. Additionally, and unexpectedly, the two different sequencing and assembly strategies (PacBio versus Illumina plus Oxford Nanopore sequencing), resulted in identical complete genome sequences, thus demonstrating the reliability of both approaches and that hybrid assemblies can circumvent the error rate of Oxford Nanopore sequencing.

Thus, Papers I to IV present reference genome sequences for species which are relevant to humans for their potential in bioremediation and role in polluted environments (i.e., *S. balearica*), for being members of clinically-relevant groups of bacteria (i.e., *S. gordonii*, within the Mitis-Group of the genus *Streptococcus*), or for being major human pathogens (i.e., *S. pyogenes*). However, a major problem that the scientific community is currently facing is the poor quality and the large amounts of errors in public genome databases. Additionally, in our experience, many users are not aware of it, and for that reason, it is vital to give visibility to this problem by putting efforts to increase the awareness of this issue among users of public sequence databases.

In Paper V, this problem with public genome databases is addressed, by demonstrating that five genome sequences which were published as genome sequences of type strains, were not actually genome sequences of type strains. Moreover, one of them was misidentified at the species level. Thus, this Paper aimed to show the presence of "false" type strain genome sequences, and once more, about the presence of misidentified genome sequences in public databases. Thus, this Paper addresses the current quality crisis of public sequence databases, and contributes to increase the awareness of the scientific community about this problem, which is paramount to prevent sterile scientific efforts, as well as errors and wrong interpretations in downstream studies and applications relying on them.

In Paper VI, the genome sequences determined in Papers I and II, the genome sequence of a strain determined in this study, and 14 publicly available genome sequences of *S. balearica* (seven derived from isolated strains and seven MAGs), were used to perform a comparative genomic analysis of *S. balearica*. Tens of genome sequences of strains listed as *S. stutzeri* in GenBank and of type strains of *Stutzerimonas* were also included in the analysis. This allowed the exploration of the genomic diversity of *S. balearica*, in its phylogenomic context. It revealed that *S. balearica* is a diverse species that has an open pan-genome, and that new strains will continue revealing novel genes and possible novel functionalities. It also enabled the reevaluation of the 16S rRNA gene signature nucleotide positions that were previously described for differentiating *S. balearica* from *S. stutzeri* (Bennasar et al., 1996), and the determination of the intra- and intra-species variability. This led to the design of a strategy to detect strains of *S. balearica* in public sequence databases. The use of this strategy revealed 16S rRNA gene sequences of 158 additional strains of *S. balearica* and elucidated, at a larger scale, the habitats and environmental distribution of the species. The genome sequences of *S. balearica* were used also to search central and peripherical routes for catabolism of aromatic compounds. The searches were complemented with growth assays on the six strains that were available, which confirmed that *S. balearica* has varied capacities of degrading aromatic compounds. The analyses also revealed multiple features that provide more insights into the diversity and lifestyle of *S. balearica*.

In Paper VII, the genome sequence presented in Paper III, together with more than 300 publicly available genome sequences of non-pneumococcal species of the Mitis-Group of the genus *Streptococcus*, and more than 300 genome sequences of *S. pneumoniae*, were used to confirm that a gene described as highly-conserved in a previous pan-genomic study (Donati et al., 2010), represents a robust gene biomarker specific for *S. pneumoniae*. Subsequently, a PCR assay was developed for reliable differentiation of *S. pneumoniae* from other species of the Mitis-Group. The assay was tested *in vitro* on the type strains of 15 species of the Mitis-Group and on 25 clinical isolates of *S. pneumoniae* and closely-related species, with available genome sequence and taxonomic identity confirmed by ANIb versus the type strains.

In Paper VIII, a bacterium isolated from a wound infection that could only be identified as member of the family *Enterobacteriaceae* at the routine clinical laboratories of the Sahlgrenska University Hospital, was characterized using a polyphasic approach. Initial phenotypic and genotypic analyses determined that the isolate was a member of the family *Enterobacteriaceae*. However, the strain could not be assigned to any previously described species or genus, and moreover, they were not able to reliably determine its most closely-related genus. Meanwhile, whole-genome sequencing and several complementary phylogenomic analyses allowed us to conclude that the strain represented a novel genus within the family *Enterobacteriaceae*. Additionally, genome sequence analysis revealed a potential novel quinolone resistance gene variant. Cloning and expression in *E. coli* demonstrated its functionality, hence the gene was proposed as a novel *qnrB* variant (*qnrB96*).

Altogether, in this thesis, whole-genome sequencing was used to establish important reference sequences and taxonomic hallmarks that contribute to improve the genomic catalogue of important reference strains. Additionally, it also contributed to increased awareness in the scientific community about errors in public sequence databases. Subsequently, several genome sequences of *S. balearica* were used to perform a comparative genomic analysis of the species and provide insides into its wide genomic diversity, habitats and potential for degradation of aromatic compounds. Also, hundreds of publicly available genome sequences were used to determine a robust pneumococcal gene biomarker and to design a species-specific PCR assay for differentiating *S. pneumoniae* from other species of the Mitis-Group. Finally, whole-genome sequencing was applied to determine the taxonomic assignment of a clinical isolate and demonstrate that it represents a novel genus and species within the family *Enterobacteriaceae*. Additionally, its whole-genome sequence revealed a novel quinolone-resistance gene variant which was subsequently proved to be functional.

Overall, whole-genome sequencing allowed us to provide the deepest genomic insights into *S. balerica*, to determine a robust gene biomarker for one of the major human pathogens and to determine that a clinical isolate represented a novel genus and species within a taxonomically complex bacterial family. The specific contributions of this thesis demonstrate once again how the latest whole-genome sequencing developments have definitely pushed the limits of microbiology and of biological sciences in general, by allowing us to dive deeper than ever and with extremely high resolution into the genomic entrails of bacteria.

2 Aims

1. To determine and characterize the genome sequences of strains of the environmental bacterium *Stutzerimonas balearica* (Papers I, II), of type strains of clinically-relevant species of the genus *Streptococcus* (Papers III, IV), and to increase awareness of the presence of misidentified genome sequences in public databases (Paper V).
2. To use genome sequences to explore the genomic diversity, habitats and potential for biodegradation of aromatic compounds of *S. balearica* (Paper VI).
3. To exploit genome sequence data to identify a species-unique gene biomarker for the human pathogen *Streptococcus pneumoniae* and design a PCR assay for differentiating it from closely-related species (Paper VII).
4. To use genome sequences to establish the phylogenetic relationships and the taxonomic identity of a clinical isolate of the family *Enterobacteriaceae* that could not be identified (Paper VIII).

4 Discussion

4.1 The impact of whole-genome sequencing in microbiology

DNA encodes the basis of the biology of any living organism; hence it is not surprising that whole-genome sequencing and, especially, the advent of high-throughput DNA sequencing methodologies have revolutionized life sciences, in general, and have had a major impact on the study of microorganisms in particular (Loman and Pallen, 2015; Shendure et al., 2017). Thus, bacterial whole-genome sequencing, alone or combined with other approaches, has become the basis for numerous downstream applications. It has opened the door to new possibilities in, for instance, forensics and outbreak investigations (Read et al., 2002), genome-wide association studies (Sheppard et al., 2013), detection of DNA chemical modifications, such as methylation (Flusberg et al., 2010), mining of new drug candidates (Medema et al., 2011), identification of new drug targets (Andries et al., 2005), or reverse vaccinology (Pizza et al., 2000). Indeed, only a few years ago, Shendure *et al*. (2017) predicted, "*the impact of DNA sequencing will be on a par with that of the microscope*". Perhaps it is still early for confirming such a strong forecast, but it is clear that the latest DNA sequencing advances are imparting a tremendous impact in life sciences.

4.2 Contributions of this thesis

In this thesis, bacterial whole-genome sequencing was applied in four main settings:

1) to establish high-quality reference genome sequences (Papers I – IV, VI and VIII), as well as to warn users about the presence of mislabeled genome sequences in public databases (Paper V); 2) to perform a comparative genomic analyses of the marine bacterium *S. balearica* and to determine its genomic diversity, habitats and potential for biodegradation of aromatic compounds (Paper VI); 3) to determine a gene biomarker and develop a specific PCR-amplification-based assay for accurate identification of the major human pathogen, *S. pneumoniae* (Paper VII); and 4) to characterize a clinical isolate of the family *Enterobacteriaceae* that could not be ascribed to any known genus by the routine clinical microbiology laboratories of the Sahlgrenska University Hospital (Paper VIII).

Cutting-edge whole-genome sequencing approaches have been employed to determine and establish draft and complete reference genome sequences of several nomenclatural type strains and additional strains of environmentally- and clinically-relevant bacteria. In total, four strains of the environmental bacterium *S. balearica* (Papers I, II and VI), two type strains of clinically-relevant species of the genus *Streptococcus* (Papers III and IV) and strains of the family *Enterobacteriaceae* (Paper VIII) were determined and analyzed. Additionally, a warning was presented about the presence of "false" type strains and other reference sequences in public databases, by reporting five publicly available genome sequences mislabeled as type strains, of which, one was misidentified at the species level (Paper V). Subsequently, the genome sequences generated in Papers I, II and VI, together with numerous, additional, publicly available genome sequences, were used to perform a pan-genomic analysis of the species *S. balearica* and other members of the genus *Stutzerimonas* (Paper VI). Furthermore, hundreds of genome sequences of the Mitis-Group of the genus *Streptococcus* were analyzed, enabling the discovery of a robust gene biomarker and the design of a specific PCR-based assay for the

detection of a major human pathogen: *S. pneumoniae* (Paper VII). Lastly, whole-genome sequencing was utilized to characterize and identify the taxonomic position of a strain that, at the routine clinical microbiology laboratories, could only be identified to the family level. The inclusion of OGRIs, core genome-based phylogenetic treeing and genome characterization in a polyphasic approach, was key in the proposal of a novel genus and species of the family *Enterobacteriaceae*, i.e., *Scandinavium goeteborgense* gen. nov., sp. nov., and in the identification of a novel genetic and functional quinolone resistance gene variant (Paper VIII).

4.3 Establishment of reference bacterial genome sequences (Papers I – V, VI and VIII)

The advent of high-throughput DNA sequencing technologies has resulted in an almost exponential growth in genome sequencing in microbiology and a consequential increase in the number of publicly available sequences. However, despite the large numbers of publicly available genome sequences, there has been a large bias towards well-studied species, such as *E. coli*, *P. aeruginosa*, *S. pneumoniae*, etc. Indeed, although several large-scale projects have aimed to determine genome sequences of type strains (Mukherjee et al., 2017; Wu et al., 2018), thousands of type strains still remain unsequenced and, moreover, there are still species without any genome sequence representing them in public databases. Additionally, the intra-species diversity can be immense (Tettelin et al., 2005; Freschi et al., 2019), and, therefore, it is also necessary to sequence the intra-species genomic diversity, because having high-quality but also comprehensive databases of species' reference genome sequences is essential for a wide range of applications and has many potential benefits (Wu et al., 2009), especially now that numerous methods relying on genome sequence data are paving their way towards being applied routinely, such as metagenomics for diagnostics of infectious diseases (Li et al., 2021a) or "shotgun proteotyping" (Karlsson et al., 2020).

As part of this thesis, a small kernel of knowledge was generated by sequencing type strains and additional strains of several selected species. Papers I and II presented the first three genome sequences of the marine bacterium *S. balearica*. Paper III presented the genome sequence of the type strain of *S. gordonii*, a commensal member of the Mitis-Group of the genus *Streptococcus*. Paper IV presented the complete genome sequence of the type strain of a major human pathogen: *S. pyogenes*. Additionally, Paper IV demonstrated that hybrid assemblies combining Illumina and Oxford Nanopore reads can generate complete, closed, genome sequences of high quality, identical to those generated with PacBio sequencing. In Paper VI, the genome sequence of an additional strain of *S. balearica* was determined. In Paper VIII, the genome sequences of the type strains of two species of the family *Enterobacteriaceae* (*Buttiauxella izardii* and *Lelliottia nimipressuralis*) and the genome sequence of a clinical isolate that represents a novel genus and species within the family were determined.

4.3.1 The genome sequence of the type strain of *Stutzerimonas balearica* (Paper I)

In Paper I, the complete genome sequence of the environmental bacterium *S. balearica* DSM 6083^T was determined and presented. In this study, Roche 454 mate pair sequencing and Illumina paired-end sequencing were combined to obtain a nearly complete genome sequence, which subsequently was closed with the aid of Sanger sequencing reads. The reason for taking this tedious approach was because, at the time, PacBio long-read sequencing was just

commencing and highly expensive, and Oxford Nanopore sequencing was still under development.

This was the first genome sequence of *S. balearica* to be determined and made publicly available and, hence, represents a benchmark, providing the first genomic insights of the species and a reference base for future studies. Additionally, as the type strain of the species, the genome sequence serves as an important taxonomic reference within the family *Pseudomonadaceae* and the recently described genus *Stutzerimonas*. In fact, this complete, closed, genome sequence is nowadays included in the major databases of type strain sequences (Federhen, 2015; Yoon et al., 2017; Meier-Kolthoff and Göker, 2019; Shi et al., 2020).

4.3.2 The genome sequences of additional strains of *Stutzerimonas balearica* (Papers II and VI)

In Paper II, the draft genome sequences of *S. balearica* LS401 and st101 were determined and presented. Despite being draft genome sequences, marked efforts were put into obtaining high quality assemblies (i.e., with low fragmentation and low error rate), by combining reference-based and *de novo* assemblies. That strategy was constructive, although, because of limited improvements over *de novo*-only genome assemblies, it was considered impractical for future studies. These were the second and third genome sequences of *S. balearica* to be made publicly available and, therefore, they opened the door to studying the genomic diversity of the species. Indeed, the ANIb values and some differences observed in the initial analyses of the genome sequences (i.e., presence/absence of certain elements) already suggested that *S. balearica* is a diverse species, in consonance with previous multiple pan-genome analyses of species of the family *Pseudomonadaceae* (Udaondo et al., 2016; Gomila et al., 2017; Freschi et al., 2019; Whelan et al., 2021). Furthermore, in Paper VI, the draft genome sequence of an additional strain of *S. balearica* was determined.

4.3.3 The genome sequence of the type strain of *Streptococcus gordonii* (Paper III)

In Paper III, the draft genome sequence of *S. gordonii* CCUG 33482^{T} was determined and stands as a taxonomic reference for a species typically considered to be a commensal bacterium and an opportunistic pathogen (Douglas et al., 1993; Yombi et al., 2012). In fact, the type strain of the species was isolated from a patient with subacute bacterial endocarditis (White and Niven, 1946) and the genome sequence revealed numerous putative virulence factors for adhesion (e.g., fibronectin-binding protein), protease activity (e.g., C5a peptidase) and immune evasion (e.g., capsule).

The main rationale for sequencing the type strain of *S. gordonii* was to contribute to the completion of the catalogue of genome sequences of type strains of the Mitis-Group of the genus *Streptococcus*, to provide a genomic basis and taxonomic anchor reference for downstream applications, such as proteotyping (Karlsson et al., 2015) of the Mitis-Group and identification of species unique biomarkers (Karlsson et al., 2018; Salvà-Serra et al., 2018a; Karlsson et al., 2020).

It is important to highlight that the draft genome sequence of *S. gordonii* strain CCUG 33482^{T} was determined in 2015 and the paper in Genome Announcements reporting it (Paper III) was published on 24th March 2016. However, another paper presenting the draft genome sequence

of another deposit of the type strain of *S. gordonii* (ATCC 10558^{T}) was published in Genome Announcements on 18th February 2016 (Rasmussen et al., 2016), just 13 days after the submission of Paper III.

At first glance, this might seem an unfortunate and redundant exercise that reinforces the idea that scientific data should be made publicly available as soon as possible (e.g., in public databases and through preprints), to minimize the number of overlapping efforts. However, this allowed confirmation that the ATCC (American Type Culture Collection) and the CCUG culture collections have the same or nearly identical strain deposit, labeled as the type strain of *S. gordonii*. For instance, the ANIb value, calculated using JSpeciesWS (Richter et al., 2016), between the two draft genome sequences is 99.92%, and a single nucleotide polymorphism (SNP) analysis performed using CSI Phylogeny (Kaas et al., 2014), as described previously (Sabat et al., 2017), showed 109 SNPs. These differences could be due to errors in the sequencing process or in the draft assemblies or, perhaps, due to real differences between the two deposits. But, in any case, the comparison suggests that the two strain deposits are, at least, nearly identical and, indeed, both genome sequences are considered to be "type material" within the NCBI Taxonomy Database (Federhen, 2015). Moreover, sequencing multiple deposits of the same type strain in different culture collections has been recommended by GenBank, as any lack of consistency in sequence data would reveal errors related to nomenclatural type material (Federhen et al., 2016).

In fact, the CCUG sometimes has observed discrepancies between type strains deposited at different culture collections, normally due to mixing up, either by the original authors, intermediaries, or by the collections (Salvà-Serra et al., 2022). Another motivation for spending the time and finances in determining the genome sequences of multiple strain deposits is that differences, (e.g., mutations, loss of plasmids, etc.), may develop, due to numerous passages and, therefore, it is convenient that researchers working with a particular strain, utilize the genome sequence of the closest strain deposit, with the minimum number of passage steps, to minimize the risk of discrepancies.

A limitation of these two studies is, in each case, that only a draft genome sequence was produced. In the case of CCUG 33482^{T}, this was because, at the time of the initial sequencing, obtaining complete, closed, genome sequences was time-consuming and prohibitively expensive (i.e., PacBio was the only available third-generation sequencing platform). Additionally, despite its limitations, a draft genome sequence provided enough data for most applications and to cover the main necessity, which was to have a genome sequence of the type strain of this member of the Mitis-Group.

More recently, the complete genome sequence of the NCTC deposit of the *S. gordonii* type strain (NCTC 7865^{T}) was determined (accession number: LR134291), as part of the NCTC 3000 project (https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx). High ANIb values and low number of SNPs also suggest that the three culture collections have the same or nearly-identical strains, with low variation between their assembled genome sequences. Evidently, the genome sequence of the NCTC deposit is currently the ultimate reference of the species because it is complete, although, as indicated by Federhan *et al*. (2016), it is convenient and valuable to have genome sequences from multiple deposits; moreover, users of the CCUG strain deposit should use the CCUG genome sequence as the closest reference to their working material. In fact, now that the latest advances in third-generation sequencing have made the obtention of complete genome sequences more accessible than ever, the possibility of completing the genome sequence of *S. gordonii* CCUG 33482^{T}, as well as those of other type strains deposited at the collection, should be considered.

4.3.4 The genome sequence of the type strain of *Streptococcus pyogenes* (Paper IV)

In Paper IV, the first complete genome sequences of the type strain of *S. pyogenes* (CCUG 4207^T and NCTC 8198^T), the type species of the genus *Streptococcus*, the type genus of the family *Streptococcaceae*, were presented. Additionally, the sequencing of the two strain deposits resulted in identical complete genome sequences. This confirms that both culture collections have the same strain deposit labeled as the type strain of *S. pyogenes*. It demonstrates the value of sequencing multiple deposits of the same type strain in different culture collections (Federhen et al., 2016; Salvà-Serra et al., 2022). It also demonstrates how different whole-genome sequencing approaches can yield identical results. However, as discussed in Paper IV, this result was not expected, since differences between strain deposits, due to passaging over time, are considered normal. Thus, if the original intention of the study would have been to evaluate the two sequencing and assembly strategies, the same strain deposit and DNA preparation should have been sequenced, in order to eliminate those two potential sources of variability. In any case, this study demonstrates the reliability of these two long-read genome sequencing approaches and should aid to reduce skepticism about hybrid approaches that use Oxford Nanopore sequence reads. It is also important to highlight that recent developments in Oxford Nanopore sequencing (especially the R10 sequencing chemistry), have enabled the obtention of near-finished bacterial isolate-derived genome sequences and MAGs, using only Oxford Nanopore sequencing (i.e., without combining with Illumina sequencing) (Sereika et al., 2022).

