UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Inter-laboratory reproducibility and expansion of the gene-by-gene typing approach in an era of whole genome sequencing

Rafael Fresca Mamede

Orientador: Professor Doutor Mário Nuno Ramos de Almeida Ramirez
Co-orientador: Doutor Simon Hubert Tausch

Documento provisório
Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade em Biologia Computacional

2025

UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA

Inter-laboratory reproducibility and expansion of the gene-by-gene typing approach in an era of whole genome sequencing

Rafael Fresca Mamede

Orientador: Professor Doutor Mário Nuno Ramos de Almeida Ramirez
Co-orientador: Doutor Simon Hubert Tausch

Documento provisório
Tese especialmente elaborada para obtenção do grau de Doutor em Ciências e Tecnologias da Saúde, especialidade em Biologia Computacional

Fundação para a Ciência e Tecnologia (2020.08493.BD)
2025

"We are all in the depths of a cave, chained by our ignorance, by our prejudices, and our weak senses reveal to us only shadows. If we try to see further, we are confused; we are unaccustomed. But we try. This is science."

- Carlo Rovelli, Reality Is Not What It Seems: The Journey to Quantum Gravity

Acknowledgements

I express my deepest gratitude to my supervisor, Prof. Dr. Mário Ramirez, for his mentorship and support throughout the years. Our invaluable discussions about the subjects presented in this thesis were crucial to reaching this point and for me to be able to implement many of the concepts that we discussed. His eagle-eyed view of the concepts and results revealed limitations in the approaches I was trying to follow, steering me towards more fruitful endeavors.

To all of the colleagues in the MRamirez lab, thank you for all the shared moments and knowledge about streptococci and microbiology in general. My work was heavily focused on the technical aspects of software development, which could make me lose sight of the approaches that are more biologically relevant. Working in an environment where people have strong knowledge about the biology of bacterial pathogens and a close link to epidemiological and clinically relevant applications helped contextualize my work and guide my efforts more effectively.

Special thanks to Inês, Pedro, and Joana for their friendship and for keeping things a little more lively and interesting. It takes a lot of energy to deal with an introvert, but you had enough patience. When I arrived at the lab, I was only a "proto-binfie". I improved in great part due to the mentorship of Inês and through the knowledge shared with her and Pedro while working on several projects. That period allowed me to learn and refine multiple skills and greatly contributed to the type of "binfie" that I have become.

To all friends and family who supported me during this journey, thank you for being present, caring, and sharing. Every casual moment, every conversation, even if fleeting, are important aspects of what makes life worthwhile. My fiancée, Joana, through her unconditional love and support, has helped me reevaluate and overcome many challenges. Our relationship has been a source of joy and motivation that has brought countless wonderful moments and has given me the strength to surpass darker ones.

Lastly, to my grandfather and my stepfather, from whom I have learnt a lot, through words and actions. I shall remember you fondly, through all the precious moments we have spent together, and hope that in me resides a mere fraction of your perseverance, intelligence, wit, and wisdom.

Summary

Bacterial infections are a huge burden for public health systems globally, causing significant health and economic losses. The emergence of accurate and cost-efficient high-throughput DNA sequencing technologies, especially by enabling whole genome sequencing of bacterial genomes, has revolutionized the characterization of bacterial strains in applications such as surveillance and outbreak investigation. These technologies have been widely adopted by research and public health institutions, providing increased resolution in a wide range of applications to complement or replace more classical phenotypic and molecular assays. The wealth and complexity of the generated data demanded greater storage capacity and improved computational methods for data analysis to make sense of the data. These demands potentiated a tremendous growth in the field of bioinformatics, which has become an integral part in omics approaches such as genomics. In bacterial genomics, in particular, bioinformatics methods have become essential to characterize bacterial strains, providing higher resolution in areas such as infectious disease surveillance and outbreak investigation, as well as for the study of the structure and evolution of bacterial populations. Whole genome sequencing allows for a purely sequence-based approach for bacterial characterization, enabling targeted approaches to identify genes of interest or the analysis of the full gene diversity. The increased availability of complete or nearly complete bacterial genome assemblies allowed researchers to study the structure and variability of bacterial genomes and encouraged the development of approaches for high-resolution bacterial typing, such as gene-by-gene and SNP-based methods. Although both approaches have been extensively applied in comparative genomics, either separately or in combination, to study the diversity of bacterial populations, gene-by-gene methods such as whole- and core-genome multilocus sequence typing (wg/cgMLST) have been adopted more frequently by research and public health institutions. wg/cgMLST allows to create schemas to capture the loci diversity of species of interest. The schemas enable the characterization of bacterial strains and can be updated over time with new alleles to maximize applicability in the long term. This is a gradual process that is commonly performed by web platforms that centralize data analysis. The efficiency of data analyses in centralized systems may be sufficient for routine surveillance, but it raises scalability concerns when it is necessary to perform large-scale analyses in reduced time, especially with the increase in the number of genome assemblies that are publicly available. Furthermore, centralized systems require users to upload their data, which may not be possible for users or institutions working under stricter data privacy policies. In addition, most analyses are performed at the core genome

level, targeting the set of loci that are present in most strains, but discarding less frequent loci that constitute a very significant part of gene diversity and may be determinant for relevant phenotypic characteristics, such as virulence and antimicrobial resistance. Thus, developing methods that allow for local and large-scale analyses, while also being able to integrate the diversity of accessory loci more accurately, can minimize scalability and data privacy concerns and considerably expand the resolution of wg/cgMLST. The present thesis presents methods that aim to improve the scalability, accuracy, and interoperability of wg/cgMLST.

The chewBBACA suite for wg/cgMLST served as the basis to explore and implement new methods for improved wg/cgMLST. The new methods were implemented into chewBBACA 3, which constitutes a complete reimplementation of its predecessor, chewBBACA 2. In contrast to chewBBACA 2, which evaluated the coding sequences predicted for each input genome separately, chewBBACA 3 identifies and stores the list of distinct coding sequences predicted from all input genomes to enable fast and non-redundant exact matching and classification at the DNA and protein levels based on sequence hash comparisons. In addition, chewBBACA 3 complements alignment-based allele identification with alignment-free methods, more specifically, minimizer-based clustering, allowing for faster and more accurate allele identification. Schema creation with chewBBACA 3 is up to 55-fold faster than with chewBBACA 2 and identifies up to 10% more loci, allowing to capture more of the diversity of bacterial species. Allele calling with chewBBACA 3 is up to 20.3- and 51.9-fold faster than with chewBBACA 2 and a comparable method, respectively. Furthermore, chewBBACA 3 classifies more coding sequences and scales better than the other methods, allowing large-scale wg/cgMLST in reduced time with computational resources typically available on a laptop. chewBBACA 3 includes functionalities to generate interactive reports that allow for an intuitive and comprehensive evaluation of wg/cgMLST schemas and results. The reports provide results and functionalities to explore loci diversity and identify groups of similar strains, which are relevant for surveillance, outbreak detection, and population studies.

A Web service, Chewie-NS, was implemented to provide broad access to wg/cgMLST schemas and allow local and private analyses based on a common allelic nomenclature. Chewie-NS leverages containerization to combine various technologies into two main components: a backend and a frontend components. The backend component includes the databases used to store and manage user data and wg/cgMLST schemas, as well as an API that accepts and processes user requests and provides data for the frontend component. The API allowed to develop a set of modules for integration with chewBBACA 3 to provide functionalities for schema download, upload, and synchronization. The integration with chewBBACA 3 allows users to quickly set up a wg/cgMLST schema for local and scalable analysis, retrieve novel alleles added to the remote schemas in Chewie-NS, and contribute with novel alleles identified locally only if desired. The synchronization process maintains the allelic nomenclatures used by local and remote schemas synchronized to ensure the comparability of the results. This decentralized approach contrasts with the centralized model adopted by other well-established web platforms for wg/cgMLST, which require users to upload their data to the platform, raising scalability and data privacy concerns. The frontend component renders the Chewie-NS

website, providing easy access to the list of available schemas for download and relevant statistics about schema composition and loci diversity. In addition, the website links to a graphical interface for the API, which allows users of any level of expertise to explore the API more intuitively and retrieve detailed schema and loci data.

A novel wgMLST schema for *Streptococcus pyogenes*, comprising 3,044 loci, was developed based on datasets representative of the species diversity. The loci in the schema were annotated by retrieving functional annotation data from several sources. A careful curation process by a domain expert allowed to validate the loci annotations and refine the schema by substituting or removing spurious loci. The solutions created to resolve the issues identified during the curation process can be integrated into workflows to improve the quality of wg/cgMLST schemas and analyses. The annotated wgMLST schema provides increased resolution compared to more classical typing methods, such as PFGE and seven-gene MLST, and displays performance comparable to SNP-based methods. Using a wgMLST schema, instead of a hard-defined cgMLST schema, enables scalable cgMSLT analysis where the set of core loci is adjusted based on the dataset under analysis. The schema provided high discriminatory power to characterize and distinguish strains in a dataset representing the global diversity of *S. pyogenes*, as well as in outbreak context to distinguish strains from recently emerged lineages.

In conclusion, the methods and results presented in this thesis seek to improve current approaches for bacterial characterization based on whole genome sequencing. chewBBACA 3 lowers the barrier for scalable and comprehensive wg/cgMLST. Chewie-NS, while not as feature-complete as other well-established platforms, provides easy access to schemas and aims to minimize scalability and data privacy concerns. The wgMLST schema for *S. pyogenes* allows detailed strain characterization at any resolution level, and its development enabled the identification of key issues and solutions to improve the quality of wg/cgMLST schemas.

Keywords: Public health, Pathogen surveillance, Bacterial genomics, Bacterial typing, wg/cgMLST

Resumo

As infecções bacterianas têm um grande impacto nos sistemas de saúde pública a um nível global, causando perdas significativas tanto do ponto de vista económico como de saúde. A emergência de tecnologias de sequenciação de ADN de alto débito de elevada precisão e com custo reduzido, especialmente por permitirem a sequenciação total do genoma bacteriano, revolucionou a caracterização de espécies bacterianas em aplicações como a vigilância e investigação de surtos. Estas tecnologias têm sido adoptadas por instituições de investigação e de saúde pública, contribuindo para o aumento da resolução das análises em diversas aplicações e complementando ou substituindo técnicas clássicas de fenotipagem e moleculares. A escala e complexidade dos dados gerados requer maior capacidade de armazenamento de dados e melhores métodos computacionais para análise de dados. Estes requisitos potenciaram um crescimento tremendo na área de bioinformática, que se tornou numa das partes integrantes das abordagens ómicas como a genómica. Na genómica de bactérias, em particular, os métodos bioinformáticos tornaram-se essenciais para caracterizar estirpes bacterianas, aumentando a resolução de aplicações em áreas como a vigilância de doenças infecciosas e investigação de surtos, bem como para o estudo da estrutura e evolução das populações bacterianas. A sequenciação total do genoma torna possível uma abordagem baseada apenas na composição das sequências para a caracterização bacteriana, permitindo abordagens mais dirigidas para identificar genes de interesse ou a análise da diversidade genética completa. O aumento do número de genomas bacterianos completos ou parciais disponíveis em bases de dados públicas permitiu aos investigadores estudar a estrutura e a variabilidade dos genomas bacterianos e encorajou o desenvolvimento de métodos de tipagem de alta resolução, como os métodos *gene-by-gene* e baseados em *SNPs*. Apesar de ambas as abordagens serem frequentemente aplicadas em genómica comparativa, quer separadamente ou em combinação, para estudar a diversidade de populações bacterianas, os métodos *gene-by-gene* como *whole-* e *core-genome multilocus sequence typing* (*wg/cgMLST*) têm sido adoptados com maior frequência por instituições de investigação e saúde pública. Os métodos *wg/cgMLST* permitem criar esquemas que capturam a diversidade genética de espécies de interesse. Estes esquemas são utilizados para a caracterização de estirpes bacterianas e podem ser actualizados com novos alelos ao longo do tempo para maximizar a sua aplicabilidade a longo prazo. Este é um processo gradual que é geralmente efectuado por plataformas *web* que centralizam as análises. A eficiência das análises em sistemas centralizados pode ser suficiente para actividades de vigilância, mas levanta questões de escalabilidade quando é necessário

efectuar análises de grande escala em tempo reduzido, especialmente à medida que o número de genomas disponíveis aumenta. Para além disso, os sistemas centralizados requerem que os utilizadores disponibilizem os seus dados, o que poderá não ser possível para utilizadores que trabalhem sob políticas de partilha de dados mais restringentes. A maior parte das análises também são efetuadas ao nível do genoma *core*, focando-se no conjunto de *loci* que estão presentes na maior parte das estirpes, mas descartando *loci* menos frequentes que constituem uma parte muito significativa da diversidade genética e que podem ser determinantes para características fenotípicas relevantes, como virulência e resistência aos antimicrobianos. Deste modo, o desenvolvimento de métodos que permitam análises de grande escala localmente, e que também sejam capazes de integrar a diversidade de *loci* do genoma acessório de forma mais precisa, pode minimizar questões relativas a escalabilidade e confidencialidade dos dados e expandir consideravelmente a resolução dos métodos *wg/cgMLST*. A presente tese apresenta métodos que procuram melhorar a escalabilidade, exactidão, e a interoperabilidade de métodos *wg/cgMLST*.

A ferramenta chewBBACA, utilizada para análises *wg/cgMLST*, serviu de base para explorar e implementar novos métodos para melhorar os métodos *wg/cgMLST*. Os novos métodos foram implementados na ferramenta chewBBACA 3, que corresponde a uma reimplementação completa do seu precursor, o chewBBACA 2. Comparativamente ao chewBBACA 2, que avaliava as sequências codificantes previstas para cada genoma separadamente, o chewBBACA 3 identifica e armazena a lista de sequências codificantes distintas previstas a partir de todos os genomas para possibilitar correspondências exactas rápidas e não redundantes e a classificação ao nível de ADN e de proteína com base em comparações de *hashes* de sequências. O chewBBACA 3 também complementa a identificação de alelos com base em alinhamento com métodos *alignment-free*, mais especificamente, *clustering* com base em *minimizers*, permitindo uma identificação mais rápida e precisa de alelos. A criação de esquemas com o chewBBACA 3 é até 55 vezes mais rápida do que com o chewBBACA 2 e identifica até 10% mais *loci*, permitindo capturar mais da diversidade de espécies bacterianas. A identificação de alelos com o chewBBACA 3 é 20.3 a 51.9 vezes mais rápida do que com o chewBBACA 2 e outro método comparável, respectivamente. Para além disso, o chewBBACA 3 classifica mais sequências codificantes e escala melhor do que os outros métodos, permitindo análises *wg/cgMLST* a grande escala em tempo reduzido e com recursos disponíveis num portátil. O chewBBACA 3 inclui funcionalidades para criar relatórios interactivos para uma análise intuitiva e detalhada dos esquemas e resultados *wg/cgMLST*. Os resultados e funcionalidades dos relatórios permitem explorar a diversidade genética e identificar grupos de estirpes semelhantes, aspectos relevantes para vigilância, detecção de surtos, e estudos populacionais.

Um *Web service*, denomeado Chewie-NS, foi implementado para disponibilizar esquemas *wg/cgMLST* e possibilitar análises locais e privadas com base numa nomenclatura alélica comum. O Chewie-NS combina várias tecnologias através de *containerization*, sendo constituído por dois componentes: um componente *backend* e um component *frontend*. O componente de *backend* inclui as bases de dados para armazenar e gerir os dados dos utilizadores e dos

esquemas *wg/cgMLST*, bem como uma API que aceita e processa os pedidos dos utilizadores e disponibiliza os dados para o componente *frontend*. A API permitiu desenvolver um conjunto de módulos para integração com o chewBBACA 3 de forma a disponibilizar funcionalidades para descarregar, carregar e sincronizar esquemas. A integração com o chewBBACA 3 simplifica a preparação de esquemas *wg/cgMLST* para análises locais e escaláveis, a obtenção de novos alelos a partir dos esquemas remotos depositados no Chewie-NS, e a contribuição de novos alelos identificados localmente. O processo de sincronização mantém as nomenclaturas alélicas utilizadas pelos esquemas locais e remotos sincronizadas para assegurar que os resultados são comparáveis. Esta abordagem descentralizada contrasta com o modelo centralizado adoptado por outras plataformas *web* para *wg/cgMLST*, que requerem que os utilizadores carreguem os seus dados para a plataforma, levantando questões de escalabilidade e confidencialidade de dados. O componente *frontend* efectua o *render* do *website* do Chewie-NS, disponibilizando a lista de esquemas disponíveis para descarregar e estatísticas relevantes sobre a composição dos esquemas e a diversidade genética. O *website* também disponibiliza uma ligação para uma interface gráfica da API, permitindo que utilizadores com diferentes níveis de proficiência explorem a API de forma mais intuitiva e tenham acesso a dados detalhados sobre os esquemas e *loci*.

Um novo esquema *wgMLST* para *Streptococcus pyogenes*, constituído por 3,044 loci, foi desenvolvido a partir de conjuntos de dados representativos da diversidade da espécie. Os *loci* do esquema foram anotados funcionalmente com base em várias fontes. Um processo de curadoria feito por uma especialista em *S. pyogenes* permitiu validar as anotações e refinar o esquema através da substituição ou remoção de *loci* espúrios. As soluções criadas para resolver os problemas identificados durante o processo de refinamento podem ser integradas em fluxos de trabalho para melhorar a qualidade de esquemas e análises *wg/cgMLST*. O esquema *wgMLST* anotado melhora a resolução das análises comparativamente com métodos de tipagem mais clássicos, como PFGE e MLST de sete genes, apresentando desempenho semelhante a métodos baseados em *SNPs*. A utilização de um esquema *wgMLST*, em vez de um esquema *cgMLST* mais estrito, permite análises *cgMLST* escaláveis em que o conjunto de *core loci* é ajustado com base no conjunto de dados em análise. O esquema proporciona alto poder discriminatório para caracterizar e distinguir estirpes num conjunto de dados representante da diversidade global de *S. pyogenes*, bem como num contexto de surto para distinguir estirpes de linhagens emergentes.

Em conclusão, os métodos e resultados apresentados nesta tese pretendem melhorar as actuais abordagens para a caracterização de bactérias com base em sequenciação total do genoma. O chewBBACA 3 reduz os requisitos para análises *wg/cgMLST* escaláveis e detalhadas. O *web service* Chewie-NS, mesmo não sendo tão completo do ponto de vista de funcionalidades comparativamente com outras plataformas *wg/cgMLST*, simplifica o acesso a esquemas e pretende minimizar problemas relacionados com escalabilidade e confidencialidade dos dados. O esquema *wgMLST* para *S. pyogenes* permite uma caracterização detalhada de estirpes a qualquer nível de resolução, e o seu desenvolvimento permitiu identificar problemas comuns e soluções para melhorar a qualidade de esquemas *wg/cgMLST*.

Keywords: Saúde pública, Vigilância de patogéneos, Genómica bacteriana, Tipagem bacteriana, wg/cgMLST

Thesis Outline

This thesis describes the development and application of improved bioinformatics methods for bacterial typing based on Whole Genome Sequencing (WGS), more specifically Gene-by-Gene (GbG) approaches such as whole- and core-genome MLST (wg/cgMLST). wg/cgMLST is widely used for the characterization of bacterial strains by research and public health institutions for surveillance, outbreak detection, and for the study of the population diversity of bacterial species. Therefore, improving wg/cgMLST methodologies is of high relevance and has a direct impact on research and public health, since any improvement can easily be adopted by the scientific community and public health authorities who already apply these methodologies. The thesis is organized into five chapters.

Chapter 1 is a general introduction to bacterial typing methods, starting with a brief explanation of some of the classical biochemical and molecular methods used in the laboratory and extending this to transition to the WGS-based approaches made possible by the DNA sequencing revolution. The main WGS-based approaches for bacterial typing are explained, with a greater focus on GbG approaches, such as wg/cgMLST, which are explained more thoroughly.