S. pyogenes is a major human pathogen causing a wide range of diseases; additionally, it is the type species of the genus *Streptococcus*, which is the type genus of the family *Streptococcaceae*. Thus, the complete, closed, genome sequence of this strain not only represents an important reference for a prominent human pathogen, but an important taxonomic reference within the order *Lactobacillales*. In fact, a genomic taxonomic reference is what was missing for a long time for *S. pyogenes*. The species, as many other major human pathogens, already had hundreds of publicly available genome sequences and, among those, the genome sequence of strain SF370 will continue to be a model reference for studying this bacterium (Ferretti et al., 2001). However, it is the genome sequence presented in Paper IV that has been included in the major databases of type strain genome sequences (Federhen, 2015; Yoon et al., 2017; Meier-Kolthoff and Göker, 2019; Shi et al., 2020). Additionally, this complete genome sequence will serve as a high-quality reference for studies on *Streptococcus* taxonomy.

4.3.5 Genome sequences of members of the family *Enterobacteriaceae* (Paper VIII)

In Paper VIII, the draft genome sequences of three strains of the family *Enterobactericeae* were determined. These included the genome sequence of the clinical isolate, CCUG 66741, which was proposed as the type strain of a novel genus and species, and the genome sequence of the type strain of *Buttiauxella izardii* as well as the type strain of *Lelliottia nimipressuralis*, the type species of the genus *Lelliottia*. The last two were sequenced because they were observed to be closely-related to the strain CCUG 66741, according to different methods. On the one hand, *B. izardii* was one of the species that were most closely-related to *S. goeteborgense*, according to the analysis of phenotypic features with StrainMatcher (CCUG software; https://ccug.se/identification/strainmatcher). On the other hand, *L. nimipressuralis* (type species of the genus *Lelliottia*) was the type species of the genus with the fourth highest 16S rRNA gene sequence identity to *S. goeteborgense*.

In this study, the three genome sequences were determined, using an Ion Torrent PGM platform; that is the technology implemented at the Department of Clinical Microbiology of the Sahlgrenska University Hospital, mainly because of its scalability and relatively short turn-around time. This sequencing approach yielded a relatively highly fragmented draft genome sequence for strain CCUG 66741^T (192 contigs and N50 = 75 kb), which was not surprising, because Ion Torrent provides single reads, with lower accuracy than, for instance, Illumina sequencing. Despite the degree of fragmentation, the assembly was adequate for the limited purposes of the study. However, soon after the publication of the study, the complete genome sequence of the type strain of *S. goeteborgense* was determined, using Illumina plus Oxford Nanopore sequencing. Ion Torrent reads were not combined with those from Oxford Nanopore because both technologies have similar problems determining homopolymers. The complete genome sequence was not included in the initial study but will serve as a more valuable and long-standing reference sequence for *S. goeteborgense* and, indeed, it revealed a putative plasmid of 143 kb that had not been noticed before, because of inherent limitations of draft genome sequences.

4.3.6 Complete and draft genome sequences

While the genome sequences of the type strains of *S. balearica* and *S. pyogenes* were completed, the other seven genome sequences determined in this thesis were left in draft status. Additionally, most of the genome sequences from public databases that were used in Papers VI, VII and VII were also in draft status. The costs and effort associated with determining complete, closed, genome sequences were, and to some extent are, much higher than those of draft genome sequences.

The use of draft genome sequences may be considered a limitation of this thesis. For instance, in Paper VIII, the genome sequence of *S. goeteborgense* CCUG 66741^T was left in draft status, which was adequate for confirming the novelty of the strain and its taxonomic position and for the valid publication of a new taxonomic name and description of a novel quinolone resistance gene variant. However, by leaving it as a draft genome sequence, the presence of the putative plasmid of 143 kb was missed, which was later revealed by completing and closing the genome (GenBank accession number: GCF_003935895.2). Likewise, in Paper VI, the use of draft genome sequences prevented the prediction of the size of multiple integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs).

Apart from the lower contiguity, draft genome sequences are usually more prone to contain foreign contamination, misassemblies and lower completeness levels (as presented in the Introduction section). This can be a problem, as errors in assemblies can give rise to errors in downstream applications, such as core and pan-genome analyses (Tonkin-Hill et al., 2020). For instance, multi-copy genes are typically collapsed to a single sequence in draft assemblies, although in some cases, they might be directly missing. This is the case for sequencing the ribosomal operons, including the 16S rRNA genes, which because of their usually multicopy nature, it is common to encounter draft genome sequences that lack a full sequence of 16S rRNA gene. In some cases, this precludes the taxonomic verification of type strain genome sequences. For example, as part of another research project, we attempted to confirm whether a genome sequence that was indicated to represent the type strain of *Eisenbergiella tayi*, really was the reference for the type strain of this species. However, this genome sequence did not contain any 16S rRNA gene sequence, which was the only gene that had been sequenced by the

authors that delineated the species phylogenetically (Amir et al., 2014). Therefore, it was impossible to do sequence comparisons between the draft genome sequence and the original study and, therefore, the authenticity of the genome sequence could not be verified.

Thus, draft genome sequences may be seen to have limitations and be a source of problems. However, the risk of having major discrepancies can be reduced by using only genome sequences from RefSeq (O'Leary et al., 2016). Moreover, despite their limitations, draft genome sequences are sufficient for numerous downstream analyses and have enabled exploration of the genomic insights into microorganisms with immense levels of resolution. Additionally, in our experience, most genes are usually represented in draft genome sequences if the coverage is larger than 20x. Thus, unless the draft genome sequences have major problems (e.g., large amounts of contamination), other variables, such as the software used for annotation and pan-genome determination, could represent larger sources of errors and variation for comparative genomic analyses. Therefore, despite their limitations, draft genome sequences are highly valuable, although to uncover the full potential of genomic analyses for the study of microorganisms, determining high-quality complete, closed, genome sequences is essential (Fraser et al., 2002; Koren and Phillippy, 2015; Omasits et al., 2017).

Until recently, determining complete genome sequences was tedious and expensive, as demonstrated in Paper I, wherein relatively large economical, technical, and human efforts were required to determine the complete, closed, genome sequence of *S. balearica* DSM 6083^{T}. However, the latest advances in high-throughput DNA sequencing and, particularly in third-generation sequencing technologies, are taking us into a new era, in which the routine determination of complete or nearly complete bacterial genome sequences is becoming feasible. For instance, the determinations of the complete, closed, genome sequences of *S. pyogenes* in Paper IV were already much more straight-forward than the determination of the complete genome sequence of *S. balearica*, thanks to the application of long-read sequencing. Despite being technically more straightforward, the costs for determining such sequences were still high, considering that a PacBio and an Oxford Nanopore run could cost nearly 1,000 €. However, those prices have already dropped by approximately one order of magnitude if we consider that Oxford Nanopore barcoding can split the cost of a run into twelve and that the recently developed Flongle Flow Cell from Oxford Nanopore can provide up to 1.8 Gb of sequence data for about 60 € (as of August 2021). Additionally, the error rate of Oxford Nanopore raw data (its main limitation) is continuously improving, at the sequencing level and at the interpretation of the signal (i.e., basecalling accuracy) (Wick et al., 2019; Sereika et al., 2022). Therefore, it is reasonable to think that, in the near future, obtaining complete, closed and error-free bacterial genome sequences on-site, within just a few hours, will be feasible, enabling the possibility of recovering complete genomic strain information almost immediately.

4.3.7 Considerations for establishing bacterial reference sequences

In this thesis, a total of ten bacterial genome sequences were determined, which served as a solid base for several studies. However, is noteworthy the fact that different sequencing and bioinformatic approaches were used to determine the draft or complete, closed, genome sequences. Indeed, six different DNA sequencing technologies were used in this thesis, including the first-generation Sanger sequencing method, three NGS and two third-generation sequencing technologies. Although this may seem inconsistent, several reasons motivated this variety: 1) the goal of each sequencing project (e.g., draft or complete genome sequence); 2)

the advantages and disadvantages of each sequencing technology; and 3) the availability, access and costs of the different methods at a given point in time and situation, which can be highly variable due to the rapid evolution of DNA sequencing.

The genome sequence of the type strain of *S. balearica* was completed, using a laborious approach, i.e., combining Illumina, Roche 454 and Sanger sequencing reads (Paper I), because access to third-generation sequencing technologies was limited at the time the strain was sequenced. The genome sequences of the type strain of *S. pyogenes* were completed, using third-generation sequencing technologies (Paper IV). Some of the draft genome sequences were determined, using Illumina (Papers II, III and VI), while some others were determined, using Ion Torrent (Paper VIII). That was simply because Ion Torrent was the technology implemented at the Department of Clinical Microbiology of the Sahlgrenska University Hospital.

These issues reflect how much the DNA sequencing field has evolved in recent years and how obtaining complete genome sequences is continuously becoming cheaper, simpler and requiring less manual work and hands-on time. Indeed, it is important to consider that high-throughput DNA sequencing is a relatively young and 'hot' field with a lot of competition, which forces companies and developers to constantly improve the existing technologies and to develop new ones if they want to continue in the field. A consequence of this high competition is that multiple high-throughput sequencing technologies have already become extinct, such as Roche 454 and other sequencing platforms (e.g., Helicos) that have not been covered in this thesis. Meanwhile, other sequencing technologies are emerging (e.g., GeneReader, from Qiagen; DNBSEQ, from the Beijing Genomics Institute; and Omniome, acquired by PacBio, which claims to provide highly accurate short reads, determined by sequencing by binding), which results in further advancements and extends the landscape of the DNA sequencing field.

A comprehensive comparative list of the main commercially available high-throughput platforms was presented a few years ago by the World Health Organization (World Health Organization (WHO), 2018). For further information, several references in recent years have or reviewed and compared the different library preparation methods and sequencing technologies (Metzker, 2010; Loman et al., 2012; van Dijk et al., 2014; Goodwin et al., 2016).

In addition to the six different DNA sequencing technologies, numerous bioinformatic programs, pipelines and different assembly strategies were used in this thesis. Even in cases where the same sequencing approach was employed, different bioinformatic strategies were used for processing and generating quality genome sequence data. This is also a consequence of the rapid evolution of the field; parallel to the developments in high-throughput DNA sequencing, a myriad of tools for sequence data analysis are continuously being developed. In fact, the bioinformatics side of DNA sequencing is more dynamic, with new scripts, algorithms, tools and new versions being posted every year in public repositories, such as GitHub (www.github.com). Additionally, maintenance and further development of bioinformatic programs depend directly on the people and resources allocated for them, which means that they can evolve at different rates or, even become discontinued if the developers run out of funding. This implies that users should 'stay tuned' for the latest advances in the field and be aware that tools can overtake each other in the 'race' for being the best and most applicable and that they may perform differently with different datasets (Wick and Holt, 2019).

These observations indicate that whole-genome sequencing approaches have many variables that can affect all stages of the sequencing workflow and that it can sometimes be difficult for users to keep up with the latest developments and to determine what is more convenient. In general, whole-genome sequence determination workflows can be organized in four main stages: DNA extraction; library preparation; nucleic acid sequence determination; and data analysis. Although simple and straightforward, each stage must consider variables that can compromise the final result, such as the starting material (e.g., hard-to-lyse cells), the read length (e.g., short- or long-reads, with or without size selection), library preparation (e.g., GC-rich or GC-poor genomes can cause problems during PCR-based library preparation methods) (Aird et al., 2011), running time of sequence (critical in clinical settings), amount of data needed (i.e., sequencing depth), cost per Gb, or access to the sequencing platforms.

Thus, to minimize risks of failure, any genome sequencing workflow should be designed or selected carefully, typically in a backwards direction (goal → data analysis → sequencing → library preparation → DNA preparation), keeping in mind the nature of the starting material and what is the final goal of the sequencing project, i.e., why a particular genome will be sequenced and what are the intended downstream analyses and applications. And, in case of doubts, it is always wise to contact the service providers or sellers, who are generally good at giving recommendations and advice regarding the design or the selection of the proper sequencing workflow.

A critical step for any whole-genome sequencing workflow is the DNA extraction and purification and, considering that, many protocols, including commercial kits, are available. However, what really matters is that the DNA quantity and quality isolated from a particular microorganism fulfills the library preparation requirements. For instance, several studies have shown that using different DNA extraction protocols has no or limited effects on the final quality of Illumina sequencing of bacterial genomes (Becker et al., 2016; Pasquali et al., 2019; Nouws et al., 2020), and, in fact, in the papers presented in this thesis, a total of five different DNA extraction protocols or commercial kits have been used. However, for sequencing procedures that are more sensitive to contaminants, such as PacBio sequencing, service providers may recommend specific DNA preparation kits, such as the QIAGEN Genomic-tip kits, which are tedious but provide large amounts of high-quality DNA and are recommended for PacBio sequencing by the National Genomic Infrastructure Uppsala, (SciLifeLab, Sweden; www.scilifelab.se). Additionally, one should consider that the outcome may vary considerably, depending on the biology of each microorganism. For instance, isolating high-molecular weight might be challenging if a strain produces DNA-degrading DNases, such as the constitutive EndA of *S. pneumoniae* (Puyet et al., 1990). Thus, the selection of a DNA extraction method should consider the requirements of the library preparation, as well as the biological properties of the starting material.

The selection of a library preparation method should also consider the properties of the DNA that will be sequenced. For instance, some library preparation methods are PCR-based (e.g., Illumina Nextera), which require low amounts of DNA, but may have an associated amplification bias, which could be problematic with high- or low-GC content microorganisms (Chen et al., 2013). Meanwhile, others are PCR-free (e.g., Illumina TrueSeq), which require higher amounts of input DNA but do not suffer from amplification bias. Additionally, most library preparation methods have a size selection step, which might miss small plasmids if large inserts are selected (e.g., PacBio 40 kb). The library preparations are often done by specialized personnel at sequencing centers (i.e., companies or core facilities). If that is the case, one can

always discuss the best approach for a particular sequencing project and what are the exact DNA requirements.

In any case, to avoid problems, the rule of thumb should always be to obtain gDNA of good enough quality (i.e., an appropriate level of contaminants, impurities and integrity) and quantity to fulfill the library preparation and sequencing requirements. For that, it is critical to thoroughly evaluate the quality of the isolated DNA without 'skimping', for instance, using a NanoDrop™ instrument to determine the absorbance ratios, Qubit™ for accurate quantification of double stranded DNA and a gel electrophoresis (e.g., Agilent Tape Station analysis, using Genomic DNA ScreenTape) for assessing the DNA integrity, which is especially critical for long-read sequencing applications that require high-molecular weight DNA. Thus, the "golden rule" should always be to stick to the DNA quality guidelines established by the company or core facilities performing the library preparation and the sequencing, or by the manufacturer of the library preparation kit and sequencing platform, if doing it in-house. Additionally, it is strongly recommend to perform a taxonomic quality control before the sequencing. This can be done by PCR-amplification of 16S rRNA genes, using universal primers, followed by Sanger sequencing, chromatogram evaluation and comparative sequence analysis, using the tool, "16S-based ID" of EzBiocloud (Yoon et al., 2017), as well as house-keeping gene PCR-amplification and sequencing, in the case of some bacterial taxa.

Finally, it is worth mentioning that the output of Oxford Nanopore sequencing, when using a rapid sequencing kit (i.e., which does not include size selection), is related directly to the molecular weight of the input DNA (as a technical specialist said at an Oxford Nanopore event in 2018: "*What you put is what you get out*"). In Paper IV, a modified version (Salvà-Serra et al., 2018b) of the "Marmur" procedure (Marmur, 1961) was used to prepare gDNA for Oxford Nanopore sequencing. This protocol is not specifically focused to provide high-molecular weight DNA, but, the method has been robust and useful for determining complete, closed, bacterial genome sequences from a wide range of taxa, using the Oxford Nanopore sequencing platform, including Paper IV and several other publications from our lab that are not included in this thesis (Méndez et al., 2018; Jaén-Luchoro et al., 2019; Grevskott et al., 2020; Jaén-Luchoro et al., 2020; Méndez et al., 2020; Fernández-Juárez et al., 2021; Salvà-Serra et al., 2021; Valenzuela et al., 2021; Marathe et al., 2022; Salvà-Serra et al., 2023). Using the "Marmur"-based protocol, one typically obtains sequence reads of approximately 10 kb (average length) with reads longer than 100 kb, which is adequate for completing most bacterial genomes (Schmid et al., 2018). However, there are many protocols available that have been developed specifically for isolating high-molecular weight DNA, which can be used to obtain ultra-long Oxford Nanopore sequence reads. Good examples are MagAttract HMW DNA Kit (Qiagen, Hilden, Germany), Quick-DNA HMW MagBead Kit (Zymo Research, Irvine, CA, USA), Monarch HMW DNA Extraction Kit for Tissue (New England Biolabs (Ipswich, MA, USA), or the protocol published by Joshua Quick (Quick, 2018). Such protocols would enable the completion of bacterial genome sequences with lower sequencing depths, which would be advantageous if using, for example, the recently developed Oxford Nanopore barcoding kit for 96 samples.

4.3.8 The quality crisis: misidentified genome sequences and false "type strains" (Paper V)

The advent of high-throughput DNA sequencing started a genomic revolution that led to an almost exponential increase in the number of genome sequencing projects. This resulted in

public databases being flooded with multitudes of genome sequences determined in a myriad of small and large-scale projects (Mukherjee et al., 2021; Sayers et al., 2022). The availability of these genome sequences opened the door to numerous possibilities and is essential for many downstream applications. For instance, the studies presented in Papers VI, VII and VIII depended upon the use of tens or hundreds of publicly available genome sequences. However, even though several measures and quality controls have been implemented in the last years (Federhen, 2015; Federhen et al., 2016; Schäffer et al., 2017; Ciufo et al., 2018), public sequence databases are riddled with errors and inaccuracies that can be dragged in applications and studies relying on them (Beaz-Hidalgo et al., 2015; Gomila et al., 2015; Mukherjee et al., 2015; Jensen et al., 2016; Breitwieser et al., 2019; Steinegger and Salzberg, 2020). Indeed, in our laboratory at the CCUG, we have experienced these problems and encountered numerous errors in publicly available sequence data, such as misidentified genome sequences within many different taxonomic groups, such as the Mitis-Group of the genus *Streptococcus* (Gonzales-Siles et al., 2019; Gonzales-Siles et al., 2020), or sequences with extremely poor quality, such as that of *Cupriavidus basilensis* B-8 (accession number AKXR00000000), which is formed by 1,418 contigs and scaffolds and contains more than 2,000 pseudogenes (almost 25% of all annotated genes).

To top it all, according to our experience, many users are not aware of this situation and work with public data assuming that it is correct, which further increases the risk of wasting time and effort with low quality or misclassified sequence data that is not what the user thinks it is, as well as reaching wrong biological interpretations in downstream studies and applications. Hence to minimize these issues, it is crucial to increase the awareness of these problems by giving them visibility and showcasing the importance of doing controls when working with public sequence data.

During this thesis projecct, five publicly available genome sequences that were erroneously reported as type strains in the journal *Genome Announcements* were encountered. This led to an invitation from the Editor in Chief of the journal *Microbiology Resource Announcements* (successor of *Genome Announcements*) to report the issue in a Letter to the Editor (Paper V). Paper V demonstrated that those five genome sequences were not type strains and that, in addition, one of them was misidentified at the species level (i.e., *L. nimipressuralis* SGAir0187 is a member of the genus *Enterobacter*, not *Lelliottia*, closely-related to *Enterobacter roggenkampii*). This exemplifies, once again, that public genome sequence databases contain unreliable and misidentified sequences, an issue that can have major consequences in downstream applications, such as shotgun metagenomics or proteotyping, potentially leading to incorrect conclusions.

Thus, in Paper V, the problem of the quality crisis was addressed, by highlighting how important it is to have correct designations for type strain genome sequences and encouraging users of public databases to perform quality controls, instead of assuming that public sequence data are correct. Paper V is intended to increase the awareness among the scientific community of the ‘quality crisis’. Previous studies already reported the presence of misclassified genome sequences in public databases (Beaz-Hidalgo et al., 2015; Gomila et al., 2015; Jensen et al., 2016; Gonzales-Siles et al., 2019). However, Paper V is unique, in the sense that it reports the presence of genome sequences erroneously reported as type strains. Among those, it is especially concerning that one of the genome sequences was, in addition, misidentified at the species level, which implies that it could be used as a wrong taxonomic reference and lead to

additional genome sequences being misidentified. In fact, its misidentification might had been caused by another erroneous type strain, since there is another publicly available genome sequence that is labeled as *L. nimipressuralis* CIP 104980^T, which is the deposit of the type strain of *L. nimipressuralis* at the Collection de l'Institut Pasteur (CIP). The ANIb value between that genome sequence and the one of strain SGAir0187 is 94.98%, which suggests that they are possibly the same genomic species. However, the 16S rRNA gene sequence of the genome sequence labeled as strain CIP 104980^T shows low identity values (< 99%) with the 16S rRNA gene sequence of the genome sequence of *L. nimipressuralis* CCUG 25894^T, the genome sequence of *L. nimipressuralis* ATCC 9912^T (available at the ATCC on-line catalogue), the partial 16S rRNA gene sequence available at the CIP on-line catalogue and the sequence used in the effective publication of the species (Brady et al., 2013). This suggests that the genome sequence labeled as *L. nimipressuralis* CIP 104980^T is also a false "type strain" (perhaps caused by a mix up during the custody of the strain, passage or sample preparation for sequencing), that probably led to the misidentification of the genome sequence of strain SGAir0187. This further reinforces the idea of how important it is to determine genome sequences from multiple (if possible, all) type strain deposits, to authenticate them and ensure that there is consistency between them (Federhen et al., 2016; Salvà-Serra et al., 2022). It also highlights the importance of having a dedicated, comprehensive, and manually curated catalogue of type strain genome sequences. The latest was somehow addressed by GenBank some years ago, by tagging sequences of type material (Federhen, 2015), but more recently, a dedicated database for genome sequences of type strains was created (Shi et al., 2020).

Thus, the presence of false "type strain" genome sequences can potentially have major consequences in public databases, as well as on studies relying on them, by serving as erroneous taxonomic references. Therefore, it is paramount to increase the general awareness of these issues and to promote the practice of doing quality controls before trusting public genome sequences. Paper V should contribute to increasing the awareness of this problem and should, therefore, help to prevent errors and wrong biological interpretations in downstream analyses. Public databases should also be encouraged to put more effort into implementing and improving automatic software quality controls, as previously suggested (Bengtsson-Palme et al., 2016). Furthermore, another effect of Paper V was that the authors of the five publications presenting false "type strains" (who had not responded to multiple requests of the Editor in Chief of Microbiology Resource Announcements before the publication of our Paper) published corrections of the articles and removed the word "type" from the title, as suggested.

In any case, it can be expected that public sequence databases should continue putting efforts into improving the quality of their data. Additionally, the development of long-read third-generation sequencing technologies, by allowing the straightforward determination of high-quality complete genome sequences, will help to compensate for any loss in quality caused by the introduction of high-throughput short-read NGS technologies. The completion of the catalogue of genome sequences of type strains, and the determinations of multiple type strain deposits, will definitely help to prevent misidentified genome sequences. Furthermore, the constant development and improvement of tools for data evaluation also facilitates the performance of proper controls (Vezzi et al., 2012; Gurevich et al., 2013; Hunt et al., 2013; Parks et al., 2015). Thus, thanks to the development of better technologies and bioinformatic tools, we can expect that in the relatively near future, the 'quality crisis' will be resolved or, at least, minimized, although this should never exempt all stakeholders from performing quality controls.

Finally, it is important to emphasize that an effective solution to, at least, reduce the magnitude of these problems could be to involve the community, by facilitating a moderated post-submission editing of public sequence entries by third party users, as previously proposed (Pennisi, 2008; Bengtsson-Palme et al., 2016). Additionally, to diminish the risk of using "false type strains", we recommend users of type strain genome sequences to check whether they are labeled as, "assembly from type material", in GenBank or to obtain them from a manually curated database of type strain genome sequences, such as gcType (Shi et al., 2020). It is also a good practice to analyze the associated 16S rRNA gene sequences, using EzBiocloud, which contains a comprehensive and manually curated database of 16S rRNA gene sequences from type strains (Yoon et al., 2017). However, not even these carefully-curated databases are always free of errors, as discrepancies may occur between different culture collection deposits and, in some cases, the original publication lacks any kind of sequence data, which implies that phenotypic comparisons are the only way to determine which strain deposit adheres to the original description of the species. Fortunately, the current requirement of a whole-genome sequence with the description of novel taxa (https://www.microbiologyresearch.org/journal/ijsem/scope) should definitely minimize the risk of future issues and help to easily resolve any discrepancy.

4.3.9 En route towards a complete catalogue of type strain genome sequences

For many years, there has been wide consensus within the scientific community about the importance of having a complete catalogue of type strain genome sequences. Indeed, several major projects have been started with the aim of sequencing genomes of type strains (listed in the Introduction section), and since January 2018, the IJSEM requires authors proposing novel taxa to provide genome sequences. However, the catalogue of type strain genome sequences is not yet complete (Shi et al., 2020).

As of 22nd October 2022, the LPSN (Parte, 2014) contained 20,487 species with validly published and correct name (Meier-Kolthoff et al., 2022). Meantime, TYGS, which does not include species without validly published and correct names, contained 17,457 microbial type-strain genome sequences. This suggests that a significant fraction of prokaryotic type strains do not have a publicly available genome sequence yet. This is in accordance with the estimates that we made at the CCUG in 2021, wherein more than 500 of the 3,843 type strains maintained did not have determined genome sequences in August 2021. Additionally, most of the species that have publicly available type strain genome sequences have only draft genome sequences and/or a single strain deposit sequenced. For instance, only 3,634 of the 19,136 bacterial genome sequences from type strains available in RefSeq on 12th February 2023 (i.e., 19%) were complete genomes.

Ideally, all type strains should have at least a complete, closed, genome sequence and, if possible, the genome sequences of all type strain deposits (i.e., 73,866, according to the GCM) should be determined, in order to demonstrate consistency or reveal discrepancies, which could then be further analyzed to determine which type strain deposits are reliable, to ensure the highest quality standards for publicly available type strain sequence data (Salvà-Serra et al., 2022). Altogether, they would serve as taxonomic reference 'anchors' that would facilitate the correct identification of additional strains and novel isolates, as well as the detection of potential novel taxa, especially thanks to tools that have been developed in the last years for high-throughput genome-based identification (Rodriguez et al., 2018; Ha et al., 2019; Meier-

Kolthoff and Göker, 2019; Shi et al., 2020). The reality is that, at least, 200 type strains will not be accessible for various reasons (Yarza et al., 2013), although, in any case, more efforts should be put into generating a comprehensive and nearly-complete catalogue of type strain complete genome sequences. Moreover, considering how vast a species pan-genome can be (Freschi et al., 2019), this catalogue should be followed by additional high-quality sequences of all taxa, to obtain as much intraspecific coverage as possible.

4.3.10 En route towards a complete catalogue that includes the uncultivated majority

Having a "complete catalogue of type strain genome sequences" of species with validly published name such as those that have been the focus of this thesis, which rely on cultivating the strains in pure cultures, will be of great value. Sequencing additional strains of all those species to cover the intraspecies diversity will also be valuable. New method developments continue allowing the cultivation of previously uncultivated microorganisms (Lewis et al., 2020). However, the vast majority of bacterial and archaeal species remain uncultivated (Steen et al., 2019). Thus, culture-based whole-genome sequencing will always be biased and far from covering the whole diversity of prokaryotes.

Consequently, in order to have a complete catalogue that covers the whole spectrum of prokaryotes, it will be essential to recover whole-genome sequences from uncultivated microbes, using sequence constructs, such as SAGs or MAGs (Alneberg et al., 2018). Recently, multiple studies have presented thousands of MAGs of various quality levels (i.e., different levels of completeness and contamination) that significantly expand our knowledge of microbial diversity and unveil novel genes and putative novel biological functions and properties (Anantharaman et al., 2016; Parks et al., 2017; Tully et al., 2018; Pasolli et al., 2019; Nayfach et al., 2021). For instance, Nayfach *et al.* presented 52,515 MAGs reconstructed from more than 10,000 diverse metagenomic datasets, which were estimated to represent 12,556 possible novel species, 5,463 novel genera, 1,525 novel families and 456 novel orders (Nayfach et al., 2021). This exemplifies how important metagenomic studies are for exploring the prokaryotic diversity. However, despite their high value, most metagenomic studies include only short reads, which therefore results in highly fragmented assemblies, formed by numerous contigs or scaffolds that should afterwards be used for binning (i.e., grouping the ones that belong to the same taxon) to obtain MAGs. For this reason, most MAGs are left in draft status, presenting various levels of completeness, gaps, assembly errors, chimeras, and contamination, which should then be critically evaluated (Parks et al., 2015; Chen et al., 2020a; Meziti et al., 2021).

Nevertheless, the latest developments and cost decreases of long-read sequencing should facilitate a wide implementation of these technologies in shotgun metagenomics. That should enable high-throughput determinations of high-quality complete genome sequences of yet uncultivated prokaryotes (Chen et al., 2020a). Indeed, the first nearly complete genome sequences assembled from metagenomic data were presented in 2004, after using shotgun Sanger sequencing on a low-complexity acid mine drainage biofilm community (Tyson et al., 2004). However, the latest developments of TGS technologies are the ones that have definitively facilitated the recovery of complete genome sequences from metagenomic data (Driscoll et al., 2017; Stewart et al., 2019; Chen et al., 2020a; Moss et al., 2020; Sereika et al., 2022). Thus, the creation of a catalogue of reference high-quality complete genome sequences that span the entire prokaryotic diversity, should be feasible in the relatively near future. Additionally, the determination and analysis of genome sequences of still uncultivated taxa,

especially if combined with other omics approaches (e.g., transcriptomics and proteomics), enable prediction of their phenotypic features and can help to predict/design a suitable media formulations and optimal physicochemical conditions for cultivation (Gutleben et al., 2018).

4.4 Comparative genomic analysis of *Stutzerimonas balearica* (Paper VI)

S. balearica was proposed as a novel species in 1996 (by that time, as *P. balearica*), after the characterization of two halotolerant, denitrifying and naphthalene-degrading strains of the genomovar 6 of *P. stutzeri*, which had been isolated from a waste water treatment plant and from polluted marine sediment (Rossello et al., 1991; Rosselló-Mora et al., 1994; Bennasar et al., 1996). After that, several strains of the species were isolated and reported throughout the following years and these reports suggested that *S. balearica* is a species that can be found in marine and polluted environments and that shows diverse and interesting capabilities such as degradation of certain aromatic compounds. The reported capacities of *S. balearica* for biodegrading certain recalcitrant pollutants such as naphthalene, made it a species of interest for potential bioremediation applications. However, the complex diversity of the former *P. stutzeri* group (currently, *Stutzerimonas*) and the limited discriminatory power of the 16S rRNA gene often resulted in strains being misclassified or unclassified to the species level (Lalucat et al., 2006; Gomila et al., 2022; Li et al., 2022; Uddin et al., 2022). This limited our knowledge and understanding of the diversity and the biology of *S. balearica*.

The genome sequences reported in Papers I and II already enabled the possibility of performing intraspecies genomic comparisons of *S. balearica* and suggested that *S. balearica* is a highly diverse species. In Paper VI, the genome sequence of an additional strain of *S. balearica* was determined. Additionally, the genome sequence of the type strain (determined in Paper I), served as a taxonomic reference point to confirm that 14 additional publicly available genome sequences represented authentic strains of *S. balearica* (seven isolated strains and seven MAGs). Subsequently, these sequences were used, together with multiple other genome sequences of the genus *Stutzerimonas*, to perform a pan-genomic analysis of *S. balearica*, within its phylogenomic context. This demonstrated that *S. balearica* is a highly diverse species and that more genome sequences will continue to reveal additional genes and possibly novel functionalities. This finding is in line with other studies focused on other members of the family *Pseudomonadaceae* (Udaondo et al., 2016; Gomila et al., 2017; Freschi et al., 2019; Whelan et al., 2021) and reflects the vast diversity of *Stutzerimonas* (Lalucat et al., 2006; Gomila et al., 2022; Lalucat et al., 2022; Li et al., 2022). In fact, pan-genomic predictions of Paper VI indicated that most of the pan-genome has not been sequenced yet, and that sequencing more strains will continue to reveal significant numbers of new genes. This implies that the 11 isolate-derived genome sequences and the seven MAGs should be considered as just a subset of the genomic diversity of the species. Future studies should aim to recover more strains and genome sequences of *S. balearica* to better understand the diversity and the biology of the species. This highlights, once again, the importance of sequencing multiple strains per species and not only type or reference strains.

The isolate-derived genome sequences of *S. balearica* and the inclusion of tens of closely-related but non-*S. balearica* strains also allowed the determination of the intra- and inter-species sequence variability of the 16S rRNA gene of *S. balearica*. It also allowed the determination of

the conservation level of the 37 signature nucleotide positions presented with the description of the species for differentiating *S. balearica* from *S. stutzeri* and other closely-related species (Bennasar et al., 1996). This enabled the design and the application of a strict strategy for detecting publicly available 16S rRNA gene sequences of *S. balearica*, which resulted in the detection of 158 additional strains of *S. balearica*. Notably, most of these public sequences were either misclassified or not classified to the species level. This highlights the necessity for curating and improving the quality of public sequence databases. The metadata associated with these sequences revealed that *S. balearica* can be found in a wide range of environments, but that has been mostly found in aquatic and oil-related environments. Similar strategies could be designed and applied in future studies on other taxonomic groups which may not have high numbers of publicly available genome sequences. However, in other groups, the 16S rRNA gene may have lower resolution power. Indeed, in some cases, different species can have identical 16S rRNA gene sequences (Pain et al., 2020). Finally, the genome sequences of *S. balearica* included in Paper VI revealed that the species has a varied potential for degradation of aromatic compounds. These observed capabilities could be experimentally validated in six of the strains, which were available at the CCUG. Some aromatic hydrocarbons are among the most predominant and persistent pollutants in the environment and are in the top 10 of the Substance Priority List of the Agency for Toxic Substrates and Disease Registry (Agency for Toxic Substances and Disease Registry (ATSDR) Division of Toxicology and Human Health Sciences, 2022). This makes *S. balearica* a species of interest for bioremediation applications, especially since it can be naturally found in environments polluted with such compounds. The use of native microorganisms has been shown increase the efficiency of bioremediation applications, as these are already well adapted and able to thrive successfully in such environments (Tyagi et al., 2011).

In conclusion, in Paper VI, whole-genome sequencing data revealed that *S. balearica* is a highly diverse species, that can be found in a wide range of environments, but that has been more commonly encountered in aquatic and oil-related environments. It also showed that strains of *S. balearica* have various capacities for degrading aromatic compounds and therefore, have potential for being useful in bioremediation applications. Additionally, their genome sequences revealed multiple features such as a large regulatory network or a vast and varied set of genes for resistance to biocides and metals. Altogether, this suggests that *S. balearica* is an adaptable bacterium which has been found mostly in polluted environments, which further increases its potential usefulness for bioremediation applications. Moreover, the predictions of the pan-genomic analysis indicate that new genes and functionalities will continue appearing as soon as more genome sequences of the species are determined, which warrants future studies aiming to explore the diversity of *S. balearica*. Paper VI has increased our knowledge and understanding about the diversity, the habitats and the potential for biodegradation of aromatic compounds of *S. balearica* and will therefore serve as a basis for future studies aiming to work with the species.

4.5 A biomarker for identification of *Streptococcus pneumoniae* (Paper VII)

S. pneumoniae is a major human pathogen which is a member of the Mitis-Group of the genus *Streptococcus*, which also encompasses numerous commensal bacteria, with typically less

pathogenic potential, such as *S. gordonii* (Kawamura et al., 1995). However, *S. pneumoniae* and closely-related species of the Mitis-Group, such as *S. mitis* or the emerging pathogen *S. pseudopneumoniae* share many genotypic and phenotypic features (Kilian et al., 2008; Garriss et al., 2019). Within the Mitis-Group, *S. pseudopneumoniae* is the most closely-related species of *S. pneumoniae* (Jensen et al., 2016; Garriss et al., 2019). When *S. pseudopneumoniae* was proposed, based on the characterization of five strains, Arbique *et al*. reported that it could be differentiated from *S. pneumoniae* because of its resistance to optochin when incubated in the presence of 5% CO_2, bile insolubility and lack of capsule (Arbique et al., 2004). However, numerous strains of the two species that do not fulfill these requirements have been isolated over the years (Keith et al., 2006; Wessels et al., 2012; Rolo et al., 2013; Fuursted et al., 2016; Ganaie et al., 2021), which has resulted in misclassifications, as reflected by the high number of misclassified genome sequences of *S. pseudopneumoniae* in GenBank (Jensen et al., 2016; Croxen et al., 2018). So it is that even MALDI-TOF MS and MLSA analyses based on six and seven housekeeping genes have recently been shown to be also problematic (Jensen et al., 2021a; Jensen et al., 2021b). Meanwhile, correct identification of *S. pneumoniae* is crucial for accurate infection diagnostics and guidance of proper treatment, to assess the burden and epidemiology of pneumococcal infections and to assess the effects of pneumococcal vaccines (Satzke et al., 2013).

In Paper VII, advantage was taken from hundreds of publicly available genome sequences of *S. pneumoniae* and other species of the Mitis-Group (taxonomically verified by ANIb calculations with the type strain of *S. pneumoniae*), to confirm that the locus SP_1992 of the genome sequence of *S. pneumoniae* TIGR4 is a robust gene biomarker for identifying *S. pneumoniae*. Since 2015, the gene has been so-called "Xisco gene" at the Sahlgrenska University Hospital (Gothenburg, Sweden). However, it is important to highlight that Escolano-Martínez *et al*. characterized the protein encoded by SP_1992 and proposed that the gene should be named, *diiA*, for dimorphic invasion-involved A (Escolano-Martínez et al., 2016). Therefore, from now on, the proposed name *diiA* will be used along this thesis. The inclusion of hundreds of pneumococcal and non-pneumococcal genome sequences was critical for evaluating the robustness of *diiA* as a biomarker, because *S. pneumoniae* has an open pan-genome (Donati et al., 2010) and presents high-rates of horizontal gene transfer with other species of the Mitis-Group (Whatmore et al., 2000; Kilian et al., 2008). This genomic comparison also facilitated the development of a specific PCR assay for detection and identification of *S. pneumoniae*, which was validated *in vitro* on 15 type strains of the Mitis-Group and 25 clinical isolates (two of *S. oralis*, two of *S. mitis*, two of *S. pseudopneumoniae* and 19 of *S. pneumoniae*), with species identification confirmed by whole-genome sequencing and ANIb analyses.

Before the publication of Paper VII, multiple other gene biomarkers had been proposed for differentiating *S. pneumoniae* from closely-related species, although many of them were afterwards shown to be non-specific for *S. pneumoniae* and others had insufficient validation against a comprehensive set of strains of closely-related species (Satzke et al., 2013). For instance, *lytA* was widely used and appeared to be highly specific for *S. pneumoniae* (Messmer et al., 2004; Llull et al., 2006; Carvalho et al., 2007). However, the gene can also be found in other non-*S. pneumoniae* strains and assays can lead to false positive results (Johnston et al., 2010; Simões et al., 2016; Tavares et al., 2019; Ganaie et al., 2021). Therefore, *diiA* and the PCR assay presented in Paper VII could be useful for replacing other less specific assays and thus improve detection and identification of *S. pneumoniae*.