The general introduction is followed by the body of the thesis, composed of three selected manuscripts that present methods and results that aim to improve the efficiency, accuracy, and interoperability of wg/cgMLST. Each manuscript has its own dedicated chapter. There is a high degree of intersection between the subjects of each chapter, and the work of each chapter contributes in part or is instrumental to advance the work presented in the other chapters.

Chapter 2 describes the implementation and evaluation of a bioinformatics tool for wg/cgMLST, chewBBACA 3. chewBBACA 3 is a reimplementation of the first published version of chewBBACA and was designed to be a complete solution for wg/cgMLST, offering improved and new functionalities that considerably expand chewBBACA's capabilities. It enables fast and efficient schema creation from multiple sources and allele calling with modest computational resources to meet current and future data processing demands. The speed and scalability of the allele calling opens up the possibility of large-scale wg/cgMLST for more users who want to study the population diversity of a species of interest based on large datasets or in surveillance and outbreak scenarios where a timely analysis is crucial. To take advantage of the extensive schema and results data produced, chewBBACA 3 includes

modules for a comprehensive analysis of the schemas and allele calling results. These modules produce interactive and easily shareable reports based on an upgraded version of the User Interface (UI) developed for the Web Service (WS) presented in **Chapter 3**, allowing users of any level of expertise to more easily explore the results and make informed decisions.

Chapter 3 presents Chewie Nomenclature Server (Chewie-NS), a Nomenclature Server (NS) that stores wg/cgMLST schemas and allows for local and private wg/cgMLST analysis based on a common allelic nomenclature. The clear and simple interface of Chewie-NS allows users to easily find, explore loci diversity, and download schemas for species of interest. For users who are more tech-savvy, the Application Programming Interface (API) provides access to all schema data, either through Swagger UI or programmatically. The schemas downloaded from Chewie-NS can be used to perform local and private analysis, minimizing scalability issues if data analysis was centralized in a WS and allowing users operating under stricter data privacy policies to still take advantage of the data and functionalities provided by Chewie-NS. This contrasts with other wg/cgMLST platforms, which typically centralize data analysis, requiring users to submit their data and establishing queues to process data submissions based on available computational resources. The local schemas downloaded from Chewie-NS can be synchronized with the remote versions to retrieve novel alleles added to the remote schemas and contribute novel alleles identified locally if desired. The synchronization process allows users to update both schema versions and the allelic nomenclature, enabling results comparison even if data analysis is not centralized. To provide functionalities for users to use the schemas deposited in Chewie-NS, a set of modules was developed for chewBBACA 3, presented in **Chapter 2**, allowing users to upload, download and synchronize the schemas. The integration with chewBBACA 3 provides access to the schemas deposited in Chewie-NS, allowing users to perform local allele calling and compare their results based on a common allelic nomenclature, while also allowing users to take advantage of chewBBACA's powerful analytic capabilities.

Chapter 4 describes the creation and evaluation of an annotated whole-genome MLST (wgMLST) schema for *S. pyogenes*, a major human pathogen with high genetic diversity. The target loci for the wgMLST schema were defined based on a dataset of high-quality complete genomes and the schema was populated by allele calling with multiple datasets representing the known diversity of *S. pyogenes*. This schema was refined based on automatic annotations and the suggestions of a domain expert. The refined wgMLST schema, comprising 3,044 loci, allows for high-resolution typing of *S. pyogenes* for the study of diverse datasets and in an outbreak context, showing performance comparable to SNP-based methods. The development of the wgMLST schema benefited from the novel functionalities that were being implemented in chewBBACA 3 concomitantly, presented in **Chapter 2**. The schema refining process also revealed key limitations and challenges of scaling current GbG approaches, often based only on core-genome MLST (cgMLST), to wgMLST, and provided crucial information to guide the development of some functionalities included in chewBBACA 3. The wgMLST schema was deposited in Chewie-NS, presented in **Chapter 3**.

Chapter 5 corresponds to the general discussion. In this chapter, the results presented in each chapter of the body of the thesis are summarized and discussed in terms of their advantages, disadvantages, and the potential to be further improved.

Abbreviations

AI Artificial Intelligence

ABR Antibiotic Resistant

AMR Antimicrobial Resistance

AP-PCR or RAPD arbitrarily primed PCR

ANI Average Nucleotide Identity

API Application Programming Interface

ALM Allele Larger than Mode

AR adjusted Rand

ASM Allele Smaller than Mode

AW adjusted Wallace

BIGSdb Bacterial Isolate Genome Sequence Database

bp base pairs

BPPL Bacterial Priority Pathogens List

BLAST Basic Local Alignment Search Tool

BLASTn Nucleotide BLAST

BLASTp Protein BLAST

BLASTx translated nucleotide BLAST

BSR BLAST Score Ratio

agMLST accessory-genome MLST

cgMLST core-genome MLST

pgMLST pan-genome MLST

wgMLST whole-genome MLST

wg/cgMLST whole- and core-genome Multilocus Sequence Typing

CC clonal complexes

CDC Centers for Disease Control and Prevention

cDNA complementary DNA

CDS coding DNA sequence

CDSs coding DNA sequences

Chewie-NS Chewie Nomenclature Server

CLI Command Line Interface

CPU Central Processing Unit

ddNTPs dideoxynucleotides

dNTPs deoxynucleotides

DNA Deoxyribonucleic acid

EU/EEA European Union and the European Economic Area

ECDC European Centre for Disease Prevention and Control

EFSA European Food Safety Authority

ENA European Nucleotide Archive

EXC Exact Match

FAIR Findable Accessible Interoperable Reusable

FCT Fundação para a Ciência e Tecnologia

FWD Food and Waterborne disease

GIGO Garbage In, Garbage Out

GAS Lancefield group A *Streptococcus*

GbG Gene-by-Gene

GPU Graphics Processing Unit

GUI General User Interface

HGT Horizontal Gene Transfer

HR Homologous Recombination

HTML HyperText Markup Language

HTTPS Hypertext Transfer Protocol Secure

HTS high-throughput sequencing

LOTSC LOTSC

MALDI-TOF matrix-assisted laser desorption/ionization-time-of-flight

MB megabyte

Mb megabases

MGEs Mobile Genetic Elements

MIC minimum inhibitory concentration

MLEE multilocus enzyme electrophoresis

MLVA multiple-locus variable number tandem repeat analysis

MLST Multilocus Sequence Typing

MPS massive parallel sequencing

MSA Multiple Sequence Alignment

MST Minimum-spanning-tree

MTC Mycobacterium tuberculosis complex

NCBI National Center for Biotechnology Information

NJ Neighbor-Joining

NIPH Non-Informative Paralogous Hit

NIPHEM Non-Informative Paralogous Hit Exact Match

NS Nomenclature Server

ONT Oxford Nanopore Technologies

ORF open-reading frame

ORFs open-reading frames

OS Operating System

PacBio Pacific Biosciences

PAMA Paralogous Match

PE paired-end

PCR Polymerase Chain Reaction

PLOT Possible Locus On the Tip

PLOT3 PLOT 3'-end

PLOT5 PLOT 5'-end

PFGE Pulse Field Gel Electrophoresis

PopPUNK POPulation Partitioning Using Nucleotide Kmers

PVL Panton-Valentine leukocidin

kb kilobases

qPCR Real-Time Quantitative PCR

RAM Random Access Memory

REs restriction enzymes

REP-PCR repetitive sequencing-based PCR

REST Representational State Transfer

RDF Resource Description Framework

RFs Restriction Fragments

RFLP Restriction Fragment Length Polymorphism

rRNA ribossomal ribonucleic acid

rps ribosomal protein subunit

RD Research and Development

rSTs ribosomal STs

rMLST Ribosomal MLST

SBS sequencing by synthesis

SBs Synteny Blocks

SID Simpson's index of diversity

SMRT Single Molecule Real-Time Sequencing

SNA Synteny Network Analysis

SNP Single Nucleotide Polymorphism

SNPs Single Nucleotide Polymorphisms

SNV Single Nucleotide Variant

SNVs Single Nucleotide Variants

SPARQL SPARQL Protocol and RDF Query Language

ssDNA single-stranded DNA

ST Sequence Type

SVG Scalable Vector Graphics

TSV Tab-Separated Values

VNTR variable number tandem repeats

wg/cgMLST whole- and core-genome MLST

WGS Whole Genome Sequencing

WHO World Health Organization

WS Web Service

WSGI Web Server Gateway Interface

UI User Interface

UPGMA unweighted pair group method with arithmetic means

Table of Contents

List of Tables

LIST OF TABLES

List of Figures

LIST OF FIGURES

Chapter 1

General Introduction

1.1 The burden of bacterial pathogens

Science has made great strides in understanding the ubiquity and complexity of microorganisms since the first observations by Antonie van Leeuwenhoek in the 17th century. The inquisitiveness of the scientific method revealed a myriad of microscopic shapes and interactions that play a pivotal role in global ecosystems and impact nearly every aspect of human activity, from beneficial applications in the food production industry to the detrimental effects on human and animal health. The latter association was discovered by Robert Koch in the 19th century, whose experiments enabled to identify bacteria as causative agents of deadly infectious diseases. Koch's contributions, along with those of other prominent scientists such as Louis Paster, established the fields of microbiology and bacteriology. The extensive research that followed revealed the enormous diversity of bacterial species, which today constitute one of the three domains of life established by Carl Woese in 1990 [1]. Bacteria are single-celled organisms that lack a nuclear membrane and divide by binary fission. Superficially, bacteria may appear as simple forms of life. In reality, life is rarely simple and bacterial species exhibit tremendous diversity. This diversity arises from the selective forces that shape bacterial evolution towards the path that maximizes competitiveness in each habitat, giving rise to varied morphologies (e.g., cocci, rods, spirilla, filamentous), sizes (from as small as about 0.2 micrometer (μm) in diameter to more than 700 μm in diameter), nutrient preferences (e.g., glucose is a good carbohydrate source for many species, but some species may prefer other sources, such as lactose in the case of lactic acid bacteria) and structures (e.g., differences in cell wall structure between Gram-positive and Gram-negative bacteria). Small structural differences can cause significant changes in the way bacteria interact with the environment or a host. For example, modifications in penicillin-binding proteins present in the cell wall,

to which the antibiotic penicillin binds, consequently weakening the cell wall and leading to cell lysis, can confer reduced susceptibility or resistance to penicillin. The mechanisms through which bacteria develop Antimicrobial Resistance (AMR) are of great concern to public health. In 2021, 4.71 million deaths were estimated to be associated with bacterial AMR [2]. Although deaths attributed to AMR among children under 5 years of age have been decreasing, deaths among older people have increased, and it is estimated that by 2050 8.22 million deaths may be associated with AMR. Reducing the number of deaths related to AMR is an important step to reduce the total number of deaths related to infection. A recent study estimated that in 2019 nearly 14 million deaths were infection-related, with 56% of those deaths associated with only 33 bacterial pathogens (Figure 1.1).

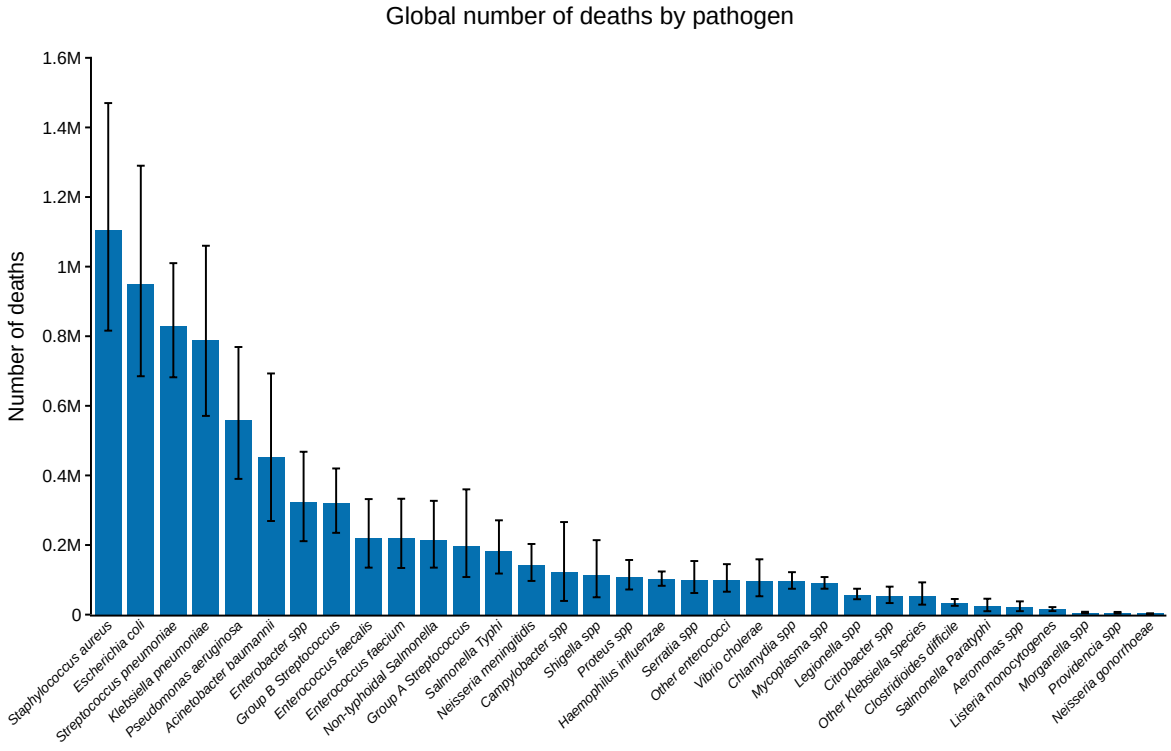


Figure 1.1: Global number of deaths, in millions, associated with 33 bacterial pathogens in 2019. The error bars represent the 95% uncertainty intervals. Adapted from [3].

In the context of global mortality, these estimates put deaths associated with bacterial infections as responsible for at least 14% of all global deaths [3]. Moreover, the top five pathogens were associated with half of all bacterial deaths. These five pathogens are included in the WHO Bacterial Priority Pathogens List (BPPL) updated in 2024 [4]. The BPPL groups 15 families of antibiotic resistant pathogens into priority levels, with the aim of serving as a compass for Research and Development (RD) and for public health action (Figure 1.2). A global concerted effort of public health authorities is of the utmost importance to implement regionally tailored strategies to reduce the burden caused by bacterial infections. The implemented measures should combine infection prevention, vaccination, adequate methods of bacterial characterization and antibiotic usage, as well as promote RD aiming to strengthen the resilience of these measures and future preparedness. Although all of these

measures must be taken into account to develop effective strategies, the following sections will focus on common techniques used to characterize bacterial pathogens, with increasing emphasis on the application and impact of high-throughput sequencing (HTS) technologies and bioinformatics methods, steering us towards the objectives of this dissertation.

Figure 1.2: WHO Bacterial Priority Pathogens List (BPPL), 2024 update. The BPPL includes 15 families of Antibiotic Resistant (ABR) pathogens, grouped into critical, high and medium categories of priority for RD and for public health measures. Adapted from [4].

1.2 Bacterial characterization

The characterization of pathogen strains is essential for the effective management of infected patients and the epidemiology of infectious diseases. Bacterial characterization or typing methodologies can be divided into phenotyping and genotyping [5]. The former characterizes bacteria based on phenotypic assays, such as colony morphology on various culture media, biochemical tests, serology, and antibiotic susceptibility. The latter distinguishes bacteria on the basis of their genetic content and has been increasingly adopted

to complement or substitute phenotypic assays. Genotyping allows inferring phenotypic characteristics through methods that are less complex and more broadly applicable than classical phenotypic assays, and in many cases has the potential to provide greater resolution. Microbiologists use both approaches to infer specific phenotypic characteristics, such as susceptibility to antimicrobial drugs, allowing to set the best course of action for patient treatment or mitigate the impact of an outbreak. Phenotypic assays rely on the expertise of clinical microbiologists who apply specialized and often species-specific techniques that were developed and optimized over years of research. These techniques involve complex and multi-step protocols that, depending on the bacterial species, can take less than a day or a few days (e.g. rapid-growing bacteria such as *Escherichia coli* [6]), to several weeks (e.g. slow-growing bacteria such as *Mycobacterium tuberculosis* [7]) [8]. The application of HTS, in combination with the development of specialized bioinformatics methods, has allowed researchers to reduce sample turnaround time by avoiding specific methodologies in favor of approaches such as WGS, which allows to accurately identify the genomic features associated with the phenotypic characteristics of interest from sequence data to provide equivalent results to more laborious and time-consuming lab protocols. Notwithstanding the impact of the latest developments in sequencing technologies and bioinformatics methods, there is still no single method for bacterial characterization that is universally ideal, with each method having to strike a balance between several desired characteristics, such as being applicable to all isolates, highly discriminatory at all levels, generating reproducible results at intra- and inter-laboratory level, while also using modest resources.

1.3 Phenotypic methods

Classical bacteriology methodologies are based on successfully isolating a bacterial pathogen on culture media. Given that different bacterial species may have different growth requirements, microbiologists had to develop a wide repertoire of techniques to account for all the variable requirements. After successfully isolating a pathogen, microbiologists may perform a series of assays to determine, for example, the species of the pathogen and its antimicrobial and virulence profiles (Figure 1.3). The culture step varies according to the complexity of a sample. For samples from usually sterile sites, such as cerebrospinal fluid, it may be possible to report all organisms present in the sample and it is simpler to identify the ones that are clinically relevant and should go through further analysis steps. In the case of complex samples, such as faeces, isolating the *micro culprit* may require a more custom approach guided by an educated guess about likely pathogens to select the appropriate media for culture and subsequent tests for a definitive diagnostic.

A correct species identification is highly informative, as it allows to deduce intrinsic characteristics from the body of knowledge and estimate the pathogenic potential, especially in the context of the isolation site. To identify the species of an isolate, microbiologists may use Gram staining, evaluate colony growth and morphology, and perform rapid biochemical

tests, such as a bile solubility test, which is used to differentiate *Streptococcus pneumoniae* from other alpha-hemolytic streptococci. Determining the biomolecule profiles of pure suspensions through matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) mass spectrometry and comparing them with known profiles is also used for rapid and cost-effective species identification, and to identify toxins and study bacterial antibiotic resistance [9–13].

Following culture and species identification, the determination of the antimicrobial resistance profile is crucial to select an effective treatment for infected patients. Antimicrobial resistance tests are mainly based on inhibition of *in vitro* bacterial growth when exposed to an antibiotic. The efficacy of testing methods, such as disc diffusion and E-TEST, is compared against gold-standard susceptibility-testing systems, such as micro-dilution, to infer *in vivo* efficacy. The level of susceptibility to a given antibiotic is based on the minimum inhibitory concentration (MIC) and on the definition of *breakpoints*, which correspond to the antibiotic concentration above which an isolate is considered to be resistant to therapy [8]. *Breakpoints* are defined based on various factors that are not necessarily universally agreed upon, making it difficult to accurately compare and assess the efficacy of susceptibility testing and associate it with clinical outcome. Moreover, it is important to note that the results of susceptibility testing may not translate into similar *in vivo* results, as resistance mechanisms may be more complex and depend on factors not adequately emulated by current susceptibility testing practices [8, 14].

Compared to antimicrobial susceptibility testing, the detection of virulence factors tends to be overlooked when selecting an effective treatment for patients. Nonetheless, knowledge of the virulence profile of pathogens can play an important role when the presence of a virulence factor is known to contribute significantly to pathogenesis and disease severity. For example, toxin-producing strains of *Clostridioides difficile*, the most common cause of nosocomial antibiotic-associated diarrhea [15], are pathogenic and may require differential treatment. Another example is the Panton-Valentine leukocidin (PVL) cytotoxin produced by some *Staphylococcus aureus* strains, which contributes to increased pathogenicity and disease severity, especially when combined with resistance to methicillin, and is associated with community-acquired severe necrotizing pneumonia and necrotic lesions of the skin and subcutaneous tissues [16–20]. In public health, virulence factors are especially important as vaccine targets. For example, the variability of the capsule polysaccharide of *Streptococcus pneumoniae* is detected by serotyping and epidemiologically relevant serotypes are targeted for vaccine development [21–24]. Serotyping is a method that uses antisera, which are sera containing antibodies against specific antigens, to identify and classify microorganisms based on their surface antigens. Serotyping provides enough discriminatory power to distinguish different strains of a pathogen, which is essential for public health interventions. Serotyping is a good example of a phenotypic assay that is routinely applied and for which there are sequence-based genotyping alternatives, generally applied after WGS to complement laboratory results or to determine the serotype of strains of interest when only sequence data is available. In the case of *S. pneumoniae*, *in silico* serotyping is possible through specialized

sequence databases[1] and software, such as SeroBA [25, 26]. Another example is serotyping of *Streptococcus pyogenes*, which measures the variability of the M protein, one of the targets of vaccine candidates currently in development [27]. The Centers for Disease Control and Prevention (CDC) hosts a database[2] with partial sequences of the M protein gene that is used by software such as emmtyper[3] for *in silico* serotyping.