Moreover, after the publication of Paper VI, we also developed a real-time PCR assay, for application in the routine clinical laboratories of the Department of Clinical Microbiology of the Sahlgrenska University Hospital (Gothenburg, Sweden). This would facilitate rapid and effective detection of *S. pneumoniae* directly in clinical samples. Robust biomarkers such as *diiA* could potentially be used also for developing other assays such as immunoassays (e.g., enzyme-linked immunosorbent assay, ELISA) or even other more novel DNA-based methods such as detection using electric field-enhanced electrochemical CRISPR (Clustered Regularly Interspaced Palindromic Repeats) biosensors (Li et al., 2021b).

Despite their practicality, using single-gene biomarkers for identification will always entail a certain risk of misidentification. This is mainly because of horizontal gene transfer events, which in *S. pneumoniae* are enhanced by its natural genetic transformation capacity (Straume et al., 2015), and also because of the vast diversity of the Mitis-Group of the genus *Streptococcus* (Kilian et al., 2008; Donati et al., 2010). A possible solution for reducing the risk of misidentifications could be the use of multiple biomarkers in parallel (Gonzales-Siles et al., 2020; Karlsson et al., 2020). Indeed, other species-specific biomarkers for *S. pneumoniae* and closely-related species were proposed after the publication of Paper VII (Croxen et al., 2018; Garriss et al., 2019; Gonzales-Siles et al., 2020). Using multiple biomarkers might be more tedious and expensive but should be considered a wise strategy even if a biomarker such as the one presented in Paper VII shows high specificity. In fact, since the implementation of the assay at the CCUG, two clinical isolates have been found to be *diiA* positive but confirmed to be not *S. pneumoniae* by whole-genome sequencing and average nucleotide identity against the type strains of the Mitis-Group (both showing ANIb values below 94% and higher with the type strain of *S. pseudopneumoniae*).

In any case, shotgun approaches such as whole-genome sequencing, metagenomic sequencing or LC-MS/MS-based proteotyping, would be ultimately the most reliable methods for doing accurate species identification and characterization of clinical isolates (Karlsson et al., 2018; Rodriguez et al., 2018; Meier-Kolthoff and Göker, 2019; Imai et al., 2020; Li et al., 2021a). By relying on thousands of genes or proteins, such methods are much less susceptible to horizontal gene transfer events and offer additional information such as the presence of antibiotic resistance genes. They also facilitate the detection of emerging or unusual pathogens (Li et al., 2021a). However, their application in routine laboratories is often limited by technical and economic reasons (Rossen et al., 2018). Additionally, assays such as real time PCR typically have higher sensitivities than for instance shotgun metagenomics (Bengtsson-Palme et al., 2017).

Apart from the usefulness of *diiA* as a robust biomarker for identification of *S. pneumoniae*, we wondered also about the function of such a conserved gene. When the study was performed, analyses of the protein sequence of DiiA revealed a domain of unknown function (DUF) DUF1542 and an LPTXG cell wall anchor motif. However, no function could be assigned to the so-far considered hypothetical protein. Meanwhile, Escolano *et al*. showed the existence of two alleles of *diiA*, containing one or two imperfect repeats, encoding a dimorphic protein that is able to interact with collagen and lactoferrin and that, in mice models, is involved in lung internalization and proliferation in the blood, suggesting that DiiA is an important virulence factor of *S. pneumoniae* (Escolano-Martínez et al., 2016). Moreover, Escolano-Martínez found *diiA* present in all 560 pneumococcal clinical isolates analyzed, which further reinforces the robustness of *diiA* as a biomarker for *S. pneumoniae*.

Numerous datasets of LC-MS/MS-based proteotyping analyses of *S. pneumoniae* strains performed in our laboratory were screened (data not shown), to search peptides of DiiA. The

protein was not detected in any of the analyses, which suggests that DiiA might not be produced under those conditions or might be produced only at low levels. This is in line with Escolano-Martínez *et al*., who did not detect *diiA* using western blot but found modest expression of *diiA*. Despite its possibly low production, a previous study showed that DiiA is immunogenic (Giefing et al., 2008) and therefore, should be considered as a potential candidate for development of a serotype-independent vaccine (Escolano-Martínez et al., 2016). Indeed, a more recent study by Martín-Galiano *et al*. confirmed that vaccination of mice models with DiiA causes a complete immunogenic response, reduces nasopharynx colonization and provides protection to sepsis (Martín-Galiano et al., 2021). This study, which demonstrated the immunoprotective capacity of DiiA, reinforces the idea that this protein could be a good candidate for addressing the current necessity of serotype-independent pneumococcal vaccines. This could be done in combination with other pneumococcal surface proteins (Masomian et al., 2020). Paper VII, with the inclusion of hundreds of taxonomically verified pneumococcal and non-pneumococcal genome sequences, further strengthens the idea that DiiA could be a good serotype-independent vaccine candidate for *S. pneumoniae.*

Altogether, this exemplifies the potential of whole-genome sequencing for identifying robust biomarkers, designing specific diagnostic assays and for applying reverse-vaccinology strategies to find novel vaccine targets. It also shows, once again, how important is to determine multiple genome sequences for each species. In Paper VII, having hundreds of genome sequences of *S. pneumoniae* and other species of the Mitis-Group of the genus *Streptococcus* was essential to determine that *diiA* is a robust biomarker for *S. pneumoniae*. This high number of genome sequences enabled the design of an *S. pneumoniae*-specific PCR assay, but additionally, supports the idea that DiiA represents a potential candidate for the development of a *S. pneumoniae* universal vaccine (Escolano-Martínez et al., 2016; Martín-Galiano et al., 2021). It exemplifies the usefulness of analyzing multiple genome sequences to identify broadly-protective protein antigens, as demonstrated by Maione *et al.* when identifying a four-antigen-based universal vaccine for Group B *Streptococcus* (Maione et al., 2005).

At this point of the discussion, it is important to highlight that the name *diiA* should had been used in Paper VII, and not the informal term "Xisco" gene that is used within the Sahlgrenska University Hospital. The name *diiA* was proposed by Escolano-Martínez *et al*. one year before the submission of Paper VII. The publication of Escolano-Martínez *et al*. was missed in Paper VII and might have been detected by doing a deeper literature search before its submission. Sequence homology searches in public sequence databases were performed, but no information about *diiA* was found. If the name *diiA* and the knowledge generated by Escolano-Martínez *et al*. had been included in public databases and not only in the peer-reviewed publication, the chances that other users would have found and tracked this information would have been higher. As discussed by Abarenkov *et al*. in 2016, "(making sequence metadata available in public sequence databases) *highlights the wealth of relevant scientific information that lies buried in the last few decades' worth of scientific publications – formally available, yet only available to those who know where to look, and reachable only to those with access to that literature.*" (Abarenkov et al., 2016).

A possible way of handling the outcome of the study published by Escolano-Martínez *et al.* would had been to do a Third Party Annotation in GenBank (https://www.ncbi.nlm.nih.gov/genbank/tpa-exp/), followed by a submission to UniProt (The UniProt Consortium, 2023) and the Virulence Factors Database (VFDB) (Chen et al., 2005), both of which actively encourage the submission of new information about sequences (https://www.uniprot.org/help/submissions) and novel virulence factors

(http://www.mgc.ac.cn/VFs/feedback.htm). Such "post-publication" efforts might have led to higher visibility and impact of such a valuable study.

In conclusion, by comparing hundreds of genome sequences of *S. pneumoniae* and other species of the Mitis-Group of the genus *Streptococcus* it was determined that *diiA* is a robust gene biomarker for detection and identification of *S. pneumoniae*. Additionally, a simple and cost-effective PCR assay was designed for detecting *diiA* and thus differentiating *S. pneumoniae* from other closely-related species of the Mitis-Group of the genus *Streptococcus*. Applying real time PCR assays such as the one presented above, would allow direct detection of *S. pneumoniae* in clinical samples. Furthermore, several studies have further proved the specificity and robustness of this biomarker (Gonzales-Siles et al., 2019; Kilian and Tettelin, 2019; Gonzales-Siles et al., 2020). Moreover, because of its specificity and immunogenicity (Giefing et al., 2008; Martín-Galiano et al., 2021), DiiA might be a good candidate for a serotype-independent vaccine for *S. pneumomoniae*.

4.6 Characterization and description of a novel genus and species within the family *Enterobacteriaceae* (Paper VIII)

In Paper VIII, a strain isolated from a wound infection of an adult patient in Sweden, which could only be assigned to the family *Enterobacteriaceae* at the routine diagnostic laboratories of the Sahlgrenska University Hospital, was characterized. The strain was typed, using several phenotypic and genotypic approaches, but still, they showed discrepancies and none of them allowed the reliable assignment to any previously described genus within the family. Meanwhile, whole-genome sequencing allowed the precise taxonomic positioning of the clinical isolate that by other methodologies could only be identified to the family level. The genome-based analyses justified the proposal of *Scandinavium goeteborgense* as a novel genus and species within the family *Enterobacteriaceae*.

Additionally, genome sequencing enabled the discovery of a novel quinolone resistance gene variant (proposed *qnrB96*), with 92% of amino acid sequence similarity to its most closely-related variant, and which was subsequently shown to be functional. When expressed in *E. coli*, it conferred a five-fold increase in the minimum inhibitory concentration of ciprofloxacin, a fluoroquinolone class broad-spectrum antibiotic. Ciprofloxacin is one of the most commonly used antibiotics against infections caused by several members of the family *Enterobacteriaceae* (Iredell et al., 2016), and therefore, knowledge about the mechanisms and distribution of resistance to fluoroquinolone antibiotics is important.

The description and proposal of *S. goeteborgense* was made using a polyphasic approach. Thus, whole-genome sequencing was accompanied by an extensive phenotypic characterization. However, one limitation of this study is that the description and proposal was based on a single strain, which was supported in the phylogenetic and phylogenomic analyses by two public genome sequences of two other strains of *Scandinavium goeteborgense*. This led to a description which in the future might not apply to all members of the genus, but in any case, it is enough to demonstrate the novelty of *S. goeteborgense* and serves as a reference point for future studies working with further strains or proposing novel species within the genus *Scandinavium*. Indeed, a recent study proposed three novel species of *Scandinavium* and emended the description of the genus and *S. goeteborgense*, based on the characterization, using a polyphasic approach, of 11 strains isolated from the rhizosphere soil of *Quercus robur*

(English oak; both healthy and affected by acute oak decline) and bleeding lesions of *Tilia* spp. (lime) and *Quercus rubra* (red oak) trees (Maddock et al., 2022).

When the study was done, the strain had to be sent by the routine clinical laboratories of the Sahlgrenska University Hospital to the typing laboratory of the CCUG, where further analyses were done to determine if it represented a novel genus. This makes the identification and characterization process more tedious and costly. However, the simplification of whole-genome sequencing workflows and their implementation in clinical routine diagnostics (Balloux et al., 2018; Rossen et al., 2018), together with the availability of tools such as TYGS (Meier-Kolthoff and Göker, 2019), MiGA (Rodriguez et al., 2018), DFAST (Tanizawa et al., 2017), TrueBac ID (Ha et al., 2019), gcType (Shi et al., 2020), will probably allow genome-based isolate identification and clear detection of potential novel species directly in the clinics. However, having a comprehensive, complete and reliable database of genome sequences of the type strains of all species with validly published names, will be the key to uncover the full potential of these tools.

An important consideration is that the implementation of such approaches in the clinics would still take precious time needed for diagnostics, since a typical whole-genome sequencing workflow can take several days. However, the possibility of recovering draft and complete genome sequences without prior cultivation (i.e., SAGs and MAGs), will open the door to performing rapid whole-genome sequencing in the clinics. Moreover, Oxford Nanopore sequencing allows real-time sequencing, which implies that the sequencing data can be analyzed on real time using pipelines for species identification and detection of antibiotic resistances such as What's in my pot (WIMP) or Antimicrobial resistance mapping application (ARMA) (https://nanoporetech.com).

It is also important to remark that in prokaryotic taxonomy, boundaries are well-established for species circumscriptions (e.g., DDH = 70% and ANI = 95%). However, higher taxonomic ranks are applied unevenly across the tree of life and highly-studied groups tend to be more split than less studied ones (Parks et al., 2018). A good example is the family *Enterobacteriaceae*, which has been overclassified for medical purposes and would probably be treated as a single genus in other regions of the tree of life like in *Bacillus* (Abbott and Janda, 2006; Parks et al., 2018). In fact, this issue motivated the proposal of the GTDB, a standardized taxonomy for bacteria, based on phylogenomic analysis using 120 universal proteins (Parks et al., 2018).

Additionally, several genus boundaries have been proposed in the last years to reduce the arbitrariness. For instance, Yarza *et al.* proposed a 16S rRNA gene identity threshold of 94.5% for genus-level delineations (Yarza et al., 2014) and at the genomic level, an AAI of approximately 70% (Konstantinidis and Tiedje, 2007) and a 50% of percentage of conserved proteins (POCP) (Qin et al., 2014), were proposed. If these guidelines were followed strictly, numerous members of the family *Enterobacteriaceae* would have to be placed within one single genus. For instance, the AAI values between the type strain of *S. goeteborgense* and those of *Pluralibacter gergoviae* and *E. coli* are 82.04% and 79.81%, respectively. This is clearly above the proposed genus boundary, but, within the 74-85% inter-genus range of the family *Enterobacteriaceae* (Alnajar and Gupta, 2017).

In any case, the purpose of this study was not to question the status of the family *Enterobacteriaceae* but to define the taxonomic status of a clinical isolate within the currently accepted classification of the family. In that sense, core genome-based phylogenomic treeing confirmed the correct assignment of *Scandinavium* as a novel genus of the family *Enterobacteriaceae*. The use of phylogenomic treeing using at least 30 genes was

recommended for classifying genera and higher taxa (Chun et al., 2018). Additionally, this was also supported by OGRIs and an MLSA that included the other species of the genus *Pluralibacter*. The recent study that proposed three additional species of *Scandinavium* also support the genus status of *Scandinavium* and confirms that its most closely-related genus is *Pluralibacter*.

4.7 Future perspectives

The number of publicly available bacterial genome sequences is growing at an unprecedented pace and is expected to continue accelerating in the coming years. At present, most genome sequences being determined are left in draft status. However, a shift in the proportion of draft and complete genome sequences will probably occur in the near future, thanks to the on-going technical developments, especially in third-generation long-read sequencing technologies (e.g., longer reads and higher accuracy), which are leading us towards a dream situation in which determining complete genome sequences is more simple, rapid and inexpensive than ever. Indeed, long-read sequencing was declared method of the year 2022 by Nature Methods (Nature Methods, 2023). We may be at the doors of a "golden age" in which obtaining numerous high-quality complete genome sequences will be almost automatic. That will provide us with the most comprehensive and accurate pictures of the basis of the biology of microorganisms, which will benefit virtually all aspects of microbiology and related fields. It will, for instance, further facilitate in-depth and high-quality intra-species sequencing and thus, large-scale and accurate pan-genomic analyses and discovery of robust biomarkers for particular taxonomic groups or features, with accurate estimations of sensitivity and specificity. Thus, it seems clear that whole-genome sequencing will continue paving its way into our societies and daily lives, although for that, further developments, automation, and simplification of the entire sequencing process (i.e., from sample preparation to data analysis) are an essential requirement.

In any case, whole-genome sequencing is a "hot" field, with continuous expansion and a growing market, a fact that motivates investments to improve all aspects of the process. Thus, in the next years, we will certainly witness the advent of new methods for rapid and automated sample preparation (e.g., high-molecular weight DNA), new disruptive sequencing technologies and innovative data analysis approaches and tools. Consequently, the "democratization" of whole-genome sequencing will continue, and it will gradually enter more aspects of our lives. However, to make the most out of it and to be able to generate correct interpretations, it will be essential to dedicate large efforts to increase our knowledge about the functions and dynamics of genes and genomes, as well as to improve and develop new methods to annotate and to accurately deduce functions *in silico*.

One of the settings in which bacterial whole-genome sequencing is expected to continue gaining importance is in clinical microbiology. Currently, it is not feasible for most clinical microbiology laboratories to sequence all isolates, mainly because of the high costs that would be associated with whole-genome sequencing hundreds or thousands of clinical samples that are processed in routine labs. However, the costs and the turn-around times are expected to continue dropping in the next years, thanks to the continuous optimization and automatization of sample preparation, sequencing, and data analyses. This allows us to be very optimistic about the implementation of whole-genome sequencing in routine clinical microbiology laboratories for accurate identification and characterization of clinical isolates, such as detection of antibiotic resistance and virulence factors. An inherent limitation of genome-based analyses is the difficulty to predict expression, even if phenotypic predictions from genomic data improve over time. However, the on-going advances in other -omic technologies, such as proteomics,

suggest that integrative multi-omic approaches might be the best for obtaining the most comprehensive picture of microorganisms during diagnostics of infectious diseases. Currently, most of these approaches are typically limited to centralized laboratories. However, the advances in shotgun metagenomics and the development of devices that fulfill the ASSURE criteria (affordable, sensitive, specific, user-friendly, rapid and robust, equipment-free and deliverable) (Kettler et al., 2004) (in the direction of the Oxford Nanopore sequencer MinION), will probably, in the long run, enable effective genomic and perhaps multi-omic point-of-care testing.

For prokaryotic genome-based identification purposes, it is crucial to have reliable genome sequences available for all type strains. Fortunately, it is expected that nearly all type strains of species with validly published name will have at least a draft genome sequence available within the next few years. Moreover, thanks to the latest advances of whole-genome sequencing, it seems realistic to think that most type strains will also have high-quality complete genome sequences and sequences from multiple deposits determined relatively soon. For instance, when having Illumina and Oxford Nanopore sequence reads, the pipeline Unicycler can provide users with highly accurate "ready-to-go" complete genome sequences, within a few hours (Wick et al., 2017). The main limitation for completing the genomic catalogue of type strains will probably be the lack of access to some of them. However, that should only affect a small proportion, and, in any case, having a nearly complete and reliable catalogue of type strain genome sequences will already be a huge landmark that will not only facilitate identification, but also more robust detection and accurate classification of potentially novel taxa. Indeed, whole-genome sequencing will continue playing a major role in prokaryotic taxonomy, by helping in the description of novel taxa and in the clarification of already described taxonomic groups.

Other important issues to be addressed are the problems of the quality of public sequence databases, although there are reasons to be optimistic about these issues. On the one hand, archiving a complete and reliable catalogue of type strain genome sequences should facilitate a reduction in the proportion of misclassified genome sequences. On the other hand, advances in DNA sequencing technologies will sooner or later facilitate, even more effectively, the routine obtention of high quality complete or nearly complete genome sequences. This will minimize the generation of highly-fragmented, poor quality and contaminated genome sequences. Additionally, the awareness of the problems in public databases is expected to increase among the scientific community. This should stimulate the search for solutions, such as new tools or database architectures and, perhaps, facilitate the participation of third users in database curation. In any case, it will still be critical to make users aware of the importance of doing their own quality controls when working with publicly-available sequence data.

The studies presented in this thesis relied on cultivation-dependent whole-genome sequencing. However, an area that currently concentrates enormous interest from the scientific community is the accurate determination of whole-genome sequences of the "uncultivated majority". The latest developments in long-read sequencing have facilitated the culture-independent determination of complete genome sequences (e.g., obtention of complete MAGs and SAGs), although that is still limited and costly. However, we can expect that upcoming developments in long-read sequencing and data analysis will further facilitate obtention of high-quality and complete MAGs and SAGs. Therefore, soon or later, culture-independent determination of complete genome sequences will be applied for clinical diagnostics of infectious diseases, massively, rapidly, and at a competitive price, perhaps in combination with other -omic approaches such as shotgun prototyping. Thus, I believe that whole-genome sequencing is still just blooming, far from reaching its full potential, and that we have exciting times ahead, in

which we will witness major advancements, landmarks, and that those, in combination with other methods, will serve as a solid ground, benefit all aspects of the study of prokaryotes and bring the entire microbiology field to a next level.

5 Conclusions

The conclusions of this thesis are:

1. The complete genome sequence of the type strain and the draft genome sequences of three additional strains of *Stutzerimonas balearica*, a species with potential for biodegradation of aromatic compounds, were determined (Papers I, II and VI).
2. The draft genome sequence of the type strain of *Streptococcus gordonii*, a member of the Mitis-Group of the genus *Streptococcus*, was determined (Paper III).
3. The first complete genome sequences of the type strain of *Streptococcus pyogenes*, a major human pathogen and the type species of the genus *Streptococcus*, the type genus of the family *Streptococcaceae*, were determined (Paper IV).
4. The combination of Illumina and Oxford Nanopore sequence reads can provide complete genome sequences as accurate and identical to those obtained using only PacBio reads, which demonstrates the very high quality of these genome sequences (Paper IV).
5. There are sequences erroneously labeled as "type strains", sometimes even misclassified, which can lead to errors in studies relying on them (Paper V).
6. It is critical to perform taxonomic controls when reporting and when using public genome sequences (Paper V).
7. Comparative genomic analyses revealed that *Stutzerimonas balearica* is a diverse species with an open pan-genome and sequencing more strains will continue revealing new genes and potentially new functionalities (Paper VI).
8. *Stutzerimonas balearica* has been found in a wide range of environments but has been principally found in aquatic and polluted environments, most of them with oil-related compounds (Paper VI).
9. Nine percent of the 158 16S rRNA gene sequences of *Stutzerimonas balearica* were misclassified and 68% were assigned to a higher rank taxa (Paper VI).
10. Strains of *Stutzerimonas balearica* have a varied potential for biodegradation of aromatic compounds (Paper VI).
11. The analyses of hundreds of genome sequences revealed that *diiA* is a robust biomarker that can be used for differentiating the major human pathogen *Streptococcus pneumoniae* from other species of the Mitis-Group of the genus *Streptococcus* (Paper VII).
12. A PCR assay for detecting the gene *diiA* was developed and proposed as a reliable method for differentiating *Streptococcus pneumoniae* from closely-related species of the Mitis-Group of the genus *Streptococcus* (Paper VII).
13. Whole-genome sequencing and comparative genomics are effective approaches for discovering biomarkers.
14. The clinical strain CCUG 66741^{T} represents a novel genus and species within the family *Enterobacteriaceae* (*Scandinavium goeteborgense*) and carries a novel functional quinolone resistance gene variant (Paper VIII).
15. Whole-genome sequencing enables accurate phylogenetic positioning and delineation of novel microbial taxa.
16. Bacterial whole-genome sequencing is still at an early and rapidly evolving stage, but on-going and future developments – e.g., point-of-care diagnostics and *in situ* monitoring in clinical and environmental settings – will progressively lead to its wider implementation and influence in many aspects of our societies.

6 References

Aanensen, D.M., Feil, E.J., Holden, M.T., Dordel, J., Yeats, C.A., Fedosejev, A., et al. (2016). Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *MBio* 7(3). doi: https://doi.org/10.1128/mBio.00444-16.

Abarenkov, K., Adams, R.I., Laszlo, I., Agan, A., Ambrosio, E., Antonelli, A., et al. (2016). Annotating public fungal ITS sequences from the built environment according to the MIxS-Built Environment standard – a report from a May 23-24, 2016 workshop (Gothenburg, Sweden). *MycoKeys* 16. doi: https://doi.org/10.3897/mycokeys.16.10000.

Abbott, S.L., and Janda, J.M. (2006). "The genus *Edwardsiella*," in *The Prokaryotes: A Handbook on the Biology of Bacteria Volume 6: Proteobacteria: Gamma Subclass*, eds. M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer & E. Stackebrandt. (New York, NY: Springer New York), 72-89.

Adeolu, M., Alnajar, S., Naushad, S., and R, S.G. (2016). Genome-based phylogeny and taxonomy of the '*Enterobacteriales*': proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology* 66(12)**,** 5575-5599. doi: https://doi.org/10.1099/ijsem.0.001485.

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* 46(W1)**,** W537-W544. doi: https://doi.org/10.1093/nar/gky379.

Agency for Toxic Substances and Disease Registry (ATSDR) Division of Toxicology and Human Health Sciences (2022). *ATSDR's Substance Priority List 2022* [Online]. Available: https://www.atsdr.cdc.gov/spl/index.html [Accessed].

Ahmadi, M., Jorfi, S., Kujlu, R., Ghafari, S., Darvishi Cheshmeh Soltani, R., and Jaafarzadeh Haghighifard, N. (2017). A novel salt-tolerant bacterial consortium for biodegradation of saline and recalcitrant petrochemical wastewater. *Journal of Environmental Management* 191**,** 198-208. doi: https://doi.org/10.1016/j.jenvman.2017.01.010.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12(2)**,** R18. doi: https://doi.org/10.1186/gb-2011-12-2-r18.

Alm, R.A., Ling, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., et al. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* 397(6715)**,** 176-180. doi: https://doi.org/10.1038/16495.

Alnajar, S., and Gupta, R.S. (2017). Phylogenomics and comparative genomic studies delineate six main clades within the family *Enterobacteriaceae* and support the reclassification of several polyphyletic members of the family. *Infection, Genetics and Evolution* 54**,** 108-127. doi: https://doi.org/10.1016/j.meegid.2017.06.024.

Alneberg, J., Karlsson, C.M.G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M.V., et al. (2018). Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* 6(1)**,** 173. doi: https://doi.org/10.1186/s40168-018-0550-0.

Amir, I., Bouvet, P., Legeay, C., Gophna, U., and Weinberger, A. (2014). *Eisenbergiella tayi* gen. nov., sp. nov., isolated from human blood. *International Journal of Systematic and Evolutionary Microbiology* 64(Pt_3)**,** 907-914. doi: https://doi.org/10.1099/ijs.0.057331-0.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* 7(1)**,** 13219. doi: https://doi.org/10.1038/ncomms13219.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290(5806)**,** 457-465. doi: https://doi.org/10.1038/290457a0.

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data.".).

Andries, K., Verhasselt, P., Guillemont, J., Göhlmann, H.W.H., Neefs, J.-M., Winkler, H., et al. (2005). A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* 307(5707)**,** 223-227. doi: https://doi.org/10.1126/science.1106753.

Antipov, D., Korobeynikov, A., McLean, J.S., and Pevzner, P.A. (2016). HybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32(7)**,** 1009-1015. doi: https://doi.org/10.1093/bioinformatics/btv688.

Arbique, J.C., Poyart, C., Trieu-Cuot, P., Quesne, G., Carvalho Mda, G., Steigerwalt, A.G., et al. (2004). Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. *Journal of Clinical Microbiology* 42(10)**,** 4686-4696. doi: https://doi.org/10.1128/JCM.42.10.4686-4696.2004.

Arita, M., Karsch-Mizrachi, I., Cochrane, G., and on behalf of the International Nucleotide Sequence Database Collaboration (2020). The international nucleotide sequence database collaboration. *Nucleic Acids Research*. doi: https://doi.org/10.1093/nar/gkaa967.

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology* 36(7)**,** 566-569. doi: https://doi.org/10.1038/nbt.4163.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J., and Weightman, A.J. (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology* 71(12)**,** 7724-7736. doi: https://doi.org/10.1128/aem.71.12.7724-7736.2005.

Balloux, F., Brønstad Brynildsrud, O., van Dorp, L., Shaw, L.P., Chen, H., Harris, K.A., et al. (2018). From theory to practice: translating whole-genome sequencing (WGS) into the clinic. *Trends in Microbiology* 26(12)**,** 1035-1048. doi: https://doi.org/10.1016/j.tim.2018.08.004.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5)**,** 455-477. doi: https://doi.org/10.1089/cmb.2012.0021.

Barnett, T.C., Bowen, A.C., and Carapetis, J.R. (2019). The fall and rise of Group A *Streptococcus* diseases. *Epidemiology & Infection***,** 1-6. doi: https://doi.org/10.1017/S0950268818002285.

Beaz-Hidalgo, R., Hossain, M.J., Liles, M.R., and Figueras, M.J. (2015). Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for *Aeromonas* genomes in the GenBank database. *PLoS One* 10(1)**,** e0115813. doi: https://doi.org/10.1371/journal.pone.0115813.

Becker, L., Steglich, M., Fuchs, S., Werner, G., and Nübel, U. (2016). Comparison of six commercial kits to extract bacterial chromosome and plasmid DNA for MiSeq sequencing. *Scientific Reports* 6(1)**,** 28063. doi: https://doi.org/10.1038/srep28063.

Bengtsson-Palme, J., Boulund, F., Edström, R., Feizi, A., Johnning, A., Jonsson, V.A., et al. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *Proteomics* 16(18)**,** 2454-2460. doi: https://doi.org/10.1002/pmic.201600034.

Bengtsson-Palme, J., Larsson, D.G.J., and Kristiansson, E. (2017). Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy* 72(10)**,** 2690-2703. doi: https://doi.org/10.1093/jac/dkx199.

Bennasar, A., Rosselló-Mora, R., Lalucat, J., and Moore, E.R. (1996). 16S rRNA gene sequence analysis relative to genomovars of *Pseudomonas stutzeri* and proposal of *Pseudomonas balearica* sp. nov. *International Journal of Systematic Bacteriology* 46(1)**,** 200-205. doi: https://doi.org/10.1099/00207713-46-1-200.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218)**,** 53-59. doi: https://doi.org/10.1038/nature07517.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277(5331)**,** 1453-1462. doi: https://doi.org/10.1126/science.277.5331.1453.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4)**,** 578-579. doi: https://doi.org/10.1093/bioinformatics/btq683.

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* 15(1)**,** 211. doi: https://doi.org/10.1186/1471-2105-15-211.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15)**,** 2114-2120. doi: https://doi.org/10.1093/bioinformatics/btu170.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* 2(1). doi: https://doi.org/10.1186/2047-217x-2-10.

Brady, C., Cleenwerck, I., Venter, S., Coutinho, T., and De Vos, P. (2013). Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia* gen. nov. as *Lelliottia nimipressuralis* comb. nov. and *Lelliottia amnigena* comb. nov., respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter* gen. nov. as *Pluralibacter gergoviae* comb. nov. and *Pluralibacter pyrinus* comb. nov., respectively, *E. cowanii*, *E. radicincitans*, *E. oryzae* and *E. arachidis* into *Kosakonia* gen. nov. as *Kosakonia cowanii* comb. nov., *Kosakonia radicincitans* comb. nov., *Kosakonia oryzae* comb. nov. and *Kosakonia arachidis* comb. nov., respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis* nom. nov., *Cronobacter helveticus* comb. nov. and *Cronobacter pulveris* comb. nov., respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. *Systematic and Applied Microbiology* 36(5)**,** 309-319. doi: https://doi.org/10.1016/j.syapm.2013.03.005.

Bravakos, P., Mandalakis, M., Nomikou, P., Anastasiou, T.I., Kristoffersen, J.B., Stavroulaki, M., et al. (2021). Genomic adaptation of *Pseudomonas* strains to acidity and antibiotics in hydrothermal vents at Kolumbo submarine volcano, Greece. *Scientific Reports* 11(1)**,** 1336. doi: https://doi.org/10.1038/s41598-020-79359-y.

Breitwieser, F.P., Pertea, M., Zimin, A.V., and Salzberg, S.L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research* 29(6)**,** 954-960. doi: https://doi.org/10.1101/gr.245373.118.

Brenner, D.J., and Farmer III, J.J. (2015). "*Enterobacteriaceae*," in *Bergey's Manual of Systematics of Archaea and Bacteria*, eds. M.E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F.A. Rainey & W.B. Whitman.), 1-24.

Briggs, N., Campbell, S., and Gupta, S. (2021). Advances in rapid diagnostics for bloodstream infections. *Diagnostic Microbiology and Infectious Disease* 99(1)**,** 115219. doi: https://doi.org/10.1016/j.diagmicrobio.2020.115219.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273(5278)**,** 1058-1073. doi: https://doi.org/10.1126/science.273.5278.1058.

Burri, R.S., A. (1895). Ueber Nitrat zerstörende Bakterien und den durch dieselben bedingten Stickstoffverlust. *Zentbl. Bakteriol. Parasitenkd.* 1**,** 257–265, 350–364, 392–398, 422–432.

Carvalho, M.d.G.S., Tondella, M.L., McCaustland, K., Weidlich, L., McGee, L., Mayer, L.W., et al. (2007). Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *Journal of Clinical Microbiology* 45(8)**,** 2460-2466. doi: https://doi.org/10.1128/JCM.02498-06.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A.M., and Banfield, J.F. (2020a). Accurate and complete genomes from metagenomes. *Genome Research*. doi: https://doi.org/10.1101/gr.258640.119.

Chen, L.H., Yang, J., Yu, J., Ya, Z.J., Sun, L.L., Shen, Y., et al. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research* 33**,** D325-D328. doi: https://doi.org/10.1093/nar/gki008.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17)**,** i884-i890. doi: https://doi.org/10.1093/bioinformatics/bty560.

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., and Hwang, C.-C. (2013). Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLOS ONE* 8(4)**,** e62856. doi: https://doi.org/10.1371/journal.pone.0062856.

Chen, Z., Erickson, D.L., and Meng, J. (2020b). Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics* 21(1)**,** 631. doi: https://doi.org/10.1186/s12864-020-07041-8.

Cherazard, R., Epstein, M., Doan, T.-L., Salim, T., Bharti, S., and Smith, M.A. (2017). Antimicrobial resistant *Streptococcus pneumoniae*: prevalence, mechanisms, and clinical implications. *American Journal of Therapeutics* 24(3). doi: https://doi.org/10.1097/MJT.0000000000000551.

Chester, F.D. (1901). *A manual of determinative bacteriology*. New York: The Macmillan Company.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 10(6)**,** 563-569. doi: https://doi.org/10.1038/nmeth.2474.

Chivian, D., Brodie, E.L., Alm, E.J., Culley, D.E., Dehal, P.S., DeSantis, T.Z., et al. (2008). Environmental genomics reveals a single-species ecosystem deep within Earth. *Science* 322(5899)**,** 275-278. doi: https://doi.org/10.1126/science.1155495.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., et al. (2018). Proposed minimal standards for the use of genome data for the taxonomy of

prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 68(1)**,** 461-466. doi: https://doi.org/10.1099/ijsem.0.002516.

Chun, J., and Rainey, F.A. (2014). Integrating genomics into the taxonomy and systematics of the *Bacteria* and *Archaea*. *International Journal of Systematic and Evolutionary Microbiology* 64(Pt 2)**,** 316-324. doi: https://doi.org/10.1099/ijs.0.054171-0.

Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., et al. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International Journal of Systematic and Evolutionary Microbiology* 68(7)**,** 2386-2392. doi: https://doi.org/10.1099/ijsem.0.002809.

Contreras-Moreira, B., and Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* 79(24)**,** 7696-7701. doi: https://doi.org/10.1128/AEM.02411-13.

Costa, S.S., Guimarães, L.C., Silva, A., Soares, S.C., and Baraúna, R.A. (2020). First steps in the analysis of prokaryotic pan-genomes. *Bioinformatics and Biology Insights* 14**,** 1177932220938064. doi: https://doi.org/10.1177/1177932220938064.

Croxen, M.A., Lee, T.D., Azana, R., and Hoang, L.M. (2018). Use of genomics to design a diagnostic assay to discriminate between *Streptococcus pneumoniae* and *Streptococcus pseudopneumoniae*. *Microbial Genomics*. doi: https://doi.org/10.1099/mgen.0.000175.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long read sequencing data. *Bioinformatics* 34(15)**,** 2666–2669. doi: https://doi.org/10.1093/bioinformatics/bty149.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S.V., et al. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* 11(10)**,** R107. doi: 10.1186/gb-2010-11-10-r107.

Douglas, C.W.I., Heath, J., Hampton, K.K., and Preston, F.E. (1993). Identity of viridans streptococci isolated from cases of infective endocarditis. *Journal of Medical Microbiology* 39(3)**,** 179-182. doi: https://doi.org/10.1099/00222615-39-3-179.

Driscoll, C.B., Otten, T.G., Brown, N.M., and Dreher, T.W. (2017). Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in Genomic Sciences* 12(1)**,** 9. doi: https://doi.org/10.1186/s40793-017-0224-8.

Dutta, J. (2001). *Isolation and characterization of polycyclic aromatic hydrocarbon degrading bacteria from the rhizosphere of salt-marsh plants.* MSc MSc thesis, Rutgers, the State University of New Jersey.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research* 21(12)**,** 2224-2241. doi: https://doi.org/10.1101/gr.126599.111.

Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., et al. (1990). Automated DNA sequencing of the human HPRT locus. *Genomics* 6(4)**,** 593-608. doi: https://doi.org/10.1016/0888-7543(90)90493-E.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910)**,** 133-138. doi: https://doi.org/10.1126/science.1162986.

Escolano-Martínez, M.S., Domenech, A., Yuste, J., Cercenado, M.I., Ardanuy, C., Liñares, J., et al. (2016). DiiA is a novel dimorphic cell wall protein of *Streptococcus pneumoniae* involved in invasive disease. *Journal of Infection* 73(1)**,** 71-81. doi: https://doi.org/10.1016/j.jinf.2016.04.010.

Federhen, S. (2015). Type material in the NCBI Taxonomy Database. *Nucleic Acids Research* 43(D1)**,** D1086-D1098. doi: https://doi.org/10.1093/nar/gku1127.

Federhen, S., Rossello-Mora, R., Klenk, H.-P., Tindall, B.J., Konstantinidis, K.T., Whitman, W.B., et al. (2016). Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015). *Standards in Genomic Sciences* 11(1)**,** 15. doi: https://doi.org/10.1186/s40793-016-0134-1.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* 34(3)**,** e22-e22. doi: https://doi.org/10.1093/nar/gnj023.

Fernández-Juárez, V., Jaén-Luchoro, D., Brito-Echeverría, J., Agawin, N.S.R., Bennasar-Figueras, A., and Echeveste, P. (2021). Everything is everywhere: physiological responses of the Mediterranean sea and Eastern Pacific ocean epiphyte *Cobetia* sp. to varying nutrient concentration. *Microbial Ecology*. doi: https://doi.org/10.1007/s00248-021-01766-z.

Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., Savic, G., Lyon, K., et al. (2001). Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proceedings of the National Academy of Sciences* 98(8)**,** 4658-4663. doi: https://doi.org/10.1073/pnas.071559398.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223)**,** 496-512. doi: https://doi.org/10.1126/science.7542800.

Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* 7(6)**,** 461-465. doi: https://doi.org/10.1038/nmeth.1459.

Fox, G.E., Wisotzkey, J.D., and Jurtshuk, P. (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *International Journal of Systematic and Evolutionary Microbiology* 42(1)**,** 166-170. doi: https://doi.org/10.1099/00207713-42-1-166.

Fraser, C.M., Eisen, J.A., Nelson, K.E., Paulsen, I.T., and Salzberg, S.L. (2002). The value of complete microbial genome sequencing (you get what you pay for). *Journal of Bacteriology* 184(23)**,** 6403-6405. doi: https://doi.org/10.1128/jb.184.23.6403-6405.2002.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235)**,** 397-404. doi: https://doi.org/10.1126/science.270.5235.397.