Although genotypic methods may be seen as viable substitutes for phenotypic methods, it is highly unlikely that a complete substitution will ever occur, especially for many of the major human pathogens for which phenotypic methods have been standardized and provide highly reliable and cost-effective results. Genotyping is relevant to complement these methods for diagnostic purposes and to provide greater resolution for applications such as research into mechanisms such as virulence and antimicrobial resistance. Genotyping may play a more dominant role for known or emerging pathogens and for the characterization of complex samples that cannot be well characterized with current phenotypic assays, opening the possibility of genotyping emerging as the gold standard in those cases.

1.4 Genotypic methods

With the introduction of molecular methods for bacterial characterization, the basis for systematics changed. The distinction based on classical phenotypic criteria was complemented or in part replaced by molecular criteria, particularly molecular sequences, because these methods can offer greater resolution, resulting in more precise phylogenetic analyses and diagnostics. Molecular methods can be divided into three main categories: i) Deoxyribonucleic acid (DNA) banding pattern-; ii) DNA hybridization-; and iii) DNA sequencing-based methods. The first differentiate bacterial strains based on the size and pattern of DNA bands/fragments generated by DNA amplification or cleavage of genomic DNA using restriction enzymes (REs). The second uses techniques such as DNA macroarrays and microarrays, which distinguish strains through hybridization to probes complementary to known sequences. The third determine and compare the DNA sequence of genomic regions of interest, often determinant for a particular feature, to discriminate bacterial strains based on sequence variation.

1.4.1 DNA banding pattern-based methods

DNA banding pattern-based methods, either through amplification by Polymerase Chain Reaction (PCR) or digestion with REs, can provide accurate and quick results and are generalizable for characterizing strains of any bacterial species.

[1] `https://www.pneumogen.net/gps/#/serobank`
[2] `https://www.cdc.gov/strep-lab/php/group-a-strep/emm-typing.html`
[3] `https://github.com/MDU-PHL/emmtyper`

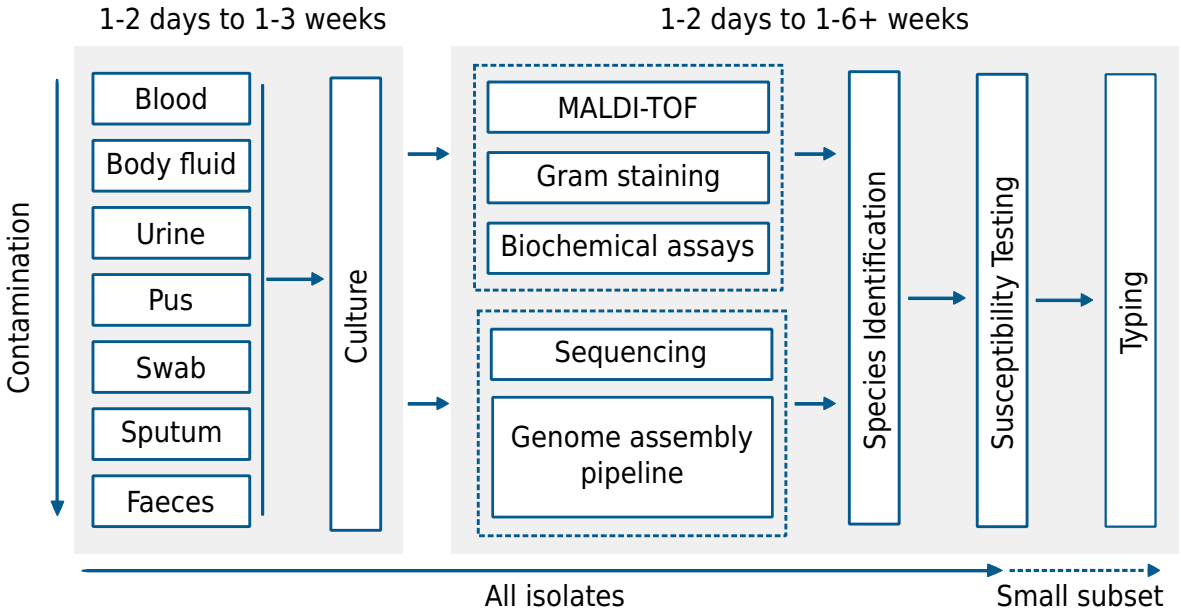


Figure 1.3: Schematic representation of the workflow for processing samples of bacterial pathogens. Samples collected from normally sterile body sites are cultured on a rich medium with the necessary nutrients to support bacterial growth. For more complex samples, such as faeces, which contain multiple bacterial species, selective media are used to favor the growth of the suspected pathogen. The growth of bacterial cultures can take from a day to several weeks depending on the growth requirements of the organism. After successful culture growth, methods such as MALDI-TOF, gram staining and other biochemical assays can be used for species determination, followed by susceptibility testing to determine the AMR profile. Depending on the species and the context of the infection, a small subset of the samples may be selected for further characterization using typing methods. Typing methods allow for a more detailed characterization of bacterial pathogens, which is especially useful in surveillance and outbreak investigation settings. DNA sequencing technologies are frequently used to substitute or complement phenotypic and molecular assays, allowing detailed characterization of bacterial pathogens through the analysis of the bacterial genomes with specialized bioinformatics methods. DNA sequencing is especially relevant as an alternative to reduce turnaround time when processing fastidious bacteria, for which susceptibility testing through more traditional methods can take from weeks to months. Adapted from [8, 28].

In this category of methods, one that allows to differentiate bacterial strains and has been widely applied in public health is Pulse Field Gel Electrophoresis (PFGE). PFGE is an electrophoretic technique that applies alternating electric fields at different angles to separate large DNA molecules ($10kb - 10Mb$) [29–31]. Prior to electrophoretic separation, REs that recognize uncommon motifs are used to cleave the bacterial DNA. The distinct banding patterns produced by PFGE reflect the DNA polymorphisms at the REs recognition sites and, ideally, can be uniquely associated to a specific bacterial strain. The resolution of PFGE depends on the choice of the REs used, with REs that recognize long and rare motifs yielding potentially more discriminatory results. The standardization of PFGE protocols and the creation of pattern databases, such as the one hosted by Pulsenet International[4], were crucial to the wide adoption of PFGE. Although widely used, PFGE is laborious and the results can be influenced by multiple factors, which hinders reproducibility and interoperability [5]. Due to these limitations and to the invention of HTS, PFGE, once considered the gold standard for bacterial typing [32], has gradually been replaced by more accurate and versatile methods based on WGS, which allow a much more detailed and increasingly cost-

[4] `https://www.pulsenetinternational.org/protocols/pfge`

effective characterization of bacterial strains based on the complete or nearly complete genome sequence.

Another method used to differentiate bacterial strains and infer relatedness is Restriction Fragment Length Polymorphism (RFLP) [33, 34]. RFLP allows to differentiate patterns of electrophoresis-separated restriction fragments by Southern Blotting with labeled probes [35]. The similarity of the patterns of Restriction Fragments (RFs) is the basis for strain differentiation. Ribotyping is a variation of RFLP that uses probes with conserved domains of ribossomal ribonucleic acid (rRNA) genes to differentiate strains based on variable regions flanking the bacterial rRNA operons [36]. The distinct banding patterns identified through this approach are named ribotypes. Since rRNA operons are universal, ribotyping is highly applicable. Furthermore, it generates fewer fragments that RFLP approaches based on frequently cutting REs, enabling easier interpretation of results and establishment of nomenclature and database systems. The potential cost-effectiveness of RFLP can be overturned by time- and labor-consuming protocols, as well as the requirement for large amounts of high-quality DNA, which is not always available. Ribotyping is an important method for the characterization and surveillance of *Clostridioides difficile*, with some ribotypes associated with greater disease severity. To overcome the limitations of RFLP specifically applied for ribotyping of *C. difficile*, prediction directly from WGS data can be performed with bioinformatics methods that estimate sequence similarity, such as sourmash [37, 38], or using machine learning [39].

It would be unacceptable to move to the next category of molecular methods without mentioning the Swiss Army Knife of molecular biology: PCR. The PCR method, originally developed by Kary Mullis, allows the amplification of any target DNA sequence in a sample in a cyclic process to generate a large number of copies of the target sequence [40]. PCR is performed by temperature cycling, with each cycle having three stages: denaturation, annealing, and elongation. Firstly, high temperature is applied during the denaturation stage to separate the DNA strands. Secondly, the temperature is lowered to allow for the annealing of two oligonucleotide primers that flank the target sequence. Lastly, the temperature is raised to the optimum level at which a heat-stable polymerase can extend the primers by incorporating deoxynucleotides (dNTPs). Over the years, a plethora of PCR-based methods were developed to expand the applicability and overcome limitations of the original PCR, firmly establishing PCR as one of the fundamental methods in molecular biology. From the vast number of PCR-based methods that were invented, some, such as multiplex PCR and Real-Time Quantitative PCR (qPCR), are broadly applicable. Multiplex PCR uses multiple primer pairs to simultaneously amplify multiple target sequences in the same PCR reaction [41]. qPCR allows to simultaneously amplify target sequences and detect the PCR product [42, 43]. This is achieved by incorporating and monitoring the fluorescence of dyes or probes, which increases proportionally to the amount of product formed. qPCR overcomes challenges related to product quantification in the original PCR and allows for faster confirmation of the presence of a target sequence. Other PCR-based methods were developed for more specific tasks, such as genotyping. Methods such as arbitrarily primed PCR (AP-PCR or RAPD) [44, 45], which uses arbitrary primers for random amplification, and repetitive sequencing-

based PCR (REP-PCR) [46, 47], which amplifies regions between interspersed repetitive elements, generate fragment patterns that can function as signatures for specific bacterial strains. The identification of variable number tandem repeats (VNTR) in bacterial genomes through multiple-locus variable number tandem repeat analysis (MLVA) enables to identify polymorphic sites [48]. VNTR elements evolve rapidly and the number of tandem repeats per locus may vary between strains. MLVA uses PCR to amplify multiple VNTR loci, followed by analysis of the banding pattern to assign a genotype and infer phylogenetic relationships. Although MLVA is a simple technique and may offer high resolution, VNTR loci in closely related strains may evolve quickly, hindering long-term surveillance. Additionally, VNTR may not be common in some species, which limits its applicability, and the accuracy of MLVA might be affected by insertions or deletions in the amplified regions. Some PCR-based methods combine PCR with other typing methods, such as methods that use REs, to overcome limitations and improve accuracy. One example is PCR-RFLP, which can amplify target regions directly from clinical or environmental samples and uses REs digestion of the PCR amplicons to generate a limited number of RFs that are more easily separated by gel electrophoresis and interpreted [49]. PCR-based methods display multiple advantages, such as being relatively inexpensive, fast, and sensitive. Notwithstanding, researchers should be mindful about the inherent limitations of each PCR-based method, and of limitations which are common to most PCR-based methods, such as the potential for contamination, artifacts caused by, for example, non-specific amplification and primer dimerization, and the need for multiple controls. PCR primer design is facilitated by tools such as Primer-BLAST[5], made freely available by the National Center for Biotechnology Information (NCBI) [50]. Multiple bioinformatics tools[6,7,8] implement PCR-like functionalities to search for sequences of interest based on flanking regions. *In silico* PCR is especially valuable to assess primer specificity in a range of applications and to identify highly variable regions whose detection is suboptimal with more common sequence comparison techniques such as alignment [51]. PCR also plays a crucial role in DNA sequencing workflows, such as being used for clonal template generation to boost signal detection and the amplification of loci of interest for target sequencing experiments. A brief overview of DNA sequencing technologies is presented in Subsection 1.4.3.

1.4.2 DNA hybridization-based methods

DNA hybridization-based methods use probes, which correspond to known DNA fragments, to detect complementary DNA sequences extracted from samples [52]. A variant of these methods uses DNA arrays to test for the presence of hundreds to tens of thousands of DNA fragments, being relevant for applications such as the study of the genetic diversity of

[5] https://www.ncbi.nlm.nih.gov/tools/primer-blast/
[6] https://www.gear-genomics.com/silica/
[7] https://bigsdb.readthedocs.io/en/latest/data_analysis/in_silico_pcr.html
[8] https://ucsc.gao-lab.org/cgi-bin/hgPcr

bacteria and in transcriptomics. Two types of DNA arrays exist: macroarrays [53, 54] and microarrays [55–57]. The former can contain up to five thousand spots, providing enough resolution to detect genes involved in AMR or for typing methods based on the detection of polymorphisms in a smaller number of loci, such as spoligotyping for Mycobacterium tuberculosis complex (MTC) bacteria. DNA microarrays are more expensive, but provide far greater discriminatory power, including up to tens of thousands of distinct probes, enabling the identification of a greater number of loci compared to macroarrays or to study the variation at the genome or transcriptome level. Since the probes included in microarrays are defined based on reference sequences, microarrays may lack probes complementary to accessory genes, leading to an underestimation of genetic diversity. DNA microarrays can use complementary DNA (cDNA) or shorter oligonucleotides as probes. The former are used to determine gene presence, while the latter is capable of detecting smaller patterns of variation, such as deletions or even Single Nucleotide Polymorphisms (SNPs). The use of DNA arrays has largely been supplanted by the use of HTS [58].

A DNA hybridization-based method that has been applied to overcome some limitations of the classic culture-dependent approach is target capture through hybridization using oligonucleotide probes. In this approach, specialized bioinformatics software is used to design a set of probes to capture bacterial DNA directly from clinical samples for subsequent HTS [59, 60]. Bypassing the culture step is especially useful in reducing the turnaround time for fastidious bacteria such as *M. tuberculosis* [61], or to capture the DNA of uncultivable bacteria, such as *Treponema pallidum*, the causative agent of syphilis [62]. In addition, target capture has the potential to provide a less biased view of within-host variation by allowing to capture DNA from multiple strains, contrasting with culture-dependent approaches, which may not meet the growth requirements of specific strains and often select a single colony from the culture plate for further characterization. In combination with HTS, target capture allows culture-independent sequencing of clinical samples for faster diagnostics and detailed bacterial characterization. Designing probes to capture taxa at multiple taxonomic levels is also a powerful approach to study complex clinical samples and for metagenomics studies, allowing capture of known and related sequences and minimizing the obscuring effect of host DNA abundance [59]. The success of target capture is highly dependent on multiple factors, such as the specificity of the designed probes, hybridization efficiency, and the abundance of the target microorganisms in the samples [61–63].

1.4.3 DNA sequencing technologies

The advent of DNA sequencing technologies represented a major milestone in biological research, finally unlocking the genetic information encoded in DNA, which had already been established as the source of genetic information in 1944 by Oswald Avery while working with *Streptococcus pneumoniae* [64] and whose three-dimensional structure was determined in 1953 by Watson and Crick based on the crystallographic data produced by Rosalind

Franklin and Maurice Wilkins [65, 66]. The potential and continuous improvement of these technologies contributed to their adoption and led to the development of highly reproducible and accurate methods used to differentiate bacterial strains and identify determinants of phenotypic features of interest. As sequencing throughput and sequence data availability increased, the diverse and highly dynamic nature of bacterial genomes was unveiled, leading to an unprecedented interest in developing sequence-based methods that could probe into the accumulated sequence data to gain new insights. The application of HTS is revolutionizing our understanding of human health and disease by elucidating fundamental biological and ecological processes.

1.4.3.1 First-generation DNA sequencing

The first major breakthrough in DNA sequencing technologies was made by Fredrick Sanger et al. in 1977 [67]. Sanger sequencing, also called dideoxy sequencing or chain termination DNA sequencing, determines the nucleotide sequence of a single-stranded template DNA using a DNA polymerase to synthesize nucleotide fragments of different lengths by incorporating radio or fluorescently labeled dideoxynucleotides (ddNTPs) and through premature termination of the DNA amplification elongation step [68, 69]. The truncated fragments resulting from the interruption of the elongation step are size-separated by capillary gel electrophoresis to reconstruct the original sequence. As the first successful DNA sequencing technology, Sanger sequencing was instrumental in projects such as the sequencing of the first bacterial genome, the genome of *Haemophilus influenzae* [70], and the Human Genome Project, which in 2003 achieved the monumental task of determining the first nearly complete sequence of a human genome [71]. Sanger sequencing was the most widely used sequencing technology until newer and cheaper HTS technologies were developed.

1.4.3.2 Second-generation DNA sequencing

The second revolution in DNA sequencing began in the early 2000s when the first massive parallel sequencing technologies were made commercially available. Pyrosequencing [72–74] was the first second generation SBS technology to reach the market (Figure 1.4). The general principles of second generation SBS technologies are the following: i) attachment of the DNA to be sequenced to a solid support, usually combined with amplification to enhance signal detection; ii) single-stranded DNA synthesis; iii) primer-dependent incorporation of complementary bases; iv) detection of each incorporated nucleotide for sequence determination. Pyrosequencing is based on real-time quantitative detection of pyrophosphate released as nucleotides are incorporated into a growing DNA sequence, yielding reads around 400-500 base pairs (bp) long. Initially, libraries of DNA molecules are attached to paramagnetic beads via adapter sequences and amplified through emulsion PCR. Ideally, on average only one DNA molecule attaches to each bead so that each bead is coated in a clonal DNA population

after emulsion PCR.

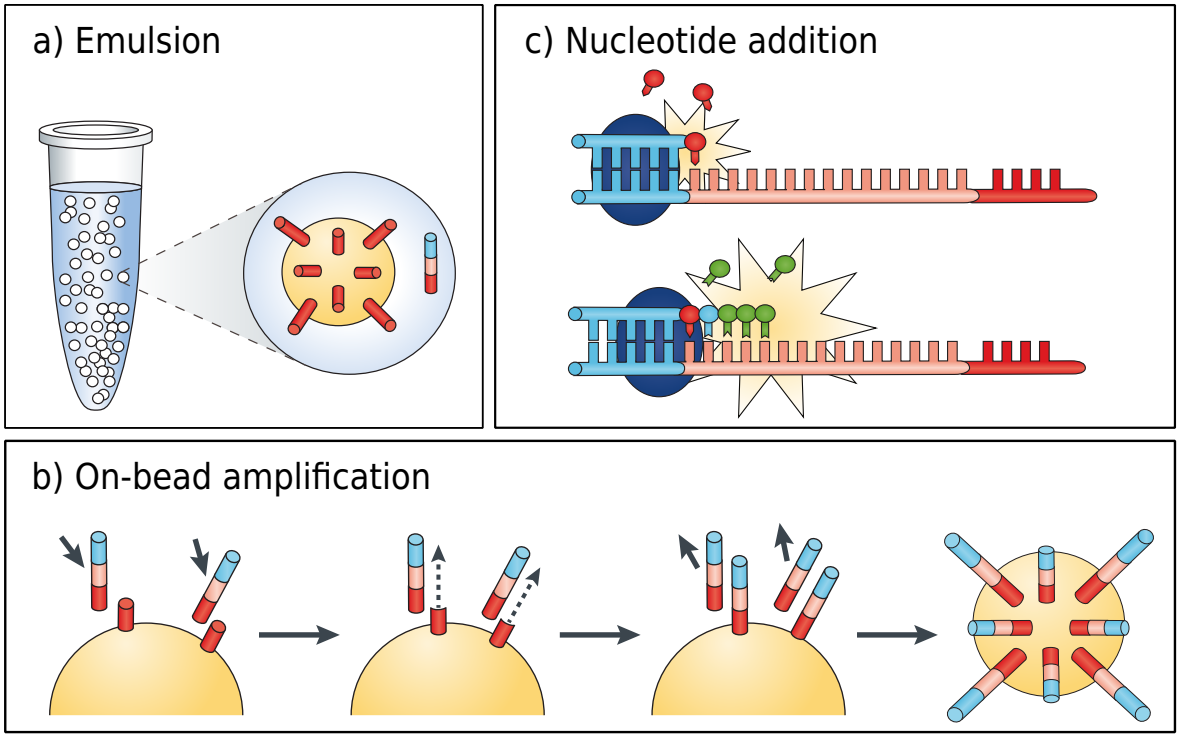


Figure 1.4: 454 pyrosequencing. In 454 pyrosequencing, fragmented DNA templates are amplified through emulsion PCR (**a**), which consists on the hybridization of the DNA templates to bead-bound primers followed by amplification to cover each bead in thousands of copies of the same DNA sequence (**b**). The beads are arrayed onto a microtitre plate along with primers and different beads that contain enzymes. Sequencing occurs in cycles. In each cycle, a single nucleotide species is added and nucleotides are incorporated into growing chains by a DNA polymerase. When a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light. Each burst of light is detected by a device to determine the bases incorporated at a particular bead, with the possibility of incorporating multiple bases of the same type in a single cycle (**c**). Adapted from [75, 76].