Freschi, L., Vincent, A.T., Jeukens, J., Emond-Rheault, J.-G., Kukavica-Ibrulj, I., Dupont, M.-J., et al. (2019). The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biology and Evolution* 11(1)**,** 109-120. doi: https://doi.org/10.1093/gbe/evy259.

Fuursted, K., Littauer, P.J., Greve, T., and Scholz, C.F.P. (2016). Septicemia with *Streptococcus pseudopneumoniae*: report of three cases with an apparent hepatic or bile duct association. *Infectious Diseases* 48(8)**,** 636-639. doi: https://doi.org/10.3109/23744235.2016.1157896.

Ganaie, F., Branche, A.R., Peasley, M., Rosch, J.W., and Nahm, M.H. (2021). Oral streptococci expressing pneumococci-like cross-reactive capsule types can affect WHO recommended pneumococcal carriage procedure. *Clinical Infectious Diseases*. doi: https://doi.org/10.1093/cid/ciab1003.

Ganaie, F., Saad, J.S., McGee, L., van Tonder, A.J., Bentley, S.D., Lo, S.W., et al. (2020). A new pneumococcal capsule type, 10D, is the 100th serotype and has a large *cps* fragment from an oral streptococcus. *mBio* 11(3)**,** e00937-00920. doi: https://doi.org/10.1128/mBio.00937-20.

Garriss, G., Nannapaneni, P., Simões, A.S., Browall, S., Subramanian, K., Sá-Leão, R., et al. (2019). Genomic characterization of the emerging pathogen *Streptococcus pseudopneumoniae*. *mBio* 10(3)**,** e01286-01219. doi: https://doi.org/10.1128/mBio.01286-19.

Giefing , C., Meinke , A.L., Hanner , M., Henics , T.s., Minh , D.B., Gelbmann , D., et al. (2008). Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. *Journal of Experimental Medicine* 205(1)**,** 117-131. doi: https://doi.org/10.1084/jem.20071168.

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

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274(5287)**,** 546-567. doi: https://doi.org/10.1126/science.274.5287.546.

Gomila, M., Busquets, A., Mulet, M., García-Valdés, E., and Lalucat, J. (2017). Clarification of taxonomic status within the *Pseudomonas syringae* species group based on a phylogenomic analysis. *Frontiers in Microbiology* 8(2422). doi: https://doi.org/10.3389/fmicb.2017.02422.

Gomila, M., Mulet, M., García-Valdés, E., and Lalucat, J. (2022). Genome-based taxonomy of the genus *Stutzerimonas* and proposal of *S. frequens* sp. nov. and *S. degradans* sp. nov. and emended descriptions of *S. perfectomarina* and *S. chloritidismutans*. *Microorganisms* 10(7)**,** 1363. doi: https://doi.org/10.3390/microorganisms10071363.

Gomila, M., Peña, A., Mulet, M., Lalucat, J., and García-Valdés, E. (2015). Phylogenomics and systematics in *Pseudomonas*. *Frontiers in Microbiology* 6**,** 214. doi: https://doi.org/10.3389/fmicb.2015.00214.

Gonzales-Siles, L., Karlsson, R., Schmidt, P., Salvà-Serra, F., Jaén-Luchoro, D., Skovbjerg, S., et al. (2020). A pangenome approach for discerning species-unique gene markers for identifications of *Streptococcus pneumoniae* and *Streptococcus pseudopneumoniae*. *Frontiers in Cellular and Infection Microbiology* 10(222). doi: https://doi.org/10.3389/fcimb.2020.00222.

Gonzales-Siles, L., Salvà-Serra, F., Degerman, A., Nordén, R., Lindh, M., Skovbjerg, S., et al. (2019). Identification and capsular serotype sequetyping of *Streptococcus pneumoniae* strains. *Journal of Medical Microbiology* 68(8)**,** 1173-1188. doi: https://doi.org/10.1099/jmm.0.001022.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17(6)**,** 333-351. doi: https://doi.org/10.1038/nrg.2016.49.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57(Pt 1)**,** 81-91. doi: https://doi.org/10.1099/ijs.0.64483-0.

Grevskott, D.H., Salvà-Serra, F., Moore, E.R.B., and Marathe, N.P. (2020). Nanopore sequencing reveals genomic map of CTX-M-type extended-spectrum β-lactamases carried by *Escherichia coli* strains isolated from blue mussels (*Mytilus edulis*) in

Norway. *BMC Microbiology* 20(1), 134. doi: https://doi.org/10.1186/s12866-020-01821-8.
Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8), 1072-1075. doi: https://doi.org/10.1093/bioinformatics/btt086.
Gutleben, J., Chaib De Mares, M., van Elsas, J.D., Smidt, H., Overmann, J., and Sipkema, D. (2018). The multi-omics promise in context: from sequence to microbial isolate. *Critical Reviews in Microbiology* 44(2), 212-229. doi: https://doi.org/10.1080/1040841X.2017.1332003.
Ha, S.M., Kim, C.K., Roh, J., Byun, J.H., Yang, S.J., Choi, S.B., et al. (2019). Application of the whole genome-based bacterial identification system, TrueBac ID, using clinical isolates that were not identified with three matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) systems. *Annals of Laboratory Medicine* 39(6), 530-536. doi: https://doi.org/10.3343/alm.2019.39.6.530.
Harrison, P.W., Ahamed, A., Aslam, R., Alako, B.T.F., Burgin, J., Buso, N., et al. (2021). The European Nucleotide Archive in 2020. *Nucleic Acids Research* 49(D1), D82-D85. doi: https://doi.org/10.1093/nar/gkaa1028.
Heather, J.M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* 107(1), 1-8. doi: https://doi.org/10.1016/j.ygeno.2015.11.003.
Hedlund, B.P., Chuvochina, M., Hugenholtz, P., Konstantinidis, K.T., Murray, A.E., Palmer, M., et al. (2022). SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nature Microbiology* 7(10), 1702-1708. doi: https://doi.org/10.1038/s41564-022-01214-9.
Hicks, L.A., Harrison, L.H., Flannery, B., Hadler, J.L., Schaffner, W., Craig, A.S., et al. (2007). Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998–2004. *The Journal of Infectious Diseases* 196(9), 1346-1354. doi: https://doi.org/10.1086/521626.
Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Research* 24(22), 4420-4449. doi: https://doi.org/10.1093/nar/24.22.4420.
Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biology* 14(5), R47. doi: https://doi.org/10.1186/gb-2013-14-5-r47.
Idury, R.M., and Waterman, M.S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology* 2(2), 291-306. doi: https://doi.org/10.1089/cmb.1995.2.291.
Ikuta, K.S., Swetschinski, L.R., Robles Aguilar, G., Sharara, F., Mestrovic, T., Gray, A.P., et al. (2022). Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 400(10369), 2221-2248. doi: https://doi.org/10.1016/S0140-6736(22)02185-7.
Imai, K., Nemoto, R., Kodana, M., Tarumoto, N., Sakai, J., Kawamura, T., et al. (2020). Rapid and accurate species identification of mitis group streptococci using the MinION Nanopore sequencer. *Frontiers in Cellular and Infection Microbiology* 10(11). doi: https://doi.org/10.3389/fcimb.2020.00011.
International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409(6822), 860-921. doi: https://doi.org/10.1038/35057062.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431(7011)**,** 931-945. doi: https://doi.org/10.1038/nature03001.

Iredell, J., Brown, J., and Tagg, K. (2016). Antibiotic resistance in *Enterobacteriaceae*: mechanisms and clinical implications. *BMJ* 352**,** h6420. doi: https://doi.org/10.1136/bmj.h6420.

Jaén-Luchoro, D., Busquets, A., Karlsson, R., Salvà-Serra, F., Åhrén, C., Karami, N., et al. (2020). Genomic and proteomic characterization of the extended-spectrum β-lactamase (ESBL)-producing *Escherichia coli* strain CCUG 73778: a virulent, nosocomial outbreak strain. *Microorganisms* 8**,** 893. doi: https://doi.org/10.3390/microorganisms8060893.

Jaén-Luchoro, D., Gonzales-Siles, L., Karlsson, R., Svensson-Stadler, L., Molin, K., Cardew, S., et al. (2019). *Corynebacterium sanguinis* sp. nov., a clinical and environmental associated corynebacterium. *Systematic and Applied Microbiology***,** 126039. doi: https://doi.org/10.1016/j.syapm.2019.126039.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 9(1)**,** 5114. doi: https://doi.org/10.1038/s41467-018-07641-9.

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

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

Jensen, C.S., Dargis, R., Shewmaker, P., Nielsen, X.C., and Christensen, J.J. (2021a). Identification of *Streptococcus pseudopneumoniae* and other mitis group streptococci using matrix assisted laser desorption/ionization - time of flight mass spectrometry. *Diagnostic Microbiology and Infectious Disease* 101(3)**,** 115487. doi: https://doi.org/10.1016/j.diagmicrobio.2021.115487.

Jensen, C.S., Iversen, K.H., Dargis, R., Shewmaker, P., Rasmussen, S., Christensen, J.J., et al. (2021b). *Streptococcus pseudopneumoniae*: use of whole-genome sequences to validate species identification methods. *Journal of Clinical Microbiology* 59(2)**,** e02503-02520. doi: https://doi.org/10.1128/jcm.02503-20.

Johnston, C., Hinds, J., Smith, A., van der Linden, M., Van Eldere, J., and Mitchell, T.J. (2010). Detection of large numbers of pneumococcal virulence genes in streptococci of the Mitis group. *Journal of Clinical Microbiology* 48(8)**,** 2762-2769. doi: https://doi.org/10.1128/JCM.01746-09.

Joshi, N.A., and Fass, J.N. (2011). "Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files". 1.33 ed.).

Kaas, R.S., Leekitcharoenphon, P., Aarestrup, F.M., and Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLOS ONE* 9(8)**,** e104984. doi: https://doi.org/10.1371/journal.pone.0104984.

Kannan, S., Sharma, S., Ciufo, S., Clark, K., Turner, S., Kitts, P.A., et al. (2023). Collection and curation of prokaryotic genome assemblies from type strains at NCBI. *International*

Journal of Systematic and Evolutionary Microbiology 73(1). doi: https://doi.org/10.1099/ijsem.0.005707.

Karlsson, R., Gonzales-Siles, L., Boulund, F., Svensson-Stadler, L., Skovbjerg, S., Karlsson, A., et al. (2015). Proteotyping: Proteomic characterization, classification and identification of microorganisms--A prospectus. *Systematic and Applied Microbiology* 38(4)**,** 246-257. doi: https://doi.org/10.1016/j.syapm.2015.03.006.

Karlsson, R., Gonzales-Siles, L., Gomila, M., Busquets, A., Salvà-Serra, F., Jaén-Luchoro, D., et al. (2018). Proteotyping bacteria: Characterization, differentiation and identification of pneumococcus and other species within the Mitis Group of the genus *Streptococcus* by tandem mass spectrometry proteomics. *PLOS ONE* 13(12)**,** e0208804. doi: https://doi.org/10.1371/journal.pone.0208804.

Karlsson, R., Thorsell, A., Gomila, M., Salvà-Serra, F., Jakobsson, H.E., Gonzales-Siles, L., et al. (2020). Discovery of species-unique peptide biomarkers of bacterial pathogens by tandem mass spectrometry-based proteotyping. *Molecular & Cellular Proteomics***,** mcp.RA119.001667. doi: https://doi.org/10.1074/mcp.RA119.001667.

Kawamura, Y., Hou, X.-G., Sultana, F., Miura, H., and Ezaki, T. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *International Journal of Systematic and Evolutionary Microbiology* 45(2)**,** 406-408. doi: https://doi.org/10.1099/00207713-45-2-406.

Keith, E.R., Podmore, R.G., Anderson, T.P., and Murdoch, D.R. (2006). Characteristics of *Streptococcus pseudopneumoniae* isolated from purulent sputum samples. *Journal of Clinical Microbiology* 44(3)**,** 923-927. doi: https://doi.org/10.1128/JCM.44.3.923-927.2006.

Kettler, H., White, K., Hawkes, S.J., and UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (2004). "Mapping the landscape of diagnostics for sexually transmitted infections: key findings and recommendations". World Health Organization).

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

Kilian, M., and Tettelin, H. (2019). Identification of virulence-associated properties by comparative genome analysis of *Streptococcus pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, three *S. oralis* subspecies, and *S. infantis*. *mBio* 10(5)**,** e01985-01919. doi: https://doi.org/10.1128/mBio.01985-19.

Kim, D., Park, S., and Chun, J. (2021). Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *Journal of Microbiology* 59(5)**,** 476-480. doi: https://doi.org/10.1007/s12275-021-1154-0.

Kingsford, C., Schatz, M.C., and Pop, M. (2010). Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics* 11(1)**,** 21. doi: https://doi.org/10.1186/1471-2105-11-21.

Klein, E. (1884). Micro-Organisms and Disease. *Practitioner* 32**,** 321-352.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37(5)**,** 540-546. doi: https://doi.org/10.1038/s41587-019-0072-8.

Konstantinidis, K.T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *The ISME Journal* 11(11)**,** 2399-2406. doi: https://doi.org/10.1038/ismej.2017.113.

Konstantinidis, K.T., and Tiedje, J.M. (2005a). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* 102(7)**,** 2567-2572. doi: https://doi.org/10.1073/pnas.0409727102.

Konstantinidis, K.T., and Tiedje, J.M. (2005b). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187(18)**,** 6258-6264. doi: 10.1128/JB.187.18.6258-6264.2005.

Konstantinidis, K.T., and Tiedje, J.M. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* 10(5)**,** 504-509. doi: https://doi.org/10.1016/j.mib.2007.08.006.

Koonin, E.V., Makarova, K.S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* 55(1)**,** 709-742. doi: https://doi.org/10.1146/annurev.micro.55.1.709.

Koren, S., Harhay, G.P., Smith, T.P.L., Bono, J.L., Harhay, D.M., McVey, S.D., et al. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* 14(9)**,** R101. doi: https://doi.org/10.1186/gb-2013-14-9-r101.

Koren, S., and Phillippy, A.M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* 23**,** 110-120. doi: https://doi.org/10.1016/j.mib.2014.11.014.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Research* 27(5)**,** 722-736. doi: https://doi.org/10.1101/gr.215087.116.

Kyrpides, N.C., Hugenholtz, P., Eisen, J.A., Woyke, T., Göker, M., Parker, C.T., et al. (2014). Genomic Encyclopedia of Bacteria and Archaea: sequencing a myriad of type strains. *PLOS Biology* 12(8)**,** e1001920. doi: https://doi.org/10.1371/journal.pbio.1001920.

Lalucat, J., Bennasar, A., Bosch, R., García-Valdés, E., and Palleroni, N.J. (2006). Biology of *Pseudomonas stutzeri*. *Microbiology and Molecular Biology Reviews* 70(2)**,** 510-547. doi: https://doi.org/10.1128/MMBR.00047-05.

Lalucat, J., Gomila, M., Mulet, M., Zaruma, A., and García-Valdés, E. (2022). Past, present and future of the boundaries of the *Pseudomonas* genus: proposal of *Stutzerimonas* gen. nov. *Systematic and Applied Microbiology***,** 126289. doi: https://doi.org/10.1016/j.syapm.2021.126289.

Lalucat, J., Mulet, M., Gomila, M., and García-Valdés, E. (2020). Genomics in bacterial taxonomy: impact on the genus *Pseudomonas*. *Genes* 11(2)**,** 139. doi: https://doi.org/10.3390/genes11020139.

Lancefield, R.C. (1933). A serological differentiation of human and other groups of hemolytic streptococci. *Journal of Experimental Medicine* 57(4)**,** 571-595. doi: https://doi.org/10.1084/jem.57.4.571.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science* 299(5607)**,** 682-686. doi: https://doi.org/10.1126/science.1079700.

Lewis, W.H., Tahon, G., Geesink, P., Sousa, D.Z., and Ettema, T.J.G. (2020). Innovations to culturing the uncultured microbial majority. *Nature Reviews Microbiology*. doi: https://doi.org/10.1038/s41579-020-00458-8.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32(14)**,** 2103-2110. doi: https://doi.org/10.1093/bioinformatics/btw152.

Li, N., Cai, Q., Miao, Q., Song, Z., Fang, Y., and Hu, B. (2021a). High-throughput metagenomics for identification of pathogens in the clinical settings. *Small Methods* 5(1)**,** 2000792. doi: https://doi.org/10.1002/smtd.202000792.

Li, X., Yang, Z., Wang, Z., Li, W., Zhang, G., and Yan, H. (2022). Comparative genomics of *Pseudomonas stutzeri* complex: taxonomic assignments and genetic diversity. *Frontiers in Microbiology* 12. doi: https://doi.org/10.3389/fmicb.2021.755874.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics* 11(1)**,** 25-37. doi: https://doi.org/10.1093/bfgp/elr035.

Li, Z., Ding, X., Yin, K., Xu, Z., Cooper, K., and Liu, C. (2021b). Electric field-enhanced electrochemical CRISPR biosensor for DNA detection. *Biosensors and Bioelectronics* 192**,** 113498. doi: https://doi.org/10.1016/j.bios.2021.113498.

Liao, X., Li, M., Zou, Y., Wu, F.-X., Yi, P., and Wang, J. (2019). Current challenges and solutions of *de novo* assembly. *Quantitative Biology* 7(2)**,** 90-109. doi: https://doi.org/10.1007/s40484-019-0166-9.

Llull, D., López, R., and García, E. (2006). Characteristic signatures of the *lytA* gene provide a basis for rapid and reliable diagnosis of *Streptococcus pneumoniae* infections. *Journal of Clinical Microbiology* 44(4)**,** 1250-1256. doi: https://doi.org/10.1128/JCM.44.4.1250-1256.2006.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., et al. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30(5)**,** 434-439. doi: https://doi.org/10.1038/nbt.2198.

Loman, N.J., and Pallen, M.J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* 13(12)**,** 787-794. doi: https://doi.org/10.1038/nrmicro3565.

Ludwig, W., Viver, T., Westram, R., Gago, J.F., Bustos-Caparros, E., Knittel, K., et al. (2021). Release LTP_12_2020, featuring a new ARB alignment and improved 16S rRNA tree for prokaryotic type strains. *Systematic and Applied Microbiology* 44(4)**,** 126218. doi: https://doi.org/10.1016/j.syapm.2021.126218.

Maddock, D., Kile, H., Denman, S., Arnold, D., and Brady, C. (2022). Description of three novel species of *Scandinavium*: *Scandinavium hiltneri* sp. nov., *Scandinavium manionii* sp. nov. and *Scandinavium tedordense* sp. nov., isolated from the oak rhizosphere and bleeding cankers of broadleaf hosts. *Frontiers in Microbiology* 13. doi: https://doi.org/10.3389/fmicb.2022.1011653.

Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., et al. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29(14)**,** 1718-1725. doi: https://doi.org/10.1093/bioinformatics/btt273.

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., et al. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* 95(6)**,** 3140-3145. doi: https://doi.org/10.1073/pnas.95.6.3140.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., et al. (2005). Identification of a universal Group B *Streptococcus* vaccine by multiple genome screen. *Science* 309(5731)**,** 148-150. doi: https://doi.org/10.1126/science.1109869.

Marathe, N.P., Salvà-Serra, F., Nimje, P.S., and Moore, E.R.B. (2022). Novel plasmid carrying mobile colistin resistance gene *mcr-4.3* and mercury resistance genes in *Shewanella baltica*: insights into mobilization of *mcr-4.3* in *Shewanella* species. *Microbiology Spectrum* 10(6)**,** e02037-02022. doi: https://doi.org/10.1128/spectrum.02037-22.

Marcy, Y., Ouverney, C., Bik, E.M., Lösekann, T., Ivanova, N., Martin, H.G., et al. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National*

Academy of Sciences 104(29)**,** 11889-11894. doi: https://doi.org/10.1073/pnas.0704662104.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057)**,** 376-380. doi: https://doi.org/10.1038/nature03959.
Marmur, J. (1961). A procedure for the isolation of deoxyribonucleic acid from micro-organisms. *Journal of Molecular Biology* 3(2)**,** 208-IN201. doi: https://doi.org/10.1016/S0022-2836(61)80047-8.
Martín-Galiano, A.J., Escolano-Martínez, M.S., Corsini, B., de la Campa, A.G., and Yuste, J. (2021). Immunization with SP_1992 (DiiA) protein of *Streptococcus pneumoniae* reduces nasopharyngeal colonization and protects against invasive disease in mice. *Vaccines* 9(3). doi: https://doi.org/10.3390/vaccines9030187.
Masomian, M., Ahmad, Z., Gew, L.T., and Poh, C.L. (2020). Development of next generation *Streptococcus pneumoniae* vaccines conferring broad protection. *Vaccines (Basel)* 8(1). doi: https://doi.org/10.3390/vaccines8010132.
Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* 74(2)**,** 560-564. doi: https://doi.org/10.1073/pnas.74.2.560.
Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., et al. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* 39**,** W339-346. doi: https://doi.org/10.1093/nar/gkr466.
Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development* 15(6)**,** 589-594. doi: https://doi.org/10.1016/j.gde.2005.09.006.
Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14(1)**,** 60. doi: https://doi.org/10.1186/1471-2105-14-60.
Meier-Kolthoff, J.P., Carbasse, J.S., Peinado-Olarte, R.L., and Göker, M. (2022). TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Research* 50(D1)**,** D801-D807. doi: https://doi.org/10.1093/nar/gkab902.
Meier-Kolthoff, J.P., and Göker, M. (2019). TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nature Communications* 10(1)**,** 2182. doi: https://doi.org/10.1038/s41467-019-10210-3.
Méndez, V., Hernández, L., Salvà-Serra, F., Jaén-Luchoro, D., Durán, R.E., Barra, B., et al. (2018). Complete genome sequence of the hydrocarbon-degrading strain *Achromobacter* sp. B7, isolated during petroleum hydrocarbon bioremediation in the Valparaiso region, Chile. *Microbiology Resource Announcements* 7(19). doi: https://doi.org/10.1128/MRA.01326-18.
Méndez, V., Valenzuela, M., Salvà-Serra, F., Jaén-Luchoro, D., Besoain, X., Moore, E.R., et al. (2020). Comparative genomics of pathogenic *Clavibacter michiganensis* subsp. *michiganensis* strains from Chile reveals potential virulence features for tomato plants. *Microorganisms* 8(11)**,** 1679. doi: https://doi.org/10.3390/microorganisms8111679.
Messmer, T.O., Sampson, J.S., Stinson, A., Wong, B., Carlone, G.M., and Facklam, R.R. (2004). Comparison of four polymerase chain reaction assays for specificity in the identification of *Streptococcus pneumoniae*. *Diagnostic Microbiology and Infectious Disease* 49(4)**,** 249-254. doi: https://doi.org/10.1016/j.diagmicrobio.2004.04.013.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11(1)**,** 31-46. doi: https://doi.org/10.1038/nrg2626.

Meziti, A., Rodriguez-R, L.M., Hatt, J.K., Peña-Gonzalez, A., Levy, K., Konstantinidis, K.T., et al. (2021). The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Applied and Environmental Microbiology* 87(6)**,** e02593-02520. doi: https://doi.org/10.1128/AEM.02593-20.

Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., et al. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24(24)**,** 2818-2824. doi: https://doi.org/10.1093/bioinformatics/btn548.

Mira, A., Martín-Cuadrado, A.B., D'Auria, G., and Rodríguez-Valera, F. (2010). The bacterial pan-genome: a new paradigm in microbiology. *International Microbiology* 13(2)**,** 45-57. doi: https://doi.org/10.2436/20.1501.01.110.

Moss, E.L., Maghini, D.G., and Bhatt, A.S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature Biotechnology* 38(6)**,** 701-707. doi: https://doi.org/10.1038/s41587-020-0422-6.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N.C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences* 10(1)**,** 18. doi: https://doi.org/10.1186/1944-3277-10-18.

Mukherjee, S., Seshadri, R., Varghese, N.J., Eloe-Fadrosh, E.A., Meier-Kolthoff, J.P., Göker, M., et al. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology* 35(7)**,** 676-683. doi: https://doi.org/10.1038/nbt.3886.

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, Jagadish C., Lee, J., et al. (2021). Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research* 49(D1)**,** D723-D733. doi: https://doi.org/10.1093/nar/gkaa983.

Mulet, M., Gomila, M., Gruffaz, C., Meyer, J.M., Palleroni, N.J., Lalucat, J., et al. (2008). Phylogenetic analysis and siderotyping as useful tools in the taxonomy of *Pseudomonas stutzeri*: description of a novel genomovar. *International Journal of Systematic and Evolutionary Microbiology* 58(Pt 10)**,** 2309-2315. doi: https://doi.org/10.1099/ijs.0.65797-0.

Murray, R.G.E., and Schleifer, K.H. (1994). Taxonomic notes: a proposal for recording the properties of putative taxa of procaryotes. *International Journal of Systematic and Evolutionary Microbiology* 44(1)**,** 174-176. doi: https://doi.org/10.1099/00207713-44-1-174.

Murray, R.G.E., and Stackebrandt, E. (1995). Taxonomic note: implementation of the provisional status *Candidatus* for incompletely described procaryotes. *International Journal of Systematic and Evolutionary Microbiology* 45(1)**,** 186-187. doi: https://doi.org/10.1099/00207713-45-1-186.

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13 Suppl 14**,** S8. doi: https://doi.org/10.1186/1471-2105-13-S14-S8.

Nakano, M., Komatsu, J., Matsuura, S.-i., Takashima, K., Katsura, S., and Mizuno, A. (2003). Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology* 102(2)**,** 117-124. doi: https://doi.org/10.1016/S0168-1656(03)00023-3.

Narzisi, G., and Mishra, B. (2011). Comparing de novo genome assembly: the long and short of it. *PLOS ONE* 6(4)**,** e19175. doi: https://doi.org/10.1371/journal.pone.0019175.

Nature Methods (2023). Method of the Year 2022: long-read sequencing. *Nature Methods* 20(1)**,** 1-1. doi: https://doi.org/10.1038/s41592-022-01759-x.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology* 39**,** 499-509. doi: https://doi.org/10.1038/s41587-020-0718-6.

Nouws, S., Bogaerts, B., Verhaegen, B., Denayer, S., Piérard, D., Marchal, K., et al. (2020). Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. *Scientific Reports* 10(1)**,** 14649. doi: https://doi.org/10.1038/s41598-020-71207-3.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44(D1)**,** D733-745. doi: https://doi.org/10.1093/nar/gkv1189.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784)**,** 299-304. doi: https://doi.org/10.1038/35012500.

Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T., and Fujisawa, T. (2020). DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Research* 48(D1)**,** D45-D50. doi: https://doi.org/10.1093/nar/gkz982.

Omasits, U., Varadarajan, A.R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., et al. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Research* 27(12)**,** 2083-2095. doi: https://doi.org/10.1101/gr.218255.116.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22)**,** 3691-3693. doi: https://doi.org/10.1093/bioinformatics/btv421.

Pain, M., Wolden, R., Jaén-Luchoro, D., Salvà-Serra, F., Iglesias, B.P., Karlsson, R., et al. (2020). *Staphylococcus borealis* sp. nov., isolated from human skin and blood. *International Journal of Systematic and Evolutionary Microbiology*. doi: https://doi.org/10.1099/ijsem.0.004499.

Palleroni, N.J. (2015). "*Pseudomonas*," in *Bergey's Manual of Systematics of Archaea and Bacteria*, eds. M.E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F.A. Rainey & W.B. Whitman.), 1-1.

Palleroni, N.J., Doudoroff, M., Stanier, R.Y., Solánes, R.E., and Mandel, M. (1970). Taxonomy of the aerobic Pseudomonads: the properties of the *Pseudomonas stutzeri* group. *Microbiology* 60(2)**,** 215-231. doi: https://doi.org/10.1099/00221287-60-2-215.

Parker, C.T., Tindall, B.J., and Garrity, G.M. (2019). International Code of Nomenclature of Prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 69(1A)**,** S1-S111. doi: https://doi.org/10.1099/ijsem.0.000778.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* 38(9)**,** 1079-1086. doi: https://doi.org/10.1038/s41587-020-0501-8.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* 36(10)**,** 996-1004. doi: https://doi.org/10.1038/nbt.4229.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25(7)**,** 1043-1055. doi: https://doi.org/10.1101/gr.186072.114.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 2(11)**,** 1533-1542. doi: https://doi.org/10.1038/s41564-017-0012-7.

Parte, A.C. (2014). LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Research* 42**,** D613-616. doi: https://doi.org/10.1093/nar/gkt1111.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176(3)**,** 649-662.e620. doi: https://doi.org/10.1016/j.cell.2019.01.001.

Pasquali, F., Do Valle, I., Palma, F., Remondini, D., Manfreda, G., Castellani, G., et al. (2019). Application of different DNA extraction procedures, library preparation protocols and sequencing platforms: impact on sequencing results. *Heliyon* 5(10)**,** e02745. doi: https://doi.org/10.1016/j.heliyon.2019.e02745.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2018). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* 35(13)**,** 2193-2198. doi: https://doi.org/10.1093/bioinformatics/bty841.

Pennisi, E. (2008). Proposal to 'Wikify' GenBank meets stiff resistance. *Science* 319(5870)**,** 1598-1599. doi: https://doi.org/10.1126/science.319.5870.1598.

Pimenta, F., Moiane, B., Gertz, R.E., Chochua, S., Snippes Vagnone, P.M., Lynfield, R., et al. (2021). New pneumococcal serotype 15D. *Journal of Clinical Microbiology* 59(5)**,** e00329-00321. doi: https://doi.org/10.1128/jcm.00329-21.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Aricò, B., Comanducci, M., et al. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287(5459)**,** 1816-1820. doi: https://doi.org/10.1126/science.287.5459.1816.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics* 10(4)**,** 354-366. doi: https://doi.org/10.1093/bib/bbp026.

Prober, J., Trainor, G., Dam, R., Hobbs, F., Robertson, C., Zagursky, R., et al. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238(4825)**,** 336-341. doi: https://doi.org/10.1126/science.2443975.

Puyet, A., Greenberg, B., and Lacks, S.A. (1990). Genetic and structural characterization of *endA*: A membrane-bound nuclease required for transformation of *Streptococcus pneumoniae*. *Journal of Molecular Biology* 213(4)**,** 727-738. doi: https://doi.org/10.1016/S0022-2836(05)80259-1.

Qin, Q.-L., Xie, B.-B., Zhang, X.-Y., Chen, X.-L., Zhou, B.-C., Zhou, J., et al. (2014). A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology* 196(12)**,** 2210-2215. doi: https://doi.org/10.1128/jb.01688-14.

Quick, J. (2018). Ultra-long read sequencing protocol for RAD004. *protocols.io*. doi: https://dx.doi.org/10.17504/protocols.io.mrxc57n.

Rahn, O. (1937). New principles for the classification of bacteria. *Zentralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheiten und Hygiene* 96**,** 273-286.

Ralph, A.P., and Carapetis, J.R. (2013). Group a streptococcal diseases and their global burden. *Current Topics in Microbiology and Immunology* 368**,** 1-27. doi: https://doi.org/10.1007/82_2012_280.

Rasmussen, L.H., Dargis, R., Christensen, J.J., Skovgaard, O., and Nielsen, X.C. (2016). Draft genome sequence of type strain *Streptococcus gordonii* ATCC 10558. *Genome Announcements* 4(1)**,** e01745-01715. doi: https://doi.org/10.1128/genomeA.01745-15.

Read, T.D., Salzberg, S.L., Pop, M., Shumway, M., Umayam, L., Jiang, L., et al. (2002). Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296(5575)**,** 2028-2033. doi: https://doi.org/10.1126/science.1071837.

Richter, M., and Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences* 106(45)**,** 19126-19131. doi: https://doi.org/10.1073/pnas.0906412106.

Richter, M., Rosselló-Móra, R., Oliver Glöckner, F., and Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32(6)**,** 929-931. doi: https://doi.org/10.1093/bioinformatics/btv681.

Richter, S.S., Heilmann, K.P., Dohrn, C.L., Riahi, F., Beekmann, S.E., and Doern, G.V. (2008). Accuracy of phenotypic methods for identification of *Streptococcus pneumoniae* isolates included in surveillance programs. *Journal of Clinical Microbiology* 46(7)**,** 2184-2188. doi: https://doi.org/10.1128/jcm.00461-08.

Ricker, N., Qian, H., and Fulthorpe, R.R. (2012). The limitations of draft assemblies for understanding prokaryotic adaptation and evolution. *Genomics* 100(3)**,** 167-175. doi: https://doi.org/10.1016/j.ygeno.2012.06.009.

Rodriguez-R, L., and Konstantinidis, K. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*. doi: https://doi.org/10.7287/peerj.preprints.1900v1.

Rodriguez, R.L., Gunturu, S., Harvey, W.T., Rosselló-Mora, R., Tiedje, J.M., Cole, J.R., et al. (2018). The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of *Archaea* and *Bacteria* at the whole genome level. *Nucleic Acids Research* 46(W1)**,** W282-W288. doi: https://doi.org/10.1093/nar/gky467.

Rolo, D., S. Simões, A., Domenech, A., Fenoll, A., Liñares, J., de Lencastre, H., et al. (2013). Disease isolates of *Streptococcus pseudopneumoniae* and non-typeable *S. pneumoniae* presumptively identified as atypical *S. pneumoniae* in Spain. *PLOS ONE* 8(2)**,** e57047. doi: https://doi.org/10.1371/journal.pone.0057047.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* 242(1)**,** 84-89. doi: https://doi.org/10.1006/abio.1996.0432.

Rosenbach, F.J. (1884). *Mikro-organismen bei den Wund-Infections-Krankheiten des Menschen.* Wiesbaden: J.F. Bergmann.

Rosselló-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiology Reviews* 25(1)**,** 39-67. doi: https://doi.org/10.1111/j.1574-6976.2001.tb00571.x.

Rosselló-Mora, R.A., Lalucat, J., Dott, W., and Kämpfer, P. (1994). Biochemical and chemotaxonomic characterization of *Pseudomonas stutzeri* genomovars. *Journal of Applied Bacteriology* 76(3)**,** 226-233. doi: https://doi.org/10.1111/j.1365-2672.1994.tb01620.x.

Rossello, R., Garcia-Valdes, E., Lalucat, J., and Ursing, J. (1991). Genotypic and phenotypic diversity of *Pseudomonas stutzeri*. *Systematic and Applied Microbiology* 14(2)**,** 150-157. doi: http://dx.doi.org/10.1016/S0723-2020(11)80294-8.

Rossen, J.W.A., Friedrich, A.W., Moran-Gilad, J., Genomic, E.S.G.f., and Molecular, D. (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clinical Microbiology and Infection* 24(4)**,** 355-360. doi: https://doi.org/10.1016/j.cmi.2017.11.001.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356)**,** 348-352. doi: https://doi.org/10.1038/nature10242.

Rouli, L., Merhej, V., Fournier, P.E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections* 7**,** 72-85. doi: https://doi.org/10.1016/j.nmni.2015.06.005.

Sabat, A.J., Hermelijn, S.M., Akkerboom, V., Juliana, A., Degener, J.E., Grundmann, H., et al. (2017). Complete-genome sequencing elucidates outbreak dynamics of CA-MRSA USA300 (ST8-*spa* t008) in an academic hospital of Paramaribo, Republic of Suriname. *Scientific Reports* 7**,** 41050. doi: https://doi.org/10.1038/srep41050.

Sadowy, E., Bojarska, A., Kuch, A., Skoczyńska, A., Jolley, K.A., Maiden, M.C.J., et al. (2020). Relationships among streptococci from the mitis group, misidentified as *Streptococcus pneumoniae*. *European Journal of Clinical Microbiology & Infectious Diseases* 39(10)**,** 1865-1878. doi: https://doi.org/10.1007/s10096-020-03916-6.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., et al. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239(4839)**,** 487-491. doi: https://doi.org/10.1126/science.2448875.

Salgar-Chaparro, S.J., Castillo-Villamizar, G., Poehlein, A., Daniel, R., Machuca, L.L., and Putonti, C. (2020). Complete genome sequence of *Pseudomonas balearica* strain EC28, an iron-Oxidizing bacterium isolated from corroded steel. *Microbiology Resource Announcements* 9(19)**,** e00275-00220. doi: https://doi.org/10.1128/MRA.00275-20.

Salipante, S.J., SenGupta, D.J., Cummings, L.A., Land, T.A., Hoogestraat, D.R., and Cookson, B.T. (2015). Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *Journal of Clinical Microbiology* 53(4)**,** 1072-1079. doi: https://doi.org/10.1128/JCM.03385-14.

Salvà-Serra, F., Cardew, S., Markopoulos, S.J., Inganäs, E., Ohlén, M., Jaén-Luchoro, D., et al. (2022). "The problem of discrepant type strains among culture collections", in: *European Culture Collections' Organization (ECCO) XL Conference.* (Braunschweig, Germany).

Salvà-Serra, F., Connolly, G., Moore, E.R.B., and Gonzales-Siles, L. (2018a). Detection of "Xisco" gene for identification of *Streptococcus pneumoniae* isolates. *Diagnostic Microbiology and Infectious Disease* 90(4)**,** 248-250. doi: https://doi.org/10.1016/j.diagmicrobio.2017.12.003.

Salvà-Serra, F., Donoso, R.A., Cho, K.H., Yoo, J.A., Lee, K., Yoon, S.-H., et al. (2021). Complete multipartite genome sequence of the *Cupriavidus basilensis* type strain, a 2,6-dichlorophenol-degrading bacterium. *Microbiology Resource Announcements* 10(19)**,** e00134-00121. doi: https://doi.org/10.1128/mra.00134-21.

Salvà-Serra, F., Jaén-Luchoro, D., Marathe, N.P., Adlerberth, I., Moore, E.R.B., and Karlsson, R. (2023). Responses of carbapenemase-producing and non-producing carbapenem-resistant *Pseudomonas aeruginosa* strains to meropenem revealed by quantitative tandem mass spectrometry proteomics. *Frontiers in Microbiology* 13. doi: https://doi.org/10.3389/fmicb.2022.1089140.

Salvà-Serra, F., Svensson-Stadler, L., Busquets, A., Jaén-Luchoro, D., Karlsson, R., R. B. Moore, E., et al. (2018b). A protocol for extraction and purification of high-quality and quantity bacterial DNA applicable for genome sequencing: a modified version of the Marmur procedure. *Protocol Exchange*. doi: https://doi.org/10.1038/protex.2018.084.

Salvà Serra, F. (2014). *Estrategias de ensamblaje y anotación del genoma de Pseudomonas balearica.* MSc. Master Thesis, University of the Balearic Islands.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22(3)**,** 557-567. doi: https://doi.org/10.1101/gr.131383.111.

Salzberg, S.L., and Yorke, J.A. (2005). Beware of mis-assembled genomes. *Bioinformatics* 21(24)**,** 4320-4321. doi: https://doi.org/10.1093/bioinformatics/bti769.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., et al. (1977a). Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 265(5596)**,** 687-695. doi: https://doi.org/10.1038/265687a0.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94(3)**,** 441-448. doi: https://doi.org/10.1016/0022-2836(75)90213-2.

Sanger, F., Coulson, A.R., Friedmann, T., Air, G.M., Barrell, B.G., Brown, N.L., et al. (1978). The nucleotide sequence of bacteriophage φX174. *Journal of Molecular Biology* 125(2)**,** 225-246. doi: https://doi.org/10.1016/0022-2836(78)90346-7.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology* 162(4)**,** 729-773. doi: https://doi.org/10.1016/0022-2836(82)90546-0.