The DNA-coated beads are distributed into a plate that fits one bead per well where pyrosequencing occurs as bead-linked enzymes and dNTPs are added and the pyrophosphate release is detected by a sensor [68, 77–79]. Compared to conventional Sanger sequencing, pyrosequencing has much higher throughput with a fraction of the cost, making it easier and more viable to scale-up. It does have some limitations over Sanger sequencing, however, as the read lengths are shorter, making downstream analysis, such as assembly, more complex. The most used pyrosequencing technology was 454 sequencing, which had major advantages over traditional Sanger sequencing, as demonstrated by its application to investigate drug resistance in *Mycobacterium tuberculosis* [80] and the whole genome sequencing of Jame Watson's genome in record time and within a fraction of the cost of the Human Genome Project [81, 82]. 454 sequencing was eventually discontinued in favor of more accurate and advanced technologies, such as Illumina's SBS. Further improvements to massive parallel SBS were introduced with the development of reversible and fluorescently labeled terminators [83]. The most widely known sequencing strategy that incorporated these improvements is Illumina's SBS (Figure 1.5) [84–86]. Illumina's SBS systems enable massive parallel sequencing (MPS) of small DNA fragments, yielding sequencing reads with up to 300 bp.

Illumina's SBS technology starts by binding adapter sequences to the DNA libraries, which contain complementary sequences that bind to the flow cell, unique indexes or barcodes for sample identification, and the sequencing primer binding sites. The DNA molecules attached to the flow cell undergo bridge amplification to generate clonal clusters. Sequencing is performed in cycles by using 'reversible-terminator' dNTPs and detecting the fluorescence resulting from the incorporation of nucleotides before proceeding to the next cycle. After sequencing the forward strand in this manner, Illumina's systems are capable of sequencing the reverse strand, generating paired-end (PE) data, which significantly improves the accuracy of downstream analysis. The advantages of Illumina's SBS systems, such as their high throughput, lower costs and accurate base calling, led to their worldwide adoption and establishment as the dominant sequencing technology for projects of any scale. The higher throughput of Illumina's SBS, especially as sequencing costs decreased, led to an explosion of sequence data, effectively pushing research into the era of *big data* and data-driven research [84]. The tremendous increase in the number of microbial genomes deposited in public databases in recent years is in great part due to the widespread application of second-generation sequencing technologies, especially of Illumina's SBS. Other examples of the successful application of this technology are the sequencing of 25,000 cancer genomes by the Cancer Genome Consortium [87] and the 100,000 Genomes Project [88].

1.4.3.3 Third-generation DNA sequencing

Advances in the early 2010s led to the third and current revolution in DNA sequencing. The third-generation of sequencing technologies provide single-molecule sequencing and eliminate the requirement of DNA amplification characteristic of second-generation sequencing technologies. Currently, the most successful technologies are HiFi sequencing from Pacific Biosciences (PacBio) [89, 90] and Nanopore sequencing from Oxford Nanopore Technologies (ONT) [91, 92]. HiFi sequencing (Figure 1.6.a) works by creating circularized DNA libraries that are sequenced in repeated passes to generate several subreads per DNA molecule, which can be compared to determine a consensus read minimizing sequencing errors. HiFi sequencing occurs inside small wells on a Single Molecule Real-Time Sequencing (SMRT) Cell microchip where DNA extension with fluorescent dNTPs is finely monitored. The sequencing technology developed by ONT (Figure 1.6.b) passes single-stranded DNA (ssDNA) through a biological nanopore embedded in a synthetic membrane, across which a voltage is applied. The passage of the ssDNA through the nanopore limits ionic flow and induces a current change for a period of time that allows to infer the sequence of the ssDNA traversing the nanopore. Both technologies generate reads with length that can far exceed the length of the reads generated by second-generation technologies, which is why they are also called long-read technologies. With HiFi sequencing, read lengths can reach 1 to 25 kilobases (kb). Nanopore sequencing is capable of generating even longer reads, from a few to more than a hundred kb, and up to several megabases (Mb). Accuracy-wise, HiFi sequencing has the high ground, but nanopore sequencing provides faster results and greater portability, crucial in out-

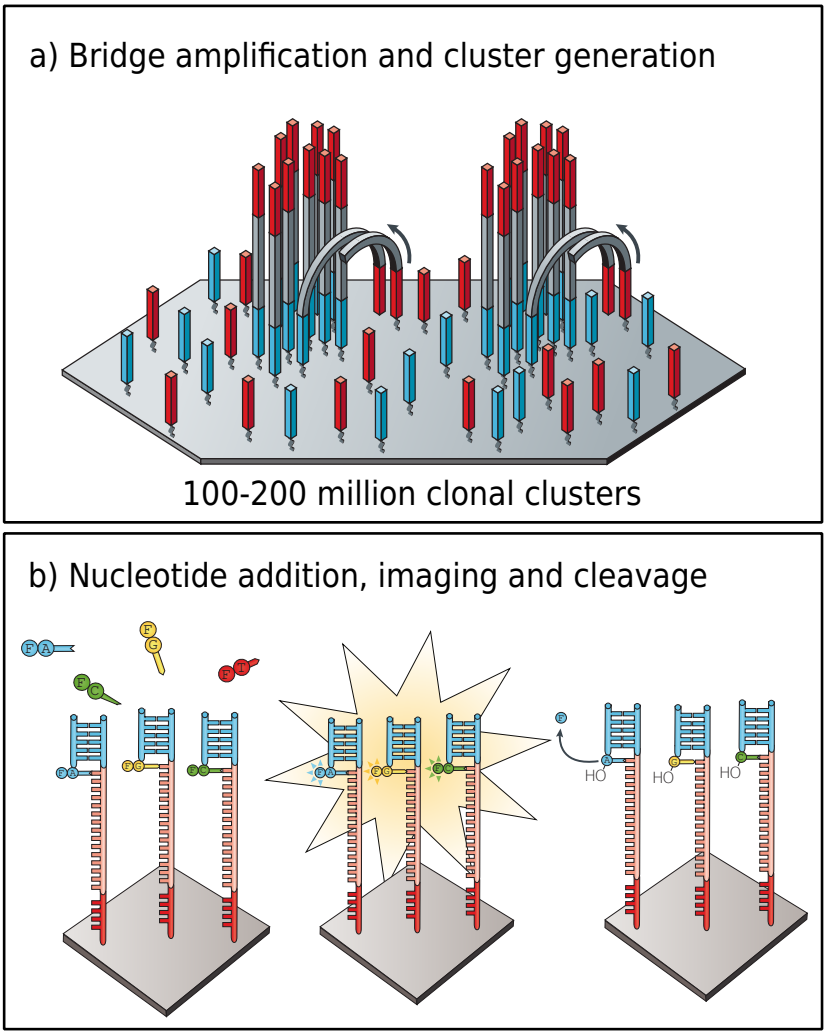


Figure 1.5: Illumina's SBS. In Illumina's SBS, DNA templates hybridize with adapters bound to a flow cell and are amplified though bridge-PCR to generate millions of clonal clusters (**a**). In each cycle, fluorophore-labelled and terminally-blocked nucleotides are added and hybridize to complementary bases. Flow cells are imaged to measure the color of the emitted light when a base is incorporated (**b**). Fluorophores are cleaved and washed before a new cycle begins. Adapted from [75, 76].

break investigation settings for fast pathogen detection and characterization. In addition, the lower entry cost, the development of new nanopores, base calling software, and experimental protocols tailored to particular applications have contributed to a gradual improvement in the accuracy of nanopore sequencing [93–96]. Despite continuous improvements in long-read sequencing technologies, their error rate is still higher than that of short-read sequencing technologies, such as Illumina's SBS. For this reason, long- and short-read data have been used in combination for so-called hybrid approaches to overcome the limitations of both sequencing technologies and obtain higher-quality genome assemblies. When applied to the assembly of bacterial genomes, this strategy allows to assemble complete and error-free genomes, which cannot be achieved by using any of the technologies separately [97–99].

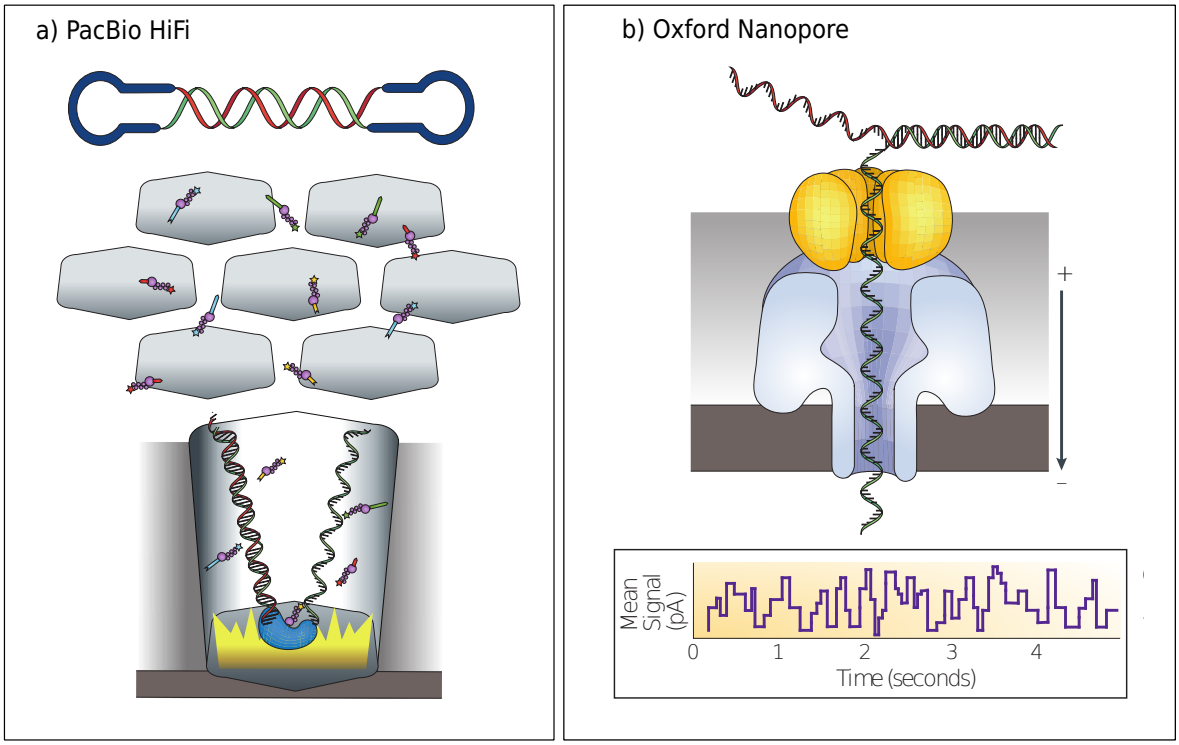


Figure 1.6: PacBio HiFi and Oxford Nanopore sequencing. In HiFi sequencing (**a**), two hairpin adapters are added to the DNA templates to allow for continuous circular sequencing. HiFi sequencing occurs inside wells where labelled dNTPs are incorporated and a camera records the emitted light to determine the incorporated bases. In Nanopore sequencing (**b**), a leader adapter is bound to the DNA templates and interacts with a motor protein and a biological nanopore, directing the DNA into the pore. As the DNA translocates the pore, a shift in voltage is measured to determine the composition of the DNA sequences. Adapted from [75, 76, 93, 100].

1.4.4 DNA sequencing-based methods

As HTS technologies became more accurate and cost-effective, wide adoption by research and public health institutions became a possibility. The application of these technologies to help resolve infectious disease events, such as the cholera epidemic in Haiti after the 2010 earthquake [101] and the international outbreak of *Escherichia coli* disease linked to contaminated fenugreek sprouts [102, 103], quickly revealed that they were an invaluable tool for the surveillance and outbreak investigation of bacterial pathogens. WGS of bacterial isolates, performed with second or third generation sequencing technologies, followed by genome assembly, allows the determination of the complete or nearly complete genome sequence, which in principle encodes most of the genetic features necessary for a detailed characterization of an isolate. Genome assembly is performed with pipelines such as Shovill[9], INNUca[10] [104–106], and Bactopia[11] [107], typically based on a *de novo* approach to determine a set of contiguous sequences, called *contigs*, resulting from the comparison and combination of overlapping sequencing reads. The application of specialized bioinformatics software allows identifying the relevant features for typing and diagnosis based solely on a

[9] `https://github.com/tseemann/shovill`
[10] `https://github.com/B-UMMI/INNUca`
[11] `https://github.com/bactopia/bactopia`

sequence approach, complementing or replacing traditional microbiological workflows [108, 109]. Moreover, WGS may also allow for the identification of emerging genetic features not tested for in routine molecular tests and the detection of uncultivable bacterial strains [109].

The surveillance of food-borne diseases has greatly benefited from the implementation of standardized WGS-based systems. An estimated 600 million people fall ill due to contaminated food annually, resulting in over 400 thousand premature deaths [110]. This puts global public health systems under strain and leads to significant costs related to medical treatment and to productivity and trade losses. Initial reports on the adoption of WGS by the PulseNet surveillance network in 2000 demonstrated improved outbreak detection and an increase in the number of solved outbreaks compared to using PFGE data [108, 111, 112]. The gradual adoption of WGS by the network participants and the standardization of analytical workflows established WGS as the gold standard for typing of foodborne pathogens tracked by the network. In 2019, the European Centre for Disease Prevention and Control (ECDC) published a strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigation [113]. The document included a progress report on the implementation of WGS for surveillance and outbreak investigations by the European Union and the European Economic Area (EU/EEA) member states and outlined key technological milestones to improve the surveillance and outbreak detection of priority pathogens/diseases. Joint work of the ECDC and European Food Safety Authority (EFSA) resulted in the implementation of the EFSA and ECDC One Health WGS System, which aims to augment surveillance capacity and coordination among EU/EEA member states [114]. This system has been instrumental in detecting and resolving multiple multi-country outbreaks [115–118].

Building WGS capacity requires significant investment in sequencing instruments, reagents, and computational resources to store and analyze WGS data. The choice of analytical methods is crucial and may be especially complex as there are multiple fundamental approaches which are not necessarily equivalent or comparable [119]. Multiple bioinformatics methods have been developed for the detailed characterization of bacterial strains. Some methods characterize strains based on the identification of a single locus, such as *in silico* serotyping or abundance estimation based on the 16S rRNA gene. Other methods offer greater resolution by identifying and measuring the variability of a greater number of loci, such as GbG, Single Nucleotide Polymorphism (SNP)-based, and *k*-mer-based methods. These methods have been increasingly integrated into WGS-based systems for surveillance, outbreak investigation, and the study of bacterial populations. For this reason and because of their superior potential for further improvement, they are more relevant to the work developed in this dissertation, and a greater focus is given to these methods in the following sections.

1.4.4.1 Multilocus sequence typing

Multilocus Sequence Typing (MLST) is a sequence-based approach that uses allele fragments, typically seven, from housekeeping genes to characterize microorganisms, with a more expressive application for bacterial species of pathogenic potential. MLST is based on the principles of multilocus enzyme electrophoresis (MLEE), but uses nucleotide sequences at each locus, taking advantage of developments in sequencing technologies and bioinformatics (Figure 1.7) [120]. Moreover, MLST allows identifying a larger number of alleles per locus, offering higher discrimination than MLEE while using a smaller number of loci. MLST was initially developed to better accommodate vertical and horizontal genetic transfer signals by targeting multiple well-conserved genes and to overcome the challenges of traditional and molecular typing methods, such as the inability to infer strain relatedness and poor reproducibility within and between laboratories [121].

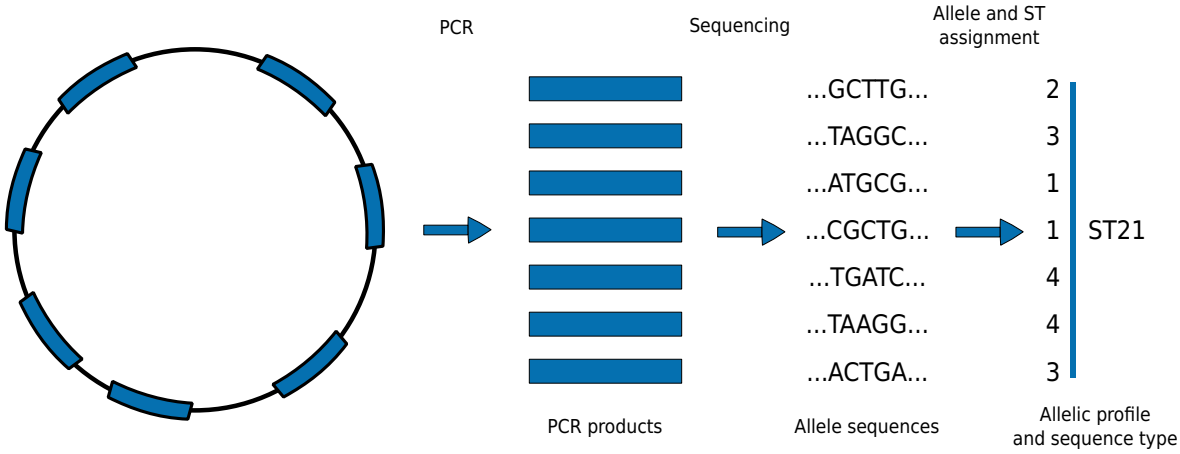


Figure 1.7: Multilocus Sequence Typing (MLST). In MLST, the internal fragments of a set of housekeeping genes, typically seven, are amplified through PCR and sequenced. Each distinct sequence is assigned an allele identifier, and the combination of allele identifiers constitutes an allelic profile. Each distinct allelic profile is assigned a Sequence Type (ST), which allows to identify groups of similar strains.

The distinct fragments identified at each locus are assigned unique integer identifiers in order of discovery, and the combination of identifiers for the allele fragments identified in all loci constitutes an allelic profile, which can be compared against a database of known allelic profiles. Each distinct allelic profile unambiguously defines a ST, assigned to isolates for direct comparisons. STs are grouped into clonal complexes (CC), a concept first introduced to describe *Neisseria meningitidis* isolates analyzed by MLEE, based on their similarity to a central ST (allelic profile or genotype). The definition of central STs is achieved through a combination of computational and experimental data obtained from public health authorities. Newly identified STs are assigned to the most similar CC based on a minimum number of shared alleles with the central ST. ST organization into CCs facilitates epidemiological analysis, often grouping most STs into a much smaller number of CCs and allowing to identify CCs of greater clinical relevance. One disadvantage of MLST is that it may not offer the same degree of discrimination within lineages or species with highly conserved housekeeping genes. Furthermore, due to the diversity of bacterial species, MLST schemes must be developed

to distinguish closely related bacteria, usually at the genus and species levels, or they may not provide sufficient resolution. Consequently, MLST cannot be applied as a combined taxonomic and typing approach at all levels of bacterial diversity [122].

By relying on the sequencing of allele fragments from multiple chromosomal locations, MLST provides unambiguous results and is more robust to recombination events, constituting a faster and more sensitive technique than most laborious lab protocols, which also tend to be more unpredictable as variation accumulates. Since allele fragments are used as a unit of comparison, single allele differences constitute a single event, regardless of the number of nucleotide polymorphisms involved. While this model may not provide resolution for every single point change, it is resistant to horizontal genetic transfer events, which introduce a lot of variation in a single event, leading to an inaccurate estimate of similarity if counted as multiple differences.

MLST aims to provide good discrimination for short- and long-term epidemiology. The original study showed that it was congruent and more discriminatory than MLEE in distinguishing hyper-virulent strains of *N. meningitidis* while also offering a clear distinction between lineages at the species level [123]. A subsequent study presented a MLST database for *Streptococcus pneumoniae*, obtaining consistent results with MLEE and PFGE for the analysis of predominantly invasive and antibiotic-resistant isolates [124]. Moreover, MLST was also congruent with serotyping, with isolates that share the same or similar STs also expressing the same serotype, except for cases where recombination at the capsular locus was suspected to have led to capsular switching [123].

Numerous MLST databases have been developed since MLST was proposed. Currently, a public collection of curated and frequently updated MLST databases is available for a great number of microbial species on the PubMLST website[12]. PubMLST integrates sequence data with sample metadata to promote the exchange of molecular typing data for epidemiological studies. As of 27 March 2025, PubMLST manages more than 130 species and genera-specific MLST databases, which contain tens of millions of alleles identified and submitted by researchers. The sheer volume of data and the range of databases in PubMLST highlight how the advantages of MLST contributed to its rapid adoption worldwide, with the technique widely used for epidemiological studies, to identify localized disease outbreaks and monitor local and global trends, and for population studies, to examine the structure of bacterial populations and perform evolutionary analyzes.

1.4.4.2 rMLST

Greater resolution and wider applicability than MLST are achieved with Ribosomal MLST (rMLST). rMLST typing indexes the variation of the genes encoding the bacterial ribosomal protein subunit (rps). The rps genes are ideal targets for universal bacterial characterization

[12] `https://pubmlst.org/`

because they are: i) universally present; ii) distributed across the genome, which makes rMLST more robust against horizontal gene transfer events that reassort loci and break phylogenetic congruence; and iii) encode proteins which are functionally conserved across the Bacteria domain. rMLST constitutes a combined taxonomic and typing approach for the whole domain of Bacteria at all taxonomic levels. rMLST allelic profiles or ribosomal STs (rSTs) determined through rMLST provide a basis for universal bacterial systematics, allowing for a precise identification of the phylogenetic position at any taxonomic rank, while also distinguishing closely-related strains for typing purposes. A database for the 53 rps genes identified in bacteria is managed by the Bacterial Isolate Genome Sequence Database (BIGSdb) platform [125].

1.4.4.3 wg/cgMLST

The level of resolution for typing depends on the desired application. On the one hand, higher resolution is necessary for the detection of outbreaks and within-patient variation. On the other hand, lower resolution is required to group strains into CCs or lineages. The GbG approach is inherently hierarchical and scalable, meaning that the number of genes used in the analyzes can be adjusted based on the desired resolution [125]. Thus, the concept and analysis methods of the highly successful seven-gene MLST can be intuitively scaled to hundreds or thousands of genes to encompass diversity at the core- or whole-genome level, giving rise to wg/cgMLST (Figure 1.8). wg/cgMLST provides higher resolution for surveillance and outbreak investigation. Furthermore, the additive nature of MLST, through the continuous update of schemas with novel alleles, ensures that wg/cgMLST can provide accurate results in the long-term while also promoting interoperability.

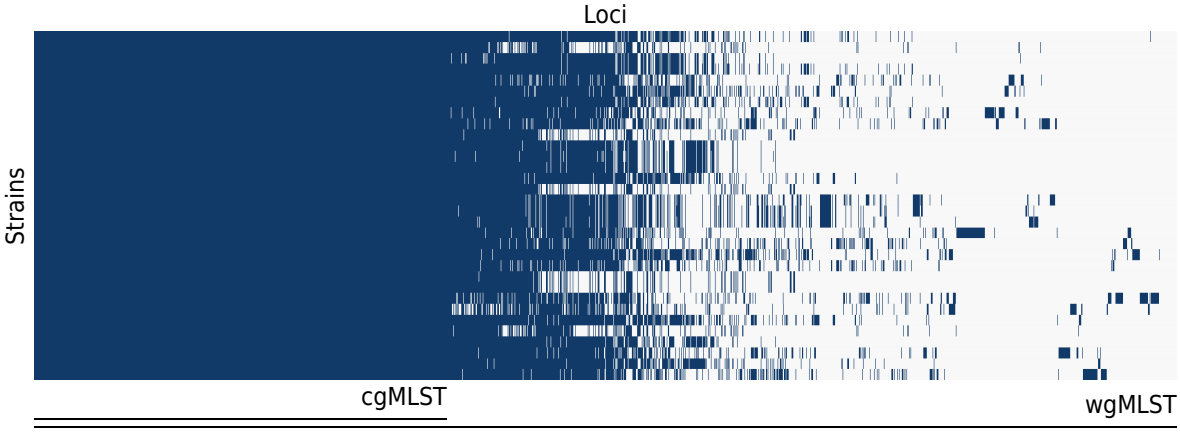


Figure 1.8: Whole- and core-genome MLST (wg/cgMLST). The heatmap represents a presence-absence matrix of the loci (columns) in a wgMLST schema that were identified in 32 bacterial strains (rows). Darker regions indicate that a locus was identified in a strain, while lighter regions indicate locus absence. cgMLST compares the set of alleles for the loci present in all strains (shorter line below the heatmap), while wgMLST incorporates loci from the accessory genome (longer line below the heatmap), providing greater resolution.

cgMLST characterizes bacterial strains based on the identification and comparison of the genes that constitute the core genome. Although the core genome is often defined as the set of

genes present in all strains of a given dataset, a more relaxed definition is needed to account for technical and biological variation. A loci presence threshold of 95% is commonly used to accommodate biases and errors introduced by, for example, the sequencing and genome assembly processes. This allows to retain genes that are present in almost all strains of a species or that are reported as absent due to misassembly. Furthermore, the set of core loci is usually determined based on the allele calling results for a specific dataset ideally representative of the diversity of the species. This means that the definition of core genome is highly dependent on the dataset used and the number of core loci detected varies according to the species and the dataset.

wgMLST further expands the set of loci used for strain typing by including loci from the accessory genome. The frequency of accessory loci is highly variable, with some accessory loci being nearly as frequent as core loci, and others being found only in a small subset of strains or even being strain-specific. In theory, a wgMLST schema should contain more loci than a cgMLST schema, but there are no hard requirements regarding the fraction of core and accessory loci of a species that should be included, allowing for wgMLST schemas with a number of loci close to that of a cgMLST schema or considerably larger wgMLST schemas encompassing all known core and accessory loci for a species. Since wgMLST identifies more loci than cgMLST, it provides greater resolution and is potentially more discriminatory when there is variation in the accessory genome. However, creating wgMLST schemas requires a more careful selection of target loci compared to cgMLST schemas. While cgMLST schemas include only core loci, which usually display lower allele diversity, the inclusion of accessory loci in wgMLST schemas increases the frequency of spurious loci due to sequencing and assembly errors or real sequence variability, affecting the accuracy of the results. Since cgMLST provides robust results and most of the available schemas are cgMLST schemas, most analyses are performed at that level, with wgMLST being recommended when higher resolution is necessary, such as when comparing closely related strains for surveillance and outbreak detection [119, 126–128].

wg/cgMLST characterizes strains by determining their allelic profiles (i.e., the set of loci and alleles identified in each strain). The comparison of the allelic profiles to determine the number of shared loci and alleles provides an estimate of strain similarity. This can be done by discarding missing loci from the analysis or by computing the absolute difference. The former is preferred, as genome fragmentation and potential sequencing and assembly errors make it impossible to be certain if a locus that was not identified is in fact absent from the genome. The cross-comparison of a set of samples allows to compute a distance matrix including the number of allelic differences between each pair of strains. The distance matrix enables phylogenetic analysis through methods such as single-linkage clustering, Neighbor-Joining (NJ) or by computing a Minimum-spanning-tree (MST). Computing a MST is frequently used in surveillance and outbreak investigation scenarios as it provides accurate results when comparing closely related strains. For the study of more diverse populations, more robust results can be obtained using methods such as maximum likelihood after computing a Multiple Sequence Alignment (MSA) for the alleles identified in all strains

for each locus and concatenating the loci MSAa to obtain a core genome MSA. Multiple software and web platforms provide functionalities to generate and visualize trees from wg/cgMLST results, including options to overlay the tree with metadata to more easily identify relevant strains. A distance threshold can be defined to identify groups of highly similar strains corresponding to lineages or potential outbreaks. Threshold definition is often empirical, depending on the diversity of the species, dataset, context, and methods used for the analysis. Consequently, threshold values are not universally applicable, which hinders comparability of the results obtained in different settings (e.g., outbreak detection in different countries.

1.4.4.3.1 wg/cgMLST platforms

There are multiple web platforms that store and manage wg/cgMLST schemas. These platforms provide access to wg/cgMLST schemas for a wide range of species and offer different functionalities for data analysis. All well-established wg/cgMLST platforms centralize data analysis by requiring users to upload their data. Platforms such as BIGSdb[13] and Enterobase[14] operate under more permissive licenses, providing wide-access to schemas and functionalities upon registration and allowing other users or institutions to set up their own instances of the platform. Other platforms, such as Ridom SeqShere+[15], are proprietary software that requires users to pay for access to the platform's functionalities, although schemas used within the platform are publicly available. The results generated within different platforms are not directly comparable, as schemas for the same species stored by different platforms may target different sets of loci and use distinct loci and allele nomenclatures, which hinders interoperability. The results are not easily comparable even when the schemas have the same origin and use the same nomenclature, since each platform applies different methods for allele identification and the nomenclatures are not synchronized, which means that different identifiers can be assigned to the same allele depending on the platform.

BIGSdb was the first platform to enable wg/cgMLST and is a prime example of a solution for wg/cgMLST that has been widely adopted and offers extensive analytical capabilities. BIGSdb extended the functionalities for MLST of the PubMLST platform to WGS data [129, 130]. BIGSdb pioneered the application of GbG methods to genome analysis, storing genomic and gene sequences, as well as associated metadata, such as provenance and phenotypic data for isolates from which the sequence data originated. Additionally, it stores allele and locus definitions, without an inherent limit to the number of records or the number of schemas into which the loci can be grouped. The loci included in a schema do not need to be associated with a single organism, enabling the creation of schemas that encompass the diversity of genes, such as accessory genes, that are distributed in diverse organisms. Known and novel alleles

[13] `https://pubmlst.org/software/bigsdb`
[14] `https://enterobase.warwick.ac.uk/`
[15] `https://www.ridom.de/seqsphere/`

are identified from sequence data uploaded to BIGSdb to maintain a record of the known diversity of genes identified in the samples stored in the database. Furthermore, genomic data are periodically rescanned as the database expands to identify variants in stored isolates that could not be detected previously based on the represented allele diversity in the database. The functionalities included in BIGSdb allow users to link isolate and sequence data with great flexibility, allowing the definition of schemas encompassing the diversity of species with utility for epidemiological investigations and population analysis or smaller schemas to study particular aspects of the biology of an organism. The genetic nomenclatures established and maintained by BIGSdb enable the definition of classification hierarchies for an effective comparison of bacterial isolates globally [130].

1.4.4.4 SNP-based methods

An accurate estimation of strain similarity for phylogenetic analyzes can be achieved by comparing the genomes of strains of interest against the genome of a reference strain to identify SNPs. This is performed by mapping sequencing reads against a reference genome to identify all variable positions in regions shared with the reference, identifying the set of core SNPs (Figure 1.9). Similarly to wg/cgMLST, pairwise SNP distances or the alignment of the core SNPs can be determined to perform phylogenetic analysis with methods such NJ or maximum likelihood, respectively. Since SNP-based approaches identify variation at the single nucleotide level, they can potentially provide higher resolution than wg/cgMLST for the regions being considered. However, it is necessary to meet two requirements for high-resolution typing with SNP-based methods. First, choosing a reference genome closely related to the strains under investigation is of the utmost importance, as SNPs are only identified in the regions shared with the reference. A more divergent strain may considerably reduce the number of shared regions, and consequently the number of detected SNPs. Strategies for choosing a reference genome include selecting a high quality genome from the same ST, CC, serogroup or determining the genome distance to a group of candidate reference genomes to select the most appropriate. Second, the group of strains under investigation cannot be very diverse, as that will also reduce the number of regions considered for SNP determination [129]. These requirements can be met for outbreak detection and investigation scenarios in which SNP-based methods have been successfully applied to resolve national and international outbreaks. However, these requirements limit the applicability of SNP-based methods compared to wg/cgMLST approaches, which can be used to characterize from very diverse datasets to closely related strains. In addition, SNP-based methods are less robust to recombination and Horizontal Gene Transfer (HGT) than wg/cgMLST, as a single event will lead to the identification of multiple SNPs, potentially overestimating the distance between strains. The congruence between SNP-based and wg/cgMLST approaches is highly dependent on the reference genome and the schema used, respectively, as well as the parameters used and the dataset under analysis. While some studies have reported good congruence between both approaches for outbreak analyses, others have highlighted

that the results are not directly comparable and that a congruence analysis is necessary to assess method equivalence [119]. Furthermore, SNP-based approaches do not scale as well as wg/cgMLST, in part because they are more computationally demanding as the size of the dataset increases, but also because it may be necessary to use multiple references and fine-tune the parameters for accurate SNP detection, making it harder to standardize or establish a reference database for consistent results. Nevertheless, variant calling data can be stored to promote reproducibility and minimize scalability concerns, an approach implemented in SnapperDB [131].

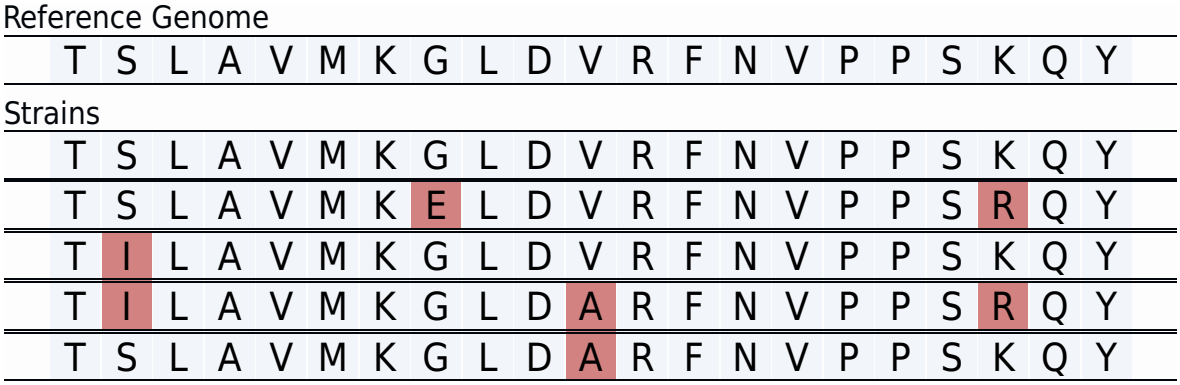
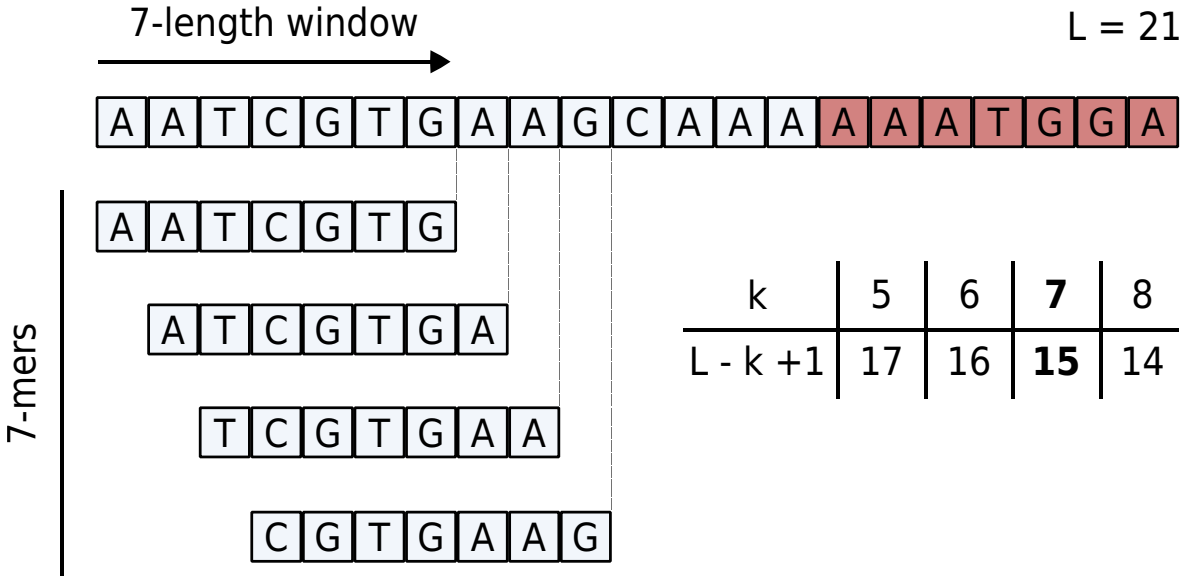


Figure 1.9: SNP-based methods map the sequencing reads or genome assemblies from bacterial strains against a reference genome to identify SNPs. The top sequence in the image represents a region of the reference genome and the bottom lines represent the same region for a group of strains compared against the reference. The positions that differ from the reference genome are colored in red.

1.4.4.5 *k*-mer-based methods

Although the concept of *k*-mer has existed for several decades, even if under different designations (e.g., N-gram, k-tuple, w-mers), its wide application to increase the efficiency and accuracy of bioinformatics methods is relatively recent. *k*-mer-based approaches break sequences into smaller subsequences of length *k* (Figure 1.10). This seemingly simple approach of *break it to understand it* has enormous potential, allowing for much more time-efficient sequence comparisons than alignment-based approaches. However, storing large sets of *k*-mers in memory can lead to high memory usage. Thus, it is important to optimize parameters such as the *k*-mer size and sampling method, also termed sketching (i.e., which *k*-mers to select from all possible *k*-mers generated from a sequence). With respect to *k*-mer size, it is important to optimize it for the desired application by balancing the trade-off between greater specificity achieved with longer *k*-mers and greater sensitivity with shorter *k*-mers. Optimizing *k*-mer selection is more complex and has been the subject of extensive research [132–137].

Ideally, it is desired to select the smallest set of *k*-mers that minimizes memory usage without sacrificing accuracy. However, depending on the application, it may be necessary to satisfy other requirements that complicate the determination of the optimal sampling method. For example, a sampling method that selects *k*-mers randomly may be preferred for unbiased



k	5	6	**7**	8
L - k +1	17	16	**15**	14

Figure 1.10: Determining *k*-mers. After defining the value of *k*, the first *k*-mer is determined by selecting the segment with the first *k* bases in the sequence. The next *k*-mer is determined by sliding one position to the right to select a new *k*-mer that differs from the first one by a single position. This process is repeated until all consecutive *k*-mers for a sequence are determined. The figure represents the determination of 7-mers from a sequence with 21 nucleotides (L). The first four 7-mers are shown below the sequence. The last 7-mer is highlighted in red. The table shows the total number of *k*-mers that can be determined from the sequence by varying the value of *k* from 5 to 8. The column corresponding to the total number of 7-mers is in bold.

sequence comparisons, while optimizing the interval between consecutively selected *k*-mers may offer better performance when sequence variability is higher. Theoretical and empirical evaluations of the performance of each sampling method are important to determine optimal parameters and limitations. Minimizers and spaced seeds are two sampling methods that have been widely applied due to their simplicity and effectiveness in a wide range of applications, such as genome assembly and taxonomic classification. *k*-mers are a centerpiece of most genome assemblers, with the construction of *de Brujin* graphs, representing the overlap between *k*-mers determined from sequencing reads, being the fundamental strategy used by many genome assemblers to solve the problem of genome assembly for both short- and long-read data [138]. For taxonomic classification, sequence data are often decomposed into *k*-mers and compared against *k*-mer indexes constructed from taxonomically annotated reference sequences. This strategy allows for ultra-fast taxonomic classification. Kraken was one of the first highly successful *k*-mer based tools for taxonomic classification [139]. Its first version enabled fast taxonomic classification based on a minimizer index, and the second version optimized the database structure to reduce memory requirements and added spaced seeds for greater accuracy [140]. In theory, *k*-mer-based methods offer several advantages over wg/cgMLST and SNP-based methods. Firstly, there is no need to define a reference database, such as a schema, or select a reference genome to compare strains against. This overcomes the reference bias limitation of SNP-based approaches and eliminates the need to store and manage a schema with a specific allelic nomenclature. However, to achieve consistent results, especially when trying to establish a cluster nomenclature, it is still recommended to create a database structure to compare and classify strains. Otherwise, the results obtained

for different datasets may not be comparable. Secondly, and contrary to wg/cgMLST, *k*-mer-based methods can include non-coding regions to estimate strain similarity. Lastly, if adequately parameterized, *k*-mer-based methods can be faster and more efficient than wg/cgMLST or SNP-based methods. A good example of an efficient and versatile *k*-mer-based method applicable to closely related samples and outbreak investigation is SKA2 [141]. SKA2 uses split *k*-mer analysis for reference-free and reference-based mapping to genotype bacterial strains using sequencing reads or genome assemblies. Another *k*-mer based tool, PopPUNK [142], uses *k*-mers to calculate core and accessory distances, which in turn are used by machine learning algorithms to cluster bacterial strains. This approach has proved useful for population analysis, and the ability to update existing clusters with new strains without having to recalculate all pairwise distances is especially useful for surveillance.

1.5 Aims of the Thesis

The advances in DNA sequencing technologies have allowed research and public health institutions to gradually switch from classical and more laborious microbiological workflows to WGS-based methods for the characterization of bacterial pathogens. With modern sequencing technologies, it is possible to streamline the genome sequencing of dozens of bacterial strains. The widespread adoption of WGS and the increased availability of bacterial genomes in public databases enabled researchers to develop and apply more advanced bioinformatics methods to gain greater insight into the structure, diversity and dynamics of bacterial genomes. GbG, SNP-based and *k*-mer-based methods allow for a more detailed characterization of bacterial pathogens and are currently widely applied for surveillance, outbreak investigation, and to study the diversity of bacterial species. GbG methods in particular, such as wg/cgMLST, have been adopted more frequently for surveillance and outbreak investigation, perhaps in part because they constitute an expansion of the classical seven-gene MLST, making it conceptually and technically easier to implement and transition to. The wide adoption of wg/cgMLST makes it relevant to explore ways in which it can be further improved. The main objective of this work is to explore concepts and implement solutions that improve the efficiency, accuracy, and interoperability of wg/cgMLST. The main goals are the following:

- Optimize the processes of schema creation and allele calling in wg/cgMLST, as well as implement solutions for comprehensive analysis of the schema and results data;

- Implement a Nomenclature Server to store and manage wg/cgMLST schemas that enables local and private analyzes based on a common allelic nomenclature;

- Identify and propose solutions for common problems found during the creation and application of wg/cgMLST schemas, either derived from low-quality data or from limitations of current methodologies used in wg/cgMLST.

The first goal will be achieved mainly through the optimization of the chewBBACA software for wg/cgMLST [143]. The second goal will focus on creating a Web service that provides easy access to wg/cgMLST schemas and minimizes scalability and data privacy concerns compared to other well-known wg/cgMLST platforms that centralize data analysis and require users to share their data. The last goal will help create strategies to identify and correct spurious loci added to wg/cgMLST schemas, improve the integration of accessory loci into wgMLST schemas, and provide valuable information to guide the development of chewBBACA.

Chapter 2

chewBBACA 3: lowering the barrier for scalable and detailed whole- and core-genome multilocus sequence typing

This chapter is a reproduction of the following article:

R. Mamede, P. Vila-Cerqueira, J. A. Carriço, M. Ramirez. chewBBACA 3: lowering the barrier for scalable and detailed whole- and core-genome multilocus sequence typing. (2025). [Manuscript submitted for publication].

The supplementary material referred to throughout the text can be consulted in the last section of this chapter, **Section 2.9**.

As mentioned in **Chapter 1**, wg/cgMLST approaches have been increasingly adopted by research and public health institutions for surveillance, outbreak detection and characterization, and for the study of the population diversity of bacterial pathogens, especially of Food and Waterborne disease (FWD) pathogens. Since wg/cgMLST is an expansion of the classical seven-gene MLST, it is conceptually and technically easier to implement and standardize. However, the successful integration of wg/cgMLST is not without challenges. Firstly, as more and more sequencing data becomes available, it is clear that wg/cgMLST methods need to be continuously improved to ensure the level of efficiency necessary to process larger volumes of data. Secondly, while we can determine the complete or nearly complete genome sequence of bacterial pathogens, most approaches focus only on the core genome (i.e., cgMLST). The core genes are present in all or most strains and generally display lower sequence diversity than the accessory genes, enabling the definition of stable cgMLST schemas that provide good resolution. However, while robust, cgMLST ignores the diversity of the accessory genome and may not offer enough resolution for outbreak-level strain discrimination. Lastly, currently there is no wg/cgMLST tool or platform that offers scalable and decentralized schema creation, allele calling, and comprehensive analysis of the results, which may hinder efforts for the global adoption of wg/cgMLST, especially in cases where computational

resources are limited and for users who lack bioinformatics training.

This chapter describes the main features of chewBBACA 3, a suite of modules developed for scalable and comprehensive wg/cgMLST. chewBBACA 3 is a complete reimplementation of the first published version of chewBBACA. By developing chewBBACA 3, I continued to support a widely used tool that was initially developed in the lab where the work presented in this thesis was carried out and had the opportunity to tackle some of the challenges related to wg/cgMLST referenced above. The results presented in this chapter highlight the efficiency and accuracy of the schema creation and allele call processes with chewBBACA 3, which enable fast large-scale analysis on a laptop. In addition, the modules for comprehensive schema and allele call results analysis included in chewBBACA 3 make it a complete solution for wg/cgMLST in surveillance and outbreak investigation settings, as well as for population studies.

The contributions of the other authors and the suggestions provided by users have been instrumental in the continuous improvement of chewBBACA. I have been maintaining and actively improving chewBBACA since 2020 to build a tool for wg/cgMLST that considerably reduces computational requirements for large-scale wg/cgMLST, enables users to create schemas that capture more of the diversity of bacterial species to go beyond cgMLST, and provides analytic capabilities that help to explore results and reach an informed decision. As the main contributor, I have been involved in all aspects of the development of chewBBACA 3 and on the study design, data analysis, and writing related to the results presented in this chapter.

chewBBACA 3: lowering the barrier for scalable and detailed whole- and core-genome multilocus sequence typing

Rafael Mamede[1,2], Pedro Vila-Cerqueira[1], João André Carriço[1], Mário Ramirez[1]

[1] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal;

[2] Gulbenkian Institute for Molecular Medicine.

2.1 Abstract

2.1.1 Background

The wide adoption of whole genome sequencing has enabled the implementation of genomics-based systems, which provide unparalleled resolution for the surveillance and outbreak investigation of bacterial pathogens. To fully exploit the wealth and complexity of genomics data, bioinformatics methods need to be highly scalable, provide accurate and extensive data for potential downstream analyses, as well as analytic capabilities. Here, we present chewBBACA 3, a suite of modules for scalable and comprehensive bacterial wg/cgMLST with built-in features to create new schemas, evaluate loci diversity and strain similarity.

2.1.2 Results

chewBBACA 3 enables faster and more accurate schema creation and allele calling by complementing an alignment-based approach with alignment-free methods, including hash-based comparisons and minimizer-based clustering. Schema creation is up to 55-fold faster and identifies up to 10% more loci than its predecessor, chewBBACA 2. Furthermore, chewBBACA 3 can quickly adapt or import schemas available on external wg/cgMLST platforms or Chewie-NS, promoting interoperability. The efficiency of allele calling allows processing larger genome collections, from thousands to tens of thousands of genomes, at the whole- and core-genome levels without requiring high computational resources and being up to 52-fold faster than similar tools. chewBBACA 3's enhanced sensitivity allows it to identify and classify more schema loci and coding sequences than the compared methods, resulting in higher resolution for strain comparison. Moreover, the allelic profiles, classification statistics and associated sequence data produced by chewBBACA 3 can be the basis for detailed analyses that provide added value in surveillance and outbreak investigation settings. New modules leverage the potential of the schema and allele call results data to create interactive

reports that enable an intuitive and in-depth analysis of allele diversity in loci of interest and allow assessing strain similarity based on loci presence, allelic distances and phylogenetic analysis.

2.1.3 Conclusions

chewBBACA 3 provides functionalities for complete wg/cgMLST analysis at scale, lowering the barrier for the use of wg/cgMLST and offering extensive results and analytic capabilities for streamlined, comprehensive, and local analyses. chewBBACA 3 is freely available at `https://github.com/B-UMMI/chewBBACA`.

2.2 Background

The burden of bacterial infections constitutes a major challenge to public health systems worldwide [3, 4]. The advances in sequencing technologies have enabled public health institutions to support and gradually transition to whole genome sequencing (WGS), increasing surveillance capacity and the effectiveness of outbreak investigations. WGS offers high-resolution discrimination of closely related bacterial strains and enables the identification of pathogens' relevant features in a timely and accurate manner, aiding in reaching an informed decision for effective disease prevention and control [113, 114, 144, 145]. The widespread use of WGS, as well as the adherence to Findable Accessible Interoperable Reusable (FAIR) principles, encouraged the development of efficient bioinformatics methods for in silico MLST, serotyping, and the identification of antimicrobial resistance and virulence determinants. It also allowed transitioning to methods with enhanced resolution that leverage the full genomic content to identify relevant features and provide a more accurate measure of strain similarity [146].

These methods are diverse but generally adopt one of three fundamental approaches: i) determining Single Nucleotide Variants (SNVs) relative to a reference genome, ii) measuring sequence similarity based on short subsequences of length k, known as k-mers, and iii) comparing the strains' gene content, referred to as GbG methods [146].

Single Nucleotide Variant (SNV) approaches detect differences at the single nucleotide level by mapping sequencing reads against a closely related reference strain. The precision level of this approach enables the identification of point mutations or more complex variants that can be determinants of phenotypic characteristics of interest, such as increased virulence and antimicrobial resistance. The choice of the reference genome is crucial as the quality and relatedness of the reference genome to the strains of interest can greatly influence the number of shared positions compared and, therefore, the extent of the variability detected [147, 148]. k-mer-based tools split genomic sequences into k-mers and compare the resulting

k-mer sets to estimate strain similarity or identify regions of interest. These approaches can estimate similarity without needing a reference genome and are potentially faster and more computationally efficient than SNV or GbG approaches. The efficiency of these approaches depends on the sampling method used to select k-mers, which should be fine-tuned to achieve a good balance between efficiency and accuracy for the desired application [134, 149]. With the wide adoption of WGS, GbG approaches have transitioned from classical MLST to wg/cgMLST. wg/cgMLST enables the creation of schemas encompassing the variability of hundreds to thousands of loci for a species of interest to accurately determine the loci and alleles present in strains of interest. Creating and maintaining wg/cgMLST schemas to capture a species' diversity is crucial for the accuracy of GbG methods and can be a laborious process. As with SNV approaches, knowing the alleles present at a given locus can be linked to phenotypic properties such as virulence or antimicrobial resistance.

It has been shown that applying any of these approaches can generate results suitable for accurate strain similarity estimation and phylogenetic analyses in surveillance and outbreak scenarios [134, 146–150]. Nevertheless, wg/cgMLST has been more frequently integrated into surveillance and outbreak detection systems, partly due to constituting an expansion of classical MLST, which conceptually and technically allows for a more straightforward implementation, specially in constantly growing datasets such as the ones used in long-term epidemiological surveillance. The capacity to update wg/cgMLST schemas with new alleles increases the diversity captured by and, consequently, the resolution of wg/cgMLST analyses. Moreover, wg/cgMLST allows establishing allelic nomenclatures for standardised comparisons. Existing solutions for wg/cgMLST analysis can vary greatly in the degree of data centralisation, analytical capabilities, and license type [129, 151, 152]. To continue to promote the adoption of wg/cgMLST, improvements should focus on interoperability to facilitate comparison of results, scalability to meet growing data processing demands, and easily performed comprehensive local analyses to offer powerful analytic capabilities to end users while complying with strict data privacy laws.

To provide a solution for scalable, detailed, and local wg/cgMLST, we developed chewBBACA 3, which vastly improves and extends the functionalities of chewBBACA 2 [143, 152], a widely used tool for wg/cgMLST which has been integrated into public health workflows such as the EFSA One Health WGS system, used for rapid detection of multi-country foodborne outbreaks in collaboration with the ECDC [114].

2.3 Implementation

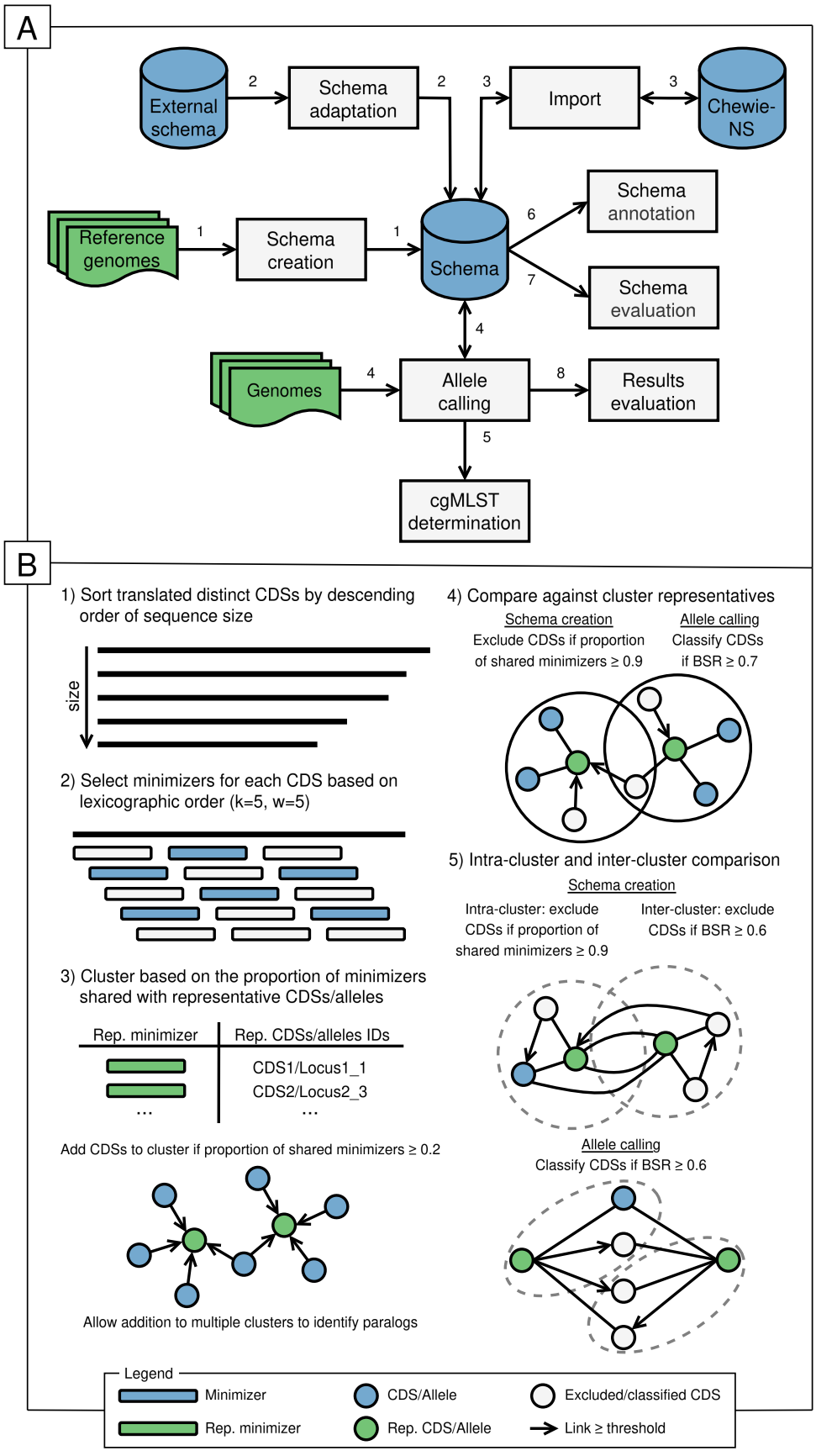
2.3.1 Overview

chewBBACA 3 is a complete reimplementation of its predecessor, chewBBACA 2 [152], which was already an upgraded version of chewBBACA's first published version [143]. chewBBACA 3 provides a modular approach for complete wg/cgMLST analysis allowing more efficient and accurate schema creation and allele calling and offering interactive reports for comprehensive schema and results analyses (Figure 2.1). chewBBACA 3 allows to set up schemas for wg/cgMLST through the use of larger collections of genome assemblies or coding DNA sequences (CDSs) in FASTA format or by adapting existing schemas from external platforms [129, 151, 153]. Additionally, the integration with Chewie-NS, which was previously described together with chewBBACA 2 [152], allows easily importing ready-to-use schemas to obtain comparable interlaboratory results based on a common allelic nomenclature. To determine the allelic profiles of strains of interest, chewBBACA 3 identifies and clusters the distinct CDSs predicted from the strains' genomes, significantly reducing the number of comparisons against the schema loci in contrast to the sequential strain processing used by chewBBACA 2. This translates into faster and more efficient allele calling and facilitates data aggregation to create output files with more detailed results. Allele calling identifies and adds new alleles to schemas, ensuring that they are gradually updated to produce accurate and comparable results over time. Similarly to chewBBACA 2, new alleles are inferred based on the BLAST Score Ratio (BSR) [154] computed from Protein BLAST (BLASTp) alignments [155], complying with minimum sequence length and allele size variation thresholds [143]. Adjustments to these parameters allow chewBBACA 3 to classify more CDSs and capture loci diversity more accurately. chewBBACA 3 increases the granularity of the results by expanding the set of special classifications assigned when the presence of a locus cannot be inferred confidently, such as when a coding DNA sequence (CDS) matching a schema locus is outside the user-specified locus size variation interval or if multiple CDSs from a genome match the same schema locus (Figures 2.5-2.8). These special classifications aid in identifying spurious alleles resulting from low-quality data, pseudogenes, and paralogous loci. The set of core loci can be determined based on the allele calling results for any locus presence threshold, and the resulting list of core loci can be used to perform allele calling at the core genome level. Schema loci can be annotated by searching for matches through UniProt's SPARQL Protocol and RDF Query Language (SPARQL) endpoint, which was already an option in chewBBACA 2, and now also by aligning a schema's alleles against UniProt's reference proteomes to retrieve annotations based on higher-quality entries [156]. New schema and results evaluation modules leverage the power of the React JavaScript library [157] to build interactive reports that enable local and comprehensive analyses of the diversity of loci contained in schemas and aid in identifying closely related strains for more effective surveillance and outbreak assessment.

2.3.2 Core modules

While all the modules in chewBBACA 2 were reimplemented to increase the scalability and comprehensiveness of the results generated by chewBBACA 3, module development concentrated primarily on the *CreateSchema* and *AlleleCall* modules (Figures 2.9 and 2.10), which handle schema creation and allele calling, respectively. Gene prediction was optimised in both modules using Pyrodigal [158, 159], a Python module that provides bindings to Prodigal for seamless integration and offers several advantages, such as faster gene prediction and greater control over gene prediction parameters and results. Additionally, a novel feature was added allowing both modules to accept FASTA files with CDSs, enabling users to leverage the vast CDS data available in public databases or provide CDSs predicted by other gene prediction tools if preferred by the user, such as GeneMarkS-2 or Balrog [160, 161]. Following the gene prediction step, CDS deduplication is performed at both DNA and protein levels to identify the set of distinct CDSs. The distinct translated CDSs are clustered based on the proportion of shared minimizers (≥ 0.2) with representative alleles [132, 162, 163] (Figure 2.1B:1-3). The minimizer parameters (k=5, w=5, lexicographic order) were chosen to select sets of k-mers that would cover most sequence positions at least once while also keeping memory usage low [164]. The low clustering threshold groups similar sequences into the same clusters, reducing the number of comparisons in subsequent steps. For schema creation, the CDSs sharing a high proportion of minimizers (≥ 0.9) with the cluster representative or larger CDSs are considered alleles of the same locus and are excluded (Figure 2.1B:4). The clustering results are complemented by intracluster and intercluster alignment with BLASTp to exclude CDSs based on the BSR threshold and select the final set of CDSs (Figure 2.1B:5). This defines the schema by creating a schema seed with one representative allele for each locus. This schema seed will be used by the *AlleleCall* module, which may add further representative alleles to capture the allelic diversity at each locus. The *AlleleCall* module uses the same functions as the schema creation process for sequence deduplication and CDS clustering. Another novel feature implemented is that each distinct CDS is hashed, mapped to the compressed list of genomes that contain it (Figure 2.11) and compared against the hashed schema alleles. This allows keeping the information about the CDSs identified in all strains in memory, enabling fast exact matching and classification of all genomes containing a CDS based on a single match. Clustering and intracluster alignment with BLASTp allow comparison of the remaining unclassified CDSs against the schema's representative alleles to find and classify inexact matches (Figure 2.1B:1-4). A final step aligns the schema's representative alleles against the remaining unclassified CDSs to find more divergent alleles and select new representative alleles (Figure 2.1B:5). The matches found throughout the process are evaluated at the end of the process to assign the final classifications, create the allelic profiles, and update the schema with novel alleles. Both core modules create several output files with detailed schema and results data that support the analyses performed by other modules and can serve as the basis for custom analyses that seek to answer relevant questions at the strain or population-wide level. A more detailed description of the implementation and functionalities included in each module is available in the supplementary material and in

chewBBACA's online documentation [165].



(Caption on next page.)

Figure 2.1: Overview of chewBBACA 3's processes and minimizer-based clustering used by the *CreateSchema* and *AlleleCall* modules. (A) chewBBACA 3 includes modules for schema setup (steps labeled with 1, 2 and 3), allele calling (steps labeled with 4), core genome determination (steps labeled with 5), schema annotation (steps labeled with 6), schema evaluation (steps labeled with 7), and results evaluation (steps labeled with 8). Blue cylinder icons represent schemas, with the central cylinder icon representing a schema created or adapted for usage with chewBBACA 3. Green document icons represent input FASTA files. Grey rectangle icons represent analysis processes available in chewBBACA 3. (B) Minimizer-based clustering and classification steps implemented in the *CreateSchema* and *AlleleCall* modules. (Step 1) The distinct translated CDSs not classified through exact matching at the DNA and protein levels are sorted based on decreasing size. (Step 2) Minimizers are selected from the set of 5-mers for each CDS based on lexicographic order and a window size of 5. (Step 3) The set of minimizers selected for each CDS is compared against the minimizers of CDSs selected as cluster representatives (*CreateSchema*) or the schema loci representative alleles (*AlleleCall*) to cluster CDSs based on a proportion of shared minimizers ≥ 0.2. (Step 4) The CDSs that share a proportion of minimizers ≥ 0.9 (*CreateSchema*) or a BSR ≥ 0.7 (AlleleCall) with the cluster representative are excluded from the analysis (*CreateSchema*) or classified (*AlleleCall*). (Step 5) Non-representative CDSs from the same cluster are compared to exclude smaller CDSs that share a proportion of minimizers ≥ 0.9 with larger CDSs (*CreateSchema*). Representative CDSs or alleles are aligned against all CDSs to exclude (*CreateSchema*) or classify (*AlleleCall*) CDSs based on a default BSR value of 0.6.

2.4 Results and discussion

2.4.1 Fast wg/cgMLST schema creation or retrieval from multiple sources

chewBBACA 3 offers three options for setting up a schema for wg/cgMLST analysis.

The first option is creating a new schema by selecting loci from a set of complete or draft genome assemblies with the *CreateSchema* module to create a schema seed (Figure 2.9). To evaluate the performance of the *CreateSchema* module, we created schema seeds with chewBBACA 3 and chewBBACA 2 based on the complete genome assemblies available on the NCBI RefSeq database [166] for three bacterial species: *Streptococcus pyogenes* (n=260), *Listeria monocytogenes* (n=309), and *Salmonella enterica* (n=1,326). Schema seed creation was 25- to 55-fold faster with chewBBACA 3 than with chewBBACA 2, with similar memory usage (Table 2.1). A comparison of the schema seeds generated with both versions revealed that the schema seeds created by chewBBACA 3 contained 98% of the loci identified by chewBBACA 2 (Table 2.2). Moreover, chewBBACA 3 identified 6% to 10% more loci than chewBBACA 2, primarily due to a more accurate identification of smaller loci. Ideally, target loci should be defined based on a set of high-quality genome assemblies to avoid the inclusion of spurious loci in the schema seed. Nonetheless, schema seed creation with chewBBACA 3 will remain efficient even when using larger genome collections, possibly including draft genomes, to adequately capture a species diversity.

A second option is to adapt schemas from external platforms with the *PrepExternalSchema* module (Figure 2.12). This module filters out incomplete alleles (i.e. alleles that contain ambiguous bases or not corresponding to valid CDSs, such as having no start/stop codon)

and selects representative alleles based on a BSR threshold to create a schema structure compatible with chewBBACA. Schema adaptation promotes the usage of schemas previously made available and adopted by the community, contributing to the integration and interoperability with other platforms. chewBBACA 3 is over three orders of magnitude faster than chewBBACA 2 when adapting the cgMLST schemas for *S. pyogenes*, *L. monocytogenes*, and *S. enterica* available on the cgMLST.org server [153], adapting any of the schemas in under five minutes (Table 2.3). Furthermore, contrarily to chewBBACA 2, chewBBACA 3 now ensures that the selected representative alleles fully capture the diversity of each locus based on the specified BSR threshold (Table 2.4). chewBBACA 3 also provides options to filter out alleles based on user-defined sequence size and size variation thresholds and outputs detailed information about the changes made while adapting a schema to inform the user of the changes introduced to the existing schema.

Lastly, the third option is the *DownloadSchema* module (Figure 2.13), one of the modules (Figures 2.13-2.15) developed to integrate with Chewie-NS [152], allowing users to import ready-to-use schemas from Chewie-NS instances. This option offers the advantage of enabling local and private analysis based on a common allelic nomenclature to facilitate the comparison of results. Schemas downloaded from Chewie-NS can be kept up-to-date by synchronizing with the remote versions to receive the latest allele data submitted by other users and, if desired, contribute novel alleles identified locally.

2.4.2 Scalable and efficient allele calling

We performed allele calling with the schema seeds created with chewBBACA 3 in the previous section and the complete genomes for each species to add new alleles to the schemas and determine the set of core loci with the *ExtractCgMLST* module (Figure 2.16). The lists of core loci (present in 100% of the genomes) were used to measure performance at the cgMLST level for datasets including between 1 and 16,384 draft genome assemblies and compared against the results obtained with chewBBACA 2 and pyMLST [167] for equivalent schemas and databases (Table 2.5). chewBBACA 3 processed all datasets faster than chewBBACA 2 and pyMLST. On average, chewBBACA 3 was 1.9- to 20.3-fold and 1.3- to 51.9-fold faster than chewBBACA 2 and pyMLST, respectively (Figure 2.2A and Table 2.6). The difference increased with dataset size, largely due to the increased redundancy (same sequence found in different genomes) of the set of CDSs extracted from the genomes (Table 2.7). For example, only 1.5% to 2.8% of the total CDSs identified in the complete datasets (n=16,384) were distinct. By identifying the set of distinct CDSs before trying to match them against the schema loci, chewBBACA 3 avoids the repeated evaluation of identical CDSs identified in multiple genomes.

Additionally, the novel minimizer-based clustering matches the remaining CDSs to the most similar schema loci, reducing comparisons between dissimilar sequences. These two

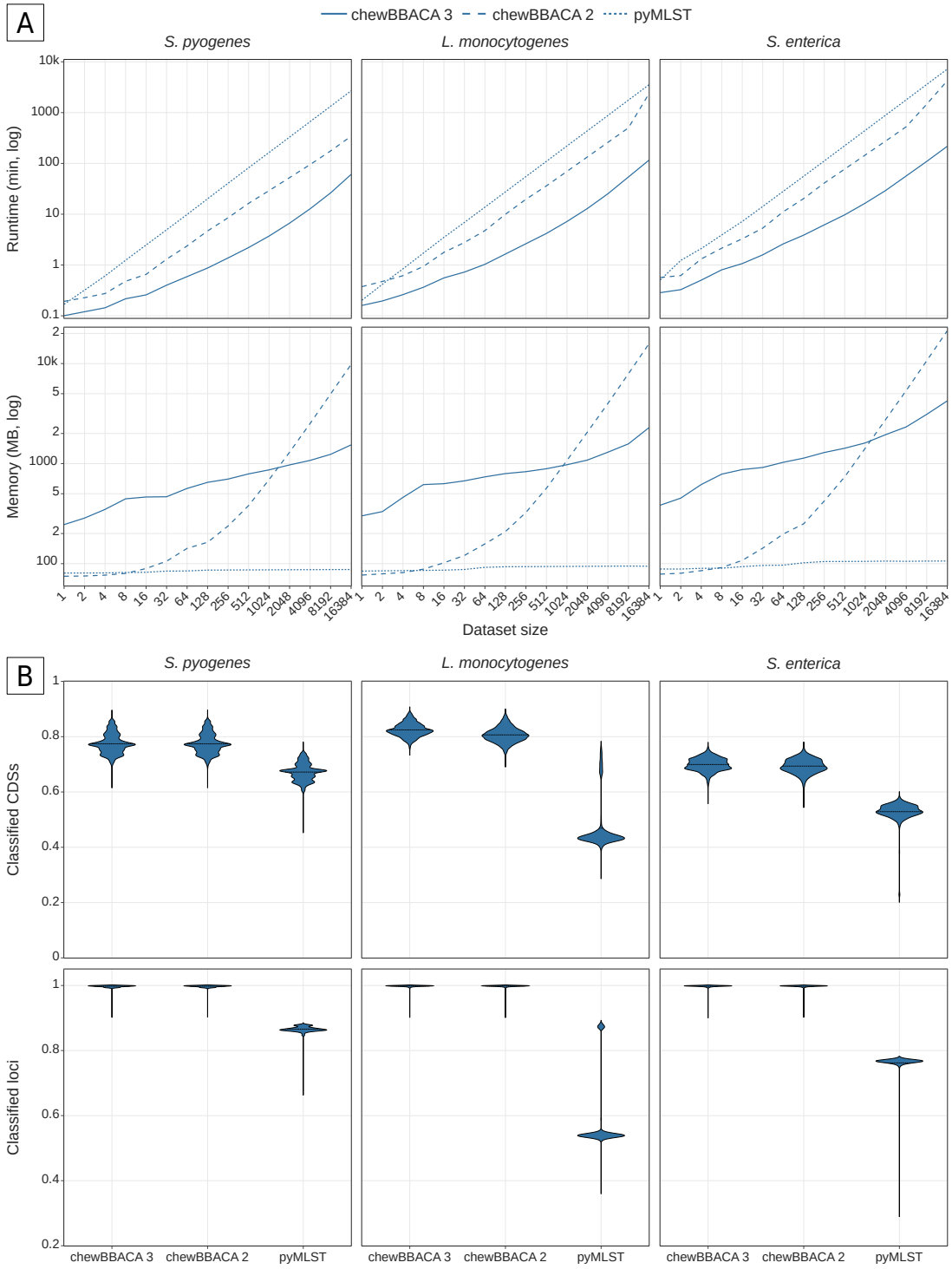


Figure 2.2: Performance comparison of chewBBACA 3, chewBBACA 2 and pyMLST. (A) Runtime and peak memory usage comparison for the allele calling of datasets with a varying number of genomes (from 1 to 16,384) for three bacterial species: *Streptococcus pyogenes*, *Listeria monocytogenes*, and *Salmonella enterica*. The benchmark was performed with five replicates per dataset size, except for the complete dataset (n=16384). The values shown are the mean of the replicate values for each dataset. Runtime was measured as the elapsed real time in minutes (logarithmic scale). Peak memory usage was measured as the maximum resident set size in megabyte (MB) (logarithmic scale). (B) Proportion of strain CDSs and schema loci classified for the complete datasets (n=16384). The proportion of classified CDSs corresponds to the number of CDSs classified by each tool divided by the total number of CDSs predicted for each strain by Pyrodigal. The proportion of classified loci corresponds to the number of schema loci identified by each tool divided by the total number of schema loci.

steps contribute the most to the increased speed compared to chewBBACA 2 and pyMLST, which process each genome separately and additionally do not take advantage of multiprocessing settings as efficiently as chewBBACA 3. Regarding peak memory usage (Figure 2.2A and Table 2.8), chewBBACA 3 used, on average, 8.6- to 1.1-fold more memory than chewBBACA 2 for datasets with up to 1,024 strains. The inverse was observed for larger datasets, with chewBBACA 2 using 1.1- to 6.9-fold more memory than chewBBACA 3. Compared to pyMLST, chewBBACA 3 used 3.1- to 40-fold more memory. pyMLST maintains low memory usage irrespective of dataset size but is single-threaded and only supports the addition of one strain per command, which limits its scalability. chewBBACA 3 enables considerably faster analyses while keeping memory usage in check to allow large-scale analysis without needing high-performance computing infrastructures. While comparing results using the entire wgMLST schema would further highlight chewBBACA 3's efficiency and accuracy, time and memory constraints related to running chewBBACA 2 and pyMLST under the same conditions invalidated such comparison.

The thoroughness of the allele calling in chewBBACA 3 can be controlled through four execution modes (Figure 2.10). Mode 1 identifies exact matches at the DNA level between the genomes' CDSs and the schema alleles. Mode 2 adds exact matching at the protein level, enabling the identification of novel alleles with synonymous substitutions. Mode 3 proceeds to clustering and intracluster alignment to identify similar alleles based on the BSR threshold. Mode 4, the default, runs the complete process to classify as many CDSs as possible and potentially selects new representative alleles, preparing the schema to better identify future novel alleles. Modes 1 and 2 offer a 4.7-fold speedup over the default mode (Figure 2.17 and Tables 2.9 and 2.10), but their capacity for allele identification is limited to only modestly divergent alleles. This makes them appropriate for applications where a less accurate but much faster strain discrimination is sufficient, or for faster allele calling with schemas that already capture most of a species' diversity for the set of loci that make up those schemas, as is the case for many publicly available cgMLST schemas. Mode 3 provides similar accuracy to the default mode in less time, with a more significant reduction in runtime for larger schemas and more diverse datasets. Mode 4 offers greater sensitivity to identify the most divergent alleles and select new representative alleles to add to schemas, which is essential to increase the diversity captured by a schema, especially in the initial phase of schema development. For schemas that already include representative alleles that capture a species diversity, Mode 3 and Mode 4 may only differ in the number of special classifications attributed, with Mode 4 identifying more.

chewBBACA 2 added new alleles to schemas automatically, not providing any option for users to prevent the allele call process from changing an existing schema. chewBBACA 3 includes the *--no-inferred* option to control this behaviour. This option can be helpful in several scenarios, including: updating schemas only periodically, in applications where frequent schema updates can compromise the reproducibility of the allele calling; classifying genomes from closely related species to identify similar loci; and avoiding adding spurious alleles to a schema when there's uncertainty about the quality level of the genome assemblies

being analyzed.

2.4.3 Comprehensive allele calling for more accurate and detailed results

We evaluated the allele calling results for the complete dataset of each of the three species chosen (consisting of 16,384 genomes) to measure the comprehensiveness of chewBBACA 3's results and compare it against chewBBACA 2 and pyMLST. Results were compared at the core and accessory genome levels, based on a locus presence threshold of 95%, of the cgMLST schemas defined above. Concordance was measured by comparing the pairwise Jaccard distances computed based on the allelic profiles. The core and accessory loci sets determined based on chewBBACA 3's and chewBBACA 2's results were highly similar, sharing over 99% and 95% of the loci at the core and accessory levels, respectively (Table 2.11). The pairwise Jaccard distances were strongly correlated and near the identity line, indicating high concordance between the results (Figure 2.3), with the pairwise allelic distances computed by both tools differing by 0 to 6 differences on average (Figure 2.18). The core loci sets determined based on pyMLST's results were considerably smaller, containing 42% to 80% of the schema loci, compared to over 94% for chewBBACA 3. The reduced number of core loci identified by pyMLST is related to an inconsistent identification of some loci in each species. This is partly due to the default identity and coverage thresholds used by pyMLST, which are more stringent than the default BSR threshold used by chewBBACA and do not allow for the same degree of allele sequence variability. Moreover, pyMLST uses a single representative allele per locus to search for matches, whereas chewBBACA 3 can add new representative alleles to schemas to better capture locus diversity. pyMLST's accessory loci sets were 4- to 11-fold larger than chewBBACA 3's (Table 2.11). The accessory pairwise Jaccard distances were weakly correlated, except for *S. pyogenes*, and the pairwise allelic distances differed by 49 to 141 differences on average. While chewBBACA 2 generates highly comparable results to chewBBACA 3, pyMLST yields considerably different loci sets and pairwise distances, indicating it is not easily comparable to chewBBACA 3. This highlights the importance of the choice of method for wg/cgMLST and how the differences detected and distance thresholds defined by different methods may not be equivalent.

chewBBACA 3 classified a similar number of CDSs than chewBBACA 2 for *S. pyogenes* and 1.8% and 0.7% more CDSs for *L. monocytogenes* and *S. enterica*, corresponding to an average of 56 and 33 more CDSs per strain (Figure 2.2B and Table 2.12). chewBBACA 3 classified 10% to 35% more CDSs than pyMLST, 177 to 1078 more CDSs per strain, on average. chewBBACA 3 and chewBBACA 2 identified over 99% of the schema loci in all strains, while pyMLST identified between 58% and 87% loci (Figure 2.2B). Running chewBBACA 3 in mode 3 provided nearly identical results to the default mode. Modes 1 and 2 classified 4% to 6% fewer CDSs and identified 6% to 7% fewer loci than the default mode, respectively, performing worse if many of the strains' alleles were not equal or highly similar to the alleles in the schemas (Figures 2.19 and 2.20).

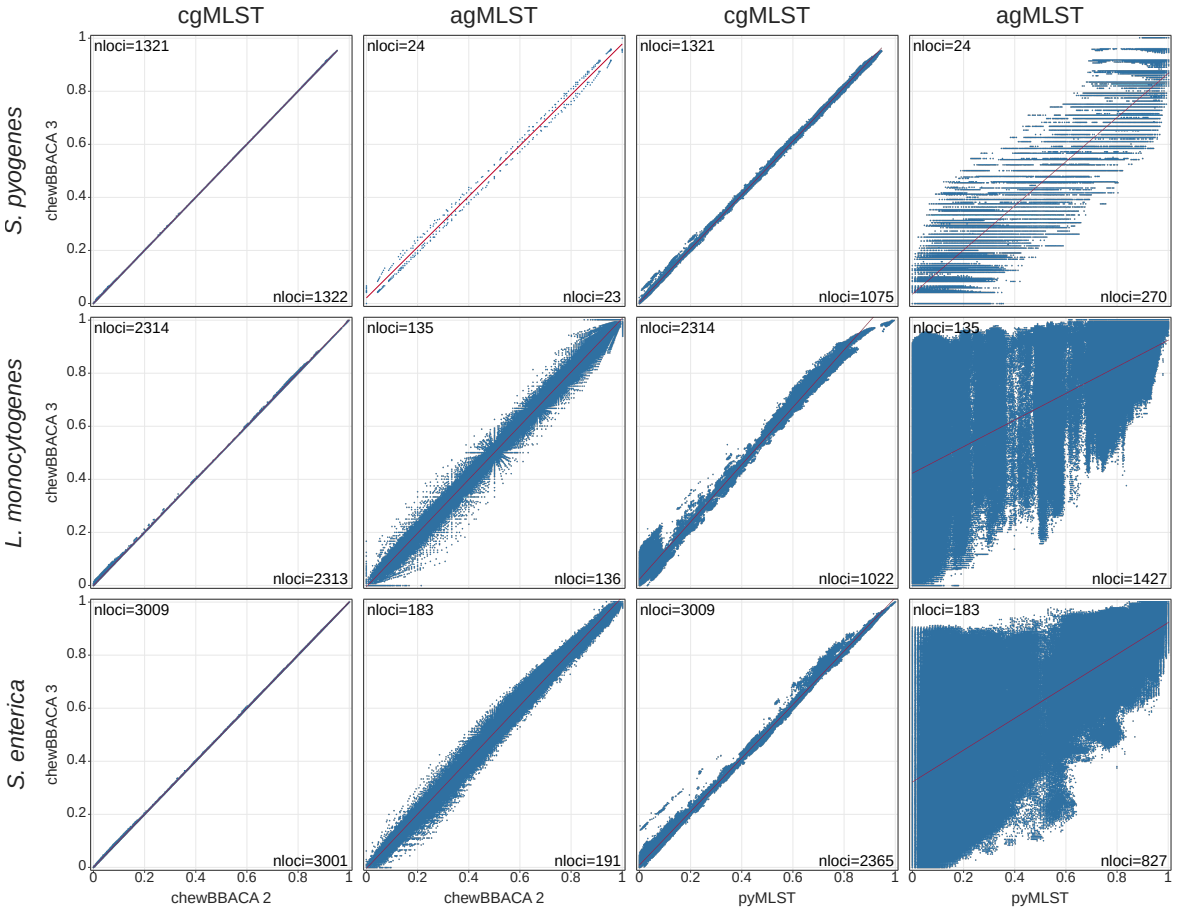


Figure 2.3: Comparison of the core (cgMLST) and accessory (accessory-genome MLST (agMLST)) pairwise Jaccard distances. The pairwise Jaccard distances computed based on chewBBACA 3's allele calling results for the complete datasets (n=16,384 genomes) of *Streptococcus pyogenes*, *Listeria monocytogenes*, and *Salmonella enterica* were compared against the the pairwise distances computed from chewBBACA 2's and pyMLST's results. The regression lines are displayed in red. The number of core or accessory loci determined based on chewBBACA 3's results are shown in the top-left corner of the plot area. The number of core or accessory loci determined based on chewBBACA 2's or pyMLST's results are shown in the bottom-right corner of the plot area.

Compared to chewBBACA 2, chewBBACA 3 identifies special classifications more accurately (Figures 2.21-2.23 and Table 2.13). The identification of paralogous loci was improved by introducing the Paralogous Match (PAMA) classification (Figure 2.7) for CDSs that match multiple loci and subdividing the Non-Informative Paralogous Hit (NIPH) classification into Non-Informative Paralogous Hit Exact Match (NIPHEM) and NIPH to differentiate between multiple exact matches or a combination of exact and inexact matches (Figure 2.6). chewBBACA 3 displays greater sensitivity for detecting multiple matches, leading to more NIPH and NIPHEM classifications than chewBBACA 2, which, in some cases, would detect a single exact match and fail to identify additional inexact matches. The Possible Locus On the Tip (PLOT) classification, used by chewBBACA 2 to classify CDSs close to contig ends, was subdivided into PLOT 5'-end (PLOT5), PLOT 3'-end (PLOT3) and LOTSC (LOTSC) to indicate if a CDS is close to the 5'-end, 3'-end, or both (Figure 2.5). New output files include the genomic coordinates for the CDSs predicted for all input genomes and relevant classification statistics per genome and locus. The DNA sequences of the CDSs assigned special

classifications or not classified can be stored in FASTA files by providing the *--output-missing* and *--output-unclassified* options, respectively. These changes improve the granularity of the results to facilitate downstream analyses, such as identifying low-quality inputs, paralogous loci, more divergent alleles, and potential new loci to add to schemas.

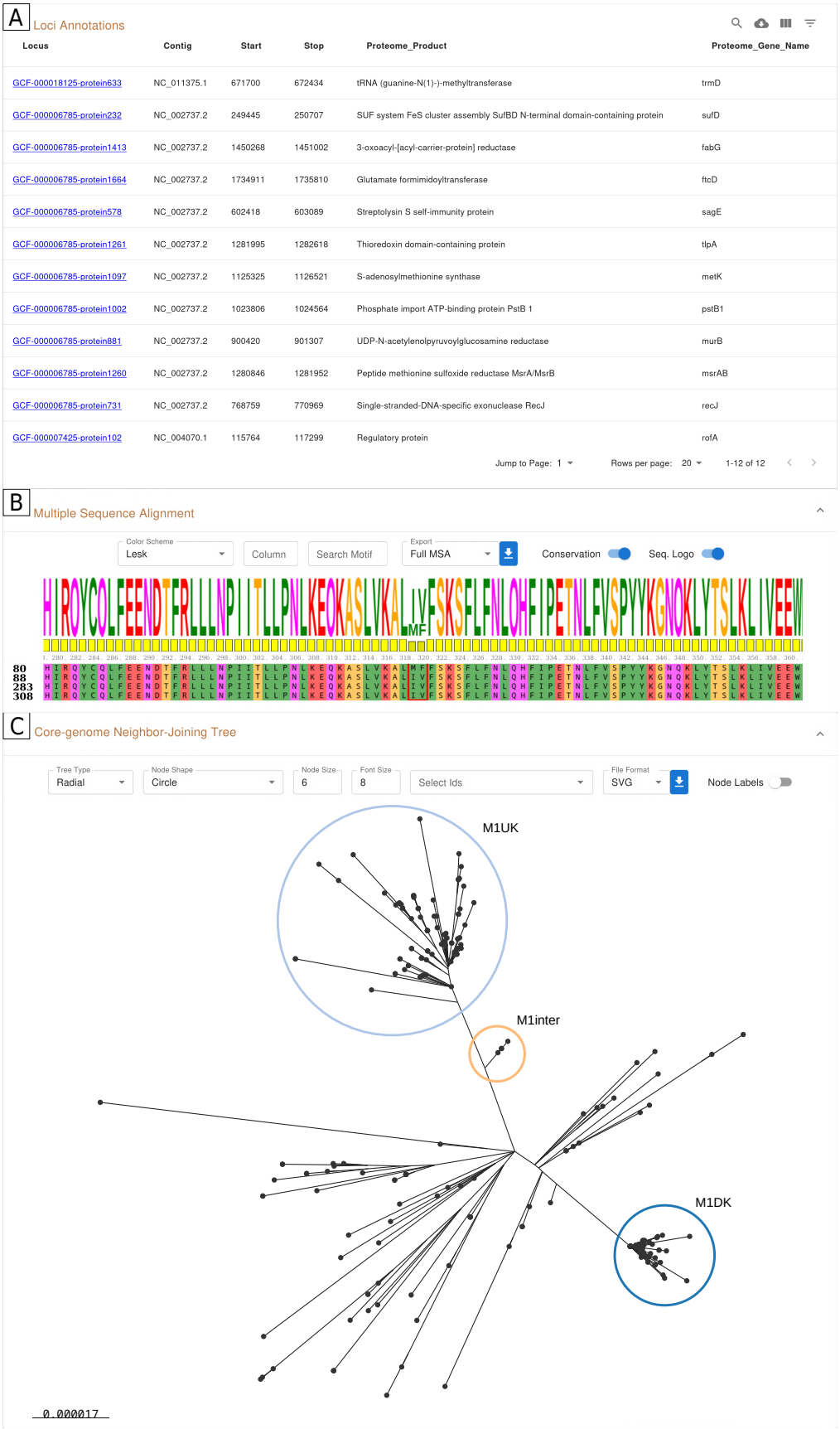
Another known issue when using wg/cgMLST approaches is that allelic profiles generated with schemas that do not share the same allele nomenclature are not directly comparable. To enable the comparison of results generated with different schemas, chewBBACA 3 includes the *--hash-profiles* option that hashes allele sequences to generate hashed allelic profiles. Since the same allele sequence will always result in the same hash value, the allelic profiles can be compared independently of the nomenclatures used by the schemas allowing also greater data privacy. chewBBACA 3 uses the SHA256 algorithm included in Python's *hashlib* module by default, but users can select any of the algorithms included in that module or the *zlib* module.

2.4.4 Interactive reports for comprehensive wg/cgMLST schema and allele call results analyses

The schemas and allele calling results generated by chewBBACA 3 can be a source of valuable data for in-depth analyses that explore the loci diversity captured by a schema and the relatedness of strains of interest. We developed modules that enable a local, scalable and comprehensive analysis of wg/cgMLST schemas and results through interactive reports to support users in performing common downstream analyses to more easily reach an informed decision. To showcase the utility of the reports' functionalities, we analysed 264 *S. pyogenes emm1* strains, including strains from the recently emerged $M1_{UK}$ and $M1_{DK}$ lineages [168, 169], and describe how some of the reports' components can help identify relevant features to distinguish the lineages.

The *SchemaEvaluator* (Figure 2.24) module evaluates wg/cgMLST schemas created with chewBBACA or from external sources to create an interactive report with detailed information about the schema composition. This module had been introduced in chewBBACA 2, however it ceased functioning due to dependency issues. The module was reimplemented and expanded in chewBBACA 3. The main page of the report includes charts that allow exploring the number of alleles and the allele size variation per locus. The module accepts a file with loci annotations to facilitate the identification of loci of interest. For example, the annotations determined by the UniprotFinder module (Figure 2.25) for the *S. pyogenes* schema can be added to a data table to identify which schema loci have the lineage-defining SNPs of the $M1_{UK}$ (Figure 2.4A) and $M1_{DK}$ lineages. Another data table displays the results of the allele integrity analysis, which identifies classes of invalid alleles per locus (e.g. incomplete CDSs, presence of ambiguous bases, absence of start and stop codons, in-frame stop codons, and minimum and locus-specific size thresholds). This can be used to identify problematic

loci or loci with unusual variability of size.



(Caption on next page.)

Figure 2.4: Report components generated for the analysis of the *S. pyogenes* schema and lineage strains. (A) Datatable component of the report generated by the *SchemaEvaluator* module including the annotations determined by the *UniprotFinder* module for 12 schema loci containing lineage-defining SNPs for the $M1_{UK}$ lineage. (B) Component of the *SchemaEvaluator* module including a MSA of the *rofA* translated alleles identified in the MGAS5005 reference strain (allele 80) and the $M1_{UK}$ strains (alleles 88, 283, and 308). Two amino acid differences caused by two SNPs in the *rofA* alleles of the $M1_{UK}$ strains are highlighted in red. (C) Component of the *AlleleCallEvaluator* module including a NJ tree computed with FastTree from the core loci MSA. The groups of strains belonging to the $M1_{UK}$ (light blue), $M1_{inter}$ (light orange), and $M1_{DK}$ (dark blue) lineages are highlighted. The full reports are available on Zenodo [170].

The *--loci-reports* option provides a more detailed analysis of each locus through dedicated locus pages, accessible by clicking on the loci identifiers in the main report data tables. Each locus page contains charts for the allele size distribution, sequence size per allele and number of DNA alleles for each distinct protein. A MSA computed with MAFFT [171] for the translated alleles allows identifying shared regions and differences caused by point mutations or indels. For example, the non-synonymous effect of two SNPs in the *rofA* gene used to define the $M1_{UK}$ lineage can be identified using the MSA by comparing the reference allele with those identified in $M1_{UK}$ strains (Figure 2.4B). The guide tree created by MAFFT is displayed with Phylocanvas.gl [172] to help identify groups of similar or divergent alleles. To provide a convenient way to identify and copy the DNA and protein sequences of the alleles, users can use the *--add-sequences* option, which adds code editor components containing the DNA and protein sequences to each locus' page.

A report with a detailed analysis of the allele calling results is obtained by running the *AlleleCallEvaluator* module (Figure 2.26). The report includes data tables with summary statistics and bar charts with the classification counts per strain and locus to explore the classification results and aid in identifying low-quality genomes (e.g. misassembled or contaminated genomes) and problematic loci (e.g. loci with a high number of special classifications). An interactive analysis of loci presence-absence is performed through a heatmap component that enables identifying the set of core loci or loci specific to certain groups of strains. Similar strains can be identified through another heatmap component that displays the matrix of pairwise core allelic distances and enables searching for similar strains based on a distance threshold. The last component in the report displays a NJ tree computed with FastTree 2 [173] from the core loci MSA. This component allows exploring phylogenetic relationships to identify groups of similar strains. For instance, the NJ tree from the analysis of the *S. pyogenes* strains allows identifying the groups of strains corresponding to each lineage of the M1 group (Figure 2.4C).

The reports' components include features to sort, select, search, and export data in tabular format, in the case of data tables, or as Scalable Vector Graphics (SVG) files, in the case of charts or trees. Some files necessary to create the components, such as the ones containing the matrix of pairwise core allelic distances and the core loci MSA, are provided in the report's folder to allow users to perform custom analyses if desired. The reports are easily shared by simply compressing the report's folder and sharing the resulting archive. The interactive

reports created with the *SchemaEvaluator* and *AlleleCallEvaluator* modules for the analysis of the *S. pyogenes* strains are available on Zenodo [170].

2.5 Conclusions

chewBBACA 3 constitutes an efficient, scalable, and comprehensive solution for wg/cgMLST. The options it provides for schema setup enable users to quickly create schemas from larger collections of genome assemblies or CDS data to capture more of the diversity of a bacterial species, or to adapt or import existing schemas created in other platforms or available in Chewie-NS to promote interoperability. The combination of alignment-based and alignment-free approaches allow for efficient and accurate allele calling, making it suitable for integration into workflows that process sample batches of any size, from sequential processing of single samples to vast genome collections for species-level population analyses. chewBBACA 3 classifies more schema loci and CDSs than the compared methods, potentially providing superior strain discrimination for surveillance and outbreak investigation. The high level of agreement with chewBBACA 2's results, while providing expanded classifications and richer results, facilitates the transition to the latest chewBBACA version. Comparisons with other wg/cgMLST methods should take into account that algorithmic differences between methods, parameter values, and input data quality can greatly affect the resolution and accuracy of the results, which might hinder results comparison and in some cases even lead to fundamentally different conclusions. The reports for schema and allele call evaluation allow a comprehensive and local analysis of locus diversity and strain similarity, enabling scalable and private analyses of the results and reducing the need to combine several tools or develop custom solutions to more fully explore the potential of wg/cgMLST schemas. The integration of chewBBACA 3 into wg/cgMLST workflows will help to further democratize wg/cgMLST by providing broader access to large-scale and detailed analyses to perform focused population studies or facilitate reaching an informed decision in outbreak or transmission investigations.

2.6 Methods

2.6.1 Download and selection of complete and draft genome assemblies

Complete and draft genome assemblies annotated as *Streptococcus pyogenes*, *Listeria monocytogenes* and *Salmonella enterica* were downloaded with the NCBI Datasets command-line tools v16.12.0 [174] on September 9, 2023. The complete genomes were downloaded from the NCBI RefSeq database [166] using the *–assembly-source "RefSeq"* and *–assembly-level complete* options. The draft genome assemblies were downloaded from the NCBI GenBank database [166] using the *–assembly-source "GenBank"* option. The *–exclude-*

atypical and *–mag exclude* options were used in both cases. The number of draft genome assemblies for *S. pyogenes* available from GenBank was insufficient to create the complete dataset (n=16384) for the benchmark. Due to that, draft genome assemblies annotated as *Streptococcus pyogenes* were also downloaded from a collection of 661K genomes available on the European Nucleotide Archive (ENA) [175]. MLST v2.23.0 [129, 176] was used to determine the ST for all assemblies. Assemblies without a known ST or assigned an ST from a different species, indicating possible misannotation, were excluded. A custom Python script was also used to filter out assemblies based on a maximum number of contigs of 100, a maximum number of ambiguous bases of 1000, and a minimum and maximum genome size. The minimum and maximum genome size values were defined based on the *min_ungapped_length* and *max_ungapped_length* values in the "species_genome_size.txt" file available on NCBI's FTP on September 9, 2023 (`https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/`) [166].

2.6.2 Dataset creation

The selected draft genome assemblies were subsampled to create datasets to evaluate the performance of chewBBACA 3, chewBBACA 2 and pyMLST. The pairwise Average Nucleotide Identity (ANI) distances for each species' selected draft genomes were computed with Skani v0.2.1 [177]. To factor in the aligned genome fraction, weighted ANI values were computed by multiplying the ANI values by the mean of the query and reference aligned fractions. The weighted ANI values were ordered to select a set of 16,384 genomes that maximized the average pairwise distance. Smaller datasets were created by randomly sampling this dataset, starting by selecting 1 genome and doubling the dataset size until reaching a dataset size of 8,192. Five replicates were created for each dataset size. The complete datasets with 16,384 genomes were compressed with AGC v3.0 [178] to allow efficient storage and fast genome retrieval based on lists of genome identifiers.

2.6.3 Creation of wg/cgMLST schemas

A total of 260, 309 and 1,326 complete genomes for *Streptococcus pyogenes*, *Listeria monocytogenes* and *Salmonella enterica*, respectively, were selected for schema creation. wgMLST schema seeds were created with the *CreateSchema* module available in chewBBACA v3.3.6 and compared against the schema seeds created by the previous *CreateSchema* implementation, available in chewBBACA v2.6.0 [143]. The schema creation processes used a minimum sequence length value of 0 (*--l 0*) and the Prodigal [159] training files bundled with chewBBACA. The schema seeds created by both versions were compared based on a BSR ≥ 0.6 and a proportion of shared minimizers ≥ 0.9 to determine sets of loci shared by the schema seeds created with both versions. Schema seeds created with chewBBACA v3.3.0 were populated with the alleles identified in the complete genomes through allele

calling. The results of the allele calling were used to determine the set of core loci with the *ExtractCgMLST* module based on a loci presence threshold of 1 (*--t 1*) and create the cgMLST schemas used to evaluate the allele calling performance. The cgMLST schemas were adapted with the *PrepExternalSchema* module implemented in chewBBACA v2.8.5 to create the cgMLST schemas for that version. To create equivalent databases for pyMLST [167], multi-FASTA files with the first representative allele for each locus in the cgMLST schemas were passed to the wgMLST create command. The wgMLST add command was used to add each complete genome to the pyMLST databases.

2.6.4 External schema adaptation

The cgMLST schemas for *S. pyogenes*, *L. monocytogenes* and *S. enterica* available on the cgMLST.org server [153] were downloaded on July 4, 2024. These schemas were adapted with the *PrepExternalSchema* module available in chewBBACA v3.3.6 and compared against the schemas adapted with the previous *PrepExternalSchema* implementation, available in chewBBACA v2.0.17.2. The representativeness of the set of representative alleles selected by the *PrepExternalSchema* module was measured by aligning the representative alleles selected for each locus against all valid locus alleles based on a BSR ≥ 0.6.

2.6.5 Evaluation of the allele calling results

The cgMLST schemas and datasets containing between 1 and 16,384 draft genome assemblies were used to evaluate the allele calling performance of chewBBACA v3.3.3, chewBBACA v2.8.5 and pyMLST v2.1.5. The number of distinct CDSs per dataset was computed based on the CDSs predicted by Pyrodigal v3.0.0. Runtime, peak memory usage, and the comprehensiveness of the allele calling were evaluated for all datasets. The allelic profiles for the strains classified by pyMLST were extracted from the databases with the *wgMLST mlst* command and converted to the format used by chewBBACA