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

Sanguinetti, L., Toti, S., Reguzzi, V., Bagnoli, F., and Donati, C. (2012). A novel computational method identifies intra- and inter-species recombination events in *Staphylococcus aureus* and *Streptococcus pneumoniae*. *PLOS Computational Biology* 8(9)**,** e1002668. doi: https://doi.org/10.1371/journal.pcbi.1002668.

Satzke, C., Turner, P., Virolainen-Julkunen, A., Adrian, P.V., Antonio, M., Hare, K.M., et al. (2013). Standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*: Updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. *Vaccine* 32(1)**,** 165-179. doi: https://doi.org/10.1016/j.vaccine.2013.08.062.

Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, Stephen T., et al. (2022). GenBank. *Nucleic Acids Research* 50(D1)**,** D161-D164. doi: https://doi.org/10.1093/nar/gkab1135.

Schäffer, A.A., Nawrocki, E.P., Choi, Y., Kitts, P.A., Karsch-Mizrachi, I., and McVeigh, R. (2017). VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* 34(5)**,** 755-759. doi: https://doi.org/10.1093/bioinformatics/btx669.

Schildkraut, C.L., Marmur, J., and Doty, P. (1961). The formation of hybrid DNA molecules and their use in studies of DNA homologies. *Journal of Molecular Biology* 3(5)**,** 595-IN516. doi: https://doi.org/10.1016/S0022-2836(61)80024-7.

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J.E., Remus-Emsermann, M.N.P., et al. (2018). Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Research***,** gky726-gky726. doi: https://doi.org/10.1093/nar/gky726.

Schopf, J.W. (1993). Microfossils of the Early Archean Apex Chert: new evidence of the antiquity of life. *Science* 260(5108)**,** 640-646. doi: https://doi.org/10.1126/science.260.5108.640.

Scotta, C., Gomila, M., Mulet, M., Lalucat, J., and García-Valdés, E. (2013). Whole-cell MALDI-TOF mass spectrometry and multilocus sequence analysis in the discrimination

of *Pseudomonas stutzeri* populations: three novel genomovars. *Microbial Ecology* 66(3)**,** 522-532. doi: https://doi.org/10.1007/s00248-013-0246-8.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14)**,** 2068-2069. doi: https://doi.org/10.1093/bioinformatics/btu153.

Sender, R., Fuchs, S., and Milo, R. (2016). Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. *Cell* 164(3)**,** 337-340. doi: https://doi.org/10.1016/j.cell.2016.01.013.

Sereika, M., Kirkegaard, R.H., Karst, S.M., Michaelsen, T.Y., Sørensen, E.A., Wollenberg, R.D., et al. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods* 19(7)**,** 823-826. doi: https://doi.org/10.1038/s41592-022-01539-7.

Shelburne, S.A., Sahasrabhojane, P., Saldana, M., Yao, H., Su, X., Horstmann, N., et al. (2014). *Streptococcus mitis* strains causing severe clinical disease in cancer patients. *Emerging Infectious Diseases* 20(5)**,** 762-771. doi: https://doi.org/10.3201/eid2005.130953.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature* 550**,** 345. doi: https://doi.org/10.1038/nature24286.

Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K.A., Kelly, D.J., et al. (2013). Genome-wide association study identifies vitamin B_5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences* 110(29)**,** 11923-11927. doi: https://doi.org/10.1073/pnas.1305559110.

Shi, W., Sun, Q., Fan, G., Hideaki, S., Moriya, O., Itoh, T., et al. (2020). gcType: a high-quality type strain genome database for microbial phylogenetic and functional research. *Nucleic Acids Research*. doi: https://doi.org/10.1093/nar/gkaa957.

Sichtig, H., Minogue, T., Yan, Y., Stefan, C., Hall, A., Tallon, L., et al. (2019). FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. *Nature Communications* 10(1)**,** 3313. doi: https://doi.org/10.1038/s41467-019-11306-6.

Simões, A.S., Tavares, D.A., Rolo, D., Ardanuy, C., Goossens, H., Henriques-Normark, B., et al. (2016). *lytA*-based identification methods can misidentify *Streptococcus pneumoniae*. *Diagnostic Microbiology and Infectious Disease* 85(2)**,** 141-148. doi: https://doi.org/10.1016/j.diagmicrobio.2016.03.018.

Simpson, J.T., and Durbin, R. (2010). Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26(12)**,** i367-i373. doi: https://doi.org/10.1093/bioinformatics/btq217.

Smith, L.M., Fung, S., Hunkapiller, M.W., Hunkapiller, T.J., and Hood, L.E. (1985). The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Research* 13(7)**,** 2399-2412. doi: https://doi.org/10.1093/nar/13.7.2399.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., et al. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* 321(6071)**,** 674-679. doi: https://doi.org/10.1038/321674a0.

Smits, T.H.M. (2019). The importance of genome sequence quality to microbial comparative genomics. *BMC Genomics* 20(1)**,** 662. doi: https://doi.org/10.1186/s12864-019-6014-5.

Sohn, J.-i., and Nam, J.-W. (2018). The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics* 19(1)**,** 23-40. doi: https://doi.org/10.1093/bib/bbw096.

Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in

bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 44(4)**,** 846-849. doi: https://doi.org/10.1099/00207713-44-4-846.
Stackebrandt, E.E., J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today* 33(4)**,** 152-155.
Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research* 6(7)**,** 2601-2610. doi: https://doi.org/10.1093/nar/6.7.2601.
Steen, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., et al. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME Journal* 13(12)**,** 3126-3130. doi: https://doi.org/10.1038/s41396-019-0484-y.
Steinegger, M., and Salzberg, S.L. (2020). Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology* 21(1)**,** 115. doi: https://doi.org/10.1186/s13059-020-02023-1.
Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., and Watson, M. (2019). Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nature Biotechnology* 37(8)**,** 953-961. doi: https://doi.org/10.1038/s41587-019-0202-3.
Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., et al. (2000). Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* 406(6799)**,** 959-964. doi: https://doi.org/10.1038/35023079.
Straume, D., Stamsås, G.A., and Håvarstein, L.S. (2015). Natural transformation and genome evolution in *Streptococcus pneumoniae*. *Infection, Genetics and Evolution* 33**,** 371-380. doi: https://doi.org/10.1016/j.meegid.2014.10.020.
Strimbu, K., and Tavel, J.A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS* 5(6)**,** 463-466. doi: https://doi.org/10.1097/COH.0b013e32833ed177.
Sutcliffe, I.C., Dijkshoorn, L., Whitman, W.B., and on behalf of the ICSP Executive Board (2020). Minutes of the International Committee on Systematics of Prokaryotes online discussion on the proposed use of gene sequences as type for naming of prokaryotes, and outcome of vote. *International Journal of Systematic and Evolutionary Microbiology* 70(7)**,** 4416-4417. doi: https://doi.org/10.1099/ijsem.0.004303.
Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., and Otto, T.D. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nature Protocols* 7(7)**,** 1260-1284. doi: https://doi.org/10.1038/nprot.2012.068.
Swerdlow, H., and Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research* 18(6)**,** 1415-1419. doi: https://doi.org/10.1093/nar/18.6.1415.
Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D.L., et al. (2018). Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet Infectious Diseases* 18(3)**,** 318-327. doi: https://doi.org/10.1016/S1473-3099(17)30753-3.
Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2017). DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* 34(6)**,** 1037-1039. doi: https://doi.org/10.1093/bioinformatics/btx713.
Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research* 44(14)**,** 6614-6624. doi: https://doi.org/10.1093/nar/gkw569.
Tavares, D.A., Handem, S., Carvalho, R.J., Paulo, A.C., de Lencastre, H., Hinds, J., et al. (2019). Identification of *Streptococcus pneumoniae* by a real-time PCR assay targeting

SP2020. *Scientific Reports* 9(1), 3285. doi: https://doi.org/10.1038/s41598-019-39791-1.

Teles, C., Smith, A., Ramage, G., and Lang, S. (2011). Identification of clinically relevant viridans group streptococci by phenotypic and genotypic analysis. *European Journal of Clinical Microbiology & Infectious Diseases* 30(2), 243-250. doi: https://doi.org/10.1007/s10096-010-1076-y.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102(39), 13950-13955. doi: https://doi.org/10.1073/pnas.0506758102.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814), 796-815. doi: https://doi.org/10.1038/35048692.

The *C. elegans* sequencing consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282(5396), 2012-2018. doi: https://doi.org/10.1126/science.282.5396.2012.

The Human Microbiome Jumpstart Reference Strains Consortium (2010). A catalog of reference genomes from the human microbiome. *Science* 328(5981), 994-999. doi: https://doi.org/10.1126/science.1183605.

The UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51(D1), D523-D531. doi: 10.1093/nar/gkac1052.

Tindall, B.J. (2008). Confirmation of deposit, but confirmation of what? *International Journal of Systematic and Evolutionary Microbiology* 58(8), 1785-1787. doi: https://doi.org/10.1099/ijs.0.2008/006023-0.

Tindall, B.J., and Garrity, G.M. (2008). Proposals to clarify how type strains are deposited and made available to the scientific community for the purpose of systematic research. *International Journal of Systematic and Evolutionary Microbiology* 58(8), 1987-1990. doi: https://doi.org/10.1099/ijs.0.2008/006155-0.

Tindall, B.J., Rosselló-Móra, R., Busse, H.-J., Ludwig, W., and Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *International Journal of Systematic and Evolutionary Microbiology* 60(1), 249-266. doi: https://doi.org/10.1099/ijs.0.016949-0.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J.A., et al. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* 21(1), 180. doi: https://doi.org/10.1186/s13059-020-02090-4.

Troeger, C., Forouzanfar, M., Rao, P.C., Khalil, I., Brown, A., Swartz, S., et al. (2017). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases* 17(11), 1133-1161. doi: https://doi.org/10.1016/S1473-3099(17)30396-1.

Tsai, I.J., Otto, T.D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology* 11(4), R41. doi: https://doi.org/10.1186/gb-2010-11-4-r41.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* 5(1), 170203. doi: https://doi.org/10.1038/sdata.2017.203.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The Human Microbiome Project. *Nature* 449(7164), 804-810. doi: https://doi.org/10.1038/nature06244.

Tyagi, M., da Fonseca, M.M.R., and de Carvalho, C.C.C.R. (2011). Bioaugmentation and biostimulation strategies to improve the effectiveness of bioremediation processes. *Biodegradation* 22(2)**,** 231-241. doi: https://doi.org/10.1007/s10532-010-9394-4.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978)**,** 37-43. doi: https://doi.org/10.1038/nature02340.

Udaondo, Z., Molina, L., Segura, A., Duque, E., and Ramos, J.L. (2016). Analysis of the core genome and pangenome of *Pseudomonas putida*. *Environmental Microbiology* 18(10)**,** 3268-3283. doi: https://doi.org/10.1111/1462-2920.13015.

Uddin, F., Sohail, M., Shaikh, Q.H., Hussain, M.T., Roulston, K., and McHugh, T.D. (2022). Verona integron-encoded metallo-Beta-lactamase (VIM) and Vietnam extended-spectrum Beta-lactamase (VEB) producing *Pseudomonas balearica* from a clinical specimen. *Journal Of Pakistan Medical Association* 72(4)**,** 761-763. doi: https://doi.org/10.47391/JPMA.3890.

Valenzuela, M., González, M., Velásquez, A., Dorta, F., Montenegro, I., Besoain, X., et al. (2021). Analyses of virulence genes of *Clavibacter michiganensis* subsp. *michiganensis* strains reveal heterogeneity and deletions that correlate with pathogenicity. *Microorganisms* 9(7)**,** 1530.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics* 30(9)**,** 418-426. doi: https://doi.org/10.1016/j.tig.2014.07.001.

van Niel, C.B., and Allen, M.B. (1952). A note on *Pseudomonas stutzeri*. *Journal of Bacteriology* 64(3)**,** 413-422. doi: https://doi.org/10.1128/jb.64.3.413-422.1952.

Vandamme, P., Pot, B., Gillis, M., de Vos, P., Kersters, K., and Swings, J. (1996). Polyphasic taxonomy, a consensus approach to bacterial systematics. *Microbiological Reviews* 60(2)**,** 407-438. doi: https://doi.org/10.1128/mr.60.2.407-438.1996.

Vaser, R., Sović, I., Nagarajan, N., Šikić, M., and 1, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* 27(5)**,** 737-746. doi: https://doi.org/10.1101/gr.214270.116.

Vezzi, F., Narzisi, G., and Mishra, B. (2012). Reevaluating assembly evaluations with Feature Response Curves: GAGE and Assemblathons. *PLOS ONE* 7(12)**,** e52210. doi: https://doi.org/10.1371/journal.pone.0052210.

Wahl, B., O'Brien, K.L., Greenbaum, A., Majumder, A., Liu, L., Chu, Y., et al. (2018). 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(7)**,** e744-e757. doi: https://doi.org/10.1016/S2214-109X(18)30247-X.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11)**,** e112963. doi: https://doi.org/10.1371/journal.pone.0112963.

Watson, J.D., and Crick, F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171(4356)**,** 737-738. doi: https://doi.org/10.1038/171737a0.

Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology* 37(2)**,** 124-126. doi: https://doi.org/10.1038/s41587-018-0004-z.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., et al. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology* 37(4)**,** 463-464. doi: https://doi.org/10.1099/00207713-37-4-463.

Weinberger, D.M., Malley, R., and Lipsitch, M. (2011). Serotype replacement in disease after pneumococcal vaccination. *The Lancet* 378(9807)**,** 1962-1973. doi: https://doi.org/10.1016/S0140-6736(10)62225-8.

Welker, M., and Moore, E.R.B. (2011). Applications of whole-cell matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry in systematic microbiology. *Systematic and Applied Microbiology* 34(1)**,** 2-11. doi: https://doi.org/10.1016/j.syapm.2010.11.013.

Wences, A.H., and Schatz, M.C. (2015). Metassembler: merging and optimizing de novo genome assemblies. *Genome Biology* 16**,** 207. doi: https://doi.org/10.1186/s13059-015-0764-4.

Wessels, E., Schelfaut, J.J.G., Bernards, A.T., and Claas, E.C.J. (2012). Evaluation of several biochemical and molecular techniques for identification of *Streptococcus pneumoniae* and *Streptococcus pseudopneumoniae* and their detection in respiratory samples. *Journal of Clinical Microbiology* 50(4)**,** 1171-1177. doi: https://doi.org/10.1128/JCM.06609-11.

Wetzel, J., Kingsford, C., and Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies. *BMC Bioinformatics* 12(1)**,** 95. doi: https://doi.org/10.1186/1471-2105-12-95.

Whatmore, A.M., Efstratiou, A., Pickerill, A.P., Broughton, K., Woodard, G., Sturgeon, D., et al. (2000). Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. *Infection and Immunity* 68(3)**,** 1374-1382. doi: https://doi.org/10.1128/iai.68.3.1374-1382.2000.

Whelan, F.J., Hall, R.J., and McInerney, J.O. (2021). Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Molecular Biology and Evolution*. doi: https://doi.org/10.1093/molbev/msab139.

Whiley, R.A., and Hardie, J.M. (2015). "*Streptococcus*," in *Bergey's Manual of Systematics of Archaea and Bacteria*, eds. M.E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F.A. Rainey & W.B. Whitman.), 1-86.

White, J.C., and Niven, C.F. (1946). *Streptococcus* s.b.e.: a *Streptococcus* associated with subacute bacterial endocarditis. *Journal of Bacteriology* 51(6)**,** 717-722. doi: https://doi.org/10.1128/jb.51.6.717-722.1946.

Whitman, W.B. (2016). Modest proposals to expand the type material for naming of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology* 66(5)**,** 2108-2112. doi: https://doi.org/10.1099/ijsem.0.000980.

Whitman, W.B., Sutcliffe, I.C., and Rossello-Mora, R. (2019). Proposal for changes in the International Code of Nomenclature of Prokaryotes: granting priority to *Candidatus* names. *International Journal of Systematic and Evolutionary Microbiology*. doi: https://doi.org/10.1099/ijsem.0.003419.

Whitney, C.G., Farley, M.M., Hadler, J., Harrison, L.H., Lexau, C., Reingold, A., et al. (2000). Increasing prevalence of multidrug-resistant *Streptococcus pneumoniae* in the United States. *New England Journal of Medicine* 343(26)**,** 1917-1924. doi: https://doi.org/10.1056/nejm200012283432603.

Wick, R. (2017). "Filtlong: quality filtering tool for long reads".).

Wick, R., and Holt, K. (2019). Benchmarking of long-read assemblers for prokaryote whole genome sequencing [version 1; peer review: 4 approved]. *F1000Research* 8(2138). doi: https://doi.org/10.12688/f1000research.21782.1.

Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology* 13(6)**,** e1005595. doi: https://doi.org/10.1371/journal.pcbi.1005595.

Wick, R.R., Judd, L.M., and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* 20(1)**,** 129. doi: https://doi.org/10.1186/s13059-019-1727-y.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* 74(11)**,** 5088-5090. doi: https://doi.org/10.1073/pnas.74.11.5088.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* 87(12)**,** 4576-4579. doi: https://doi.org/10.1073/pnas.87.12.4576.

World Health Organization (2014). *Antimicrobial resistance: global report on surveillance.* Geneva, Switzerland.

World Health Organization (WHO) (2018). "The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in *Mycobacterium tuberculosis* complex: technical guide". (Geneva: World Health Organization).

Woyke, T., Doud, D.F.R., and Schulz, F. (2017). The trajectory of microbial single-cell sequencing. *Nature Methods* 14(11)**,** 1045-1054. doi: https://doi.org/10.1038/nmeth.4469.

Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., et al. (2010). One bacterial cell, one complete genome. *PLOS ONE* 5(4)**,** e10314. doi: https://doi.org/10.1371/journal.pone.0010314.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462(7276)**,** 1056-1060. doi: https://doi.org/10.1038/nature08656.

Wu, L., and Ma, J. (2019). The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation. *International Journal of Systematic and Evolutionary Microbiology* 69(4)**,** 895-898. doi: https://doi.org/10.1099/ijsem.0.003276.

Wu, L., McCluskey, K., Desmeth, P., Liu, S., Hideaki, S., Yin, Y., et al. (2018). The Global Catalogue of Microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. *GigaScience***,** giy026-giy026. doi: https://doi.org/10.1093/gigascience/giy026.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., et al. (2008). The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology* 31(4)**,** 241-250. doi: https://doi.org/10.1016/j.syapm.2008.07.001.

Yarza, P., Spröer, C., Swiderski, J., Mrotzek, N., Spring, S., Tindall, B.J., et al. (2013). Sequencing orphan species initiative (SOS): filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. *Systematic and Applied Microbiology* 36(1)**,** 69-73. doi: https://doi.org/10.1016/j.syapm.2012.12.006.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., et al. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S

rRNA gene sequences. *Nature Reviews Microbiology* 12(9), 635-645. doi: https://doi.org/10.1038/nrmicro3330.

Yombi, J.c., Belkhir, L., Jonckheere, S., Wilmes, D., Cornu, O., Vandercam, B., et al. (2012). *Streptococcus gordonii* septic arthritis: two cases and review of literature. *BMC Infectious Diseases* 12(1), 215. doi: https://doi.org/10.1186/1471-2334-12-215.

Yoon, S.H., Ha, S.M., Kwon, S., Lim, J., Kim, Y., Seo, H., et al. (2017). Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International Journal of Systematic and Evolutionary Microbiology* 67(5), 1613-1617. doi: https://doi.org/10.1099/ijsem.0.001755.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18(5), 821-829. doi: https://doi.org/10.1101/gr.074492.107.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLOS ONE* 6(3), e17915. doi: https://doi.org/10.1371/journal.pone.0017915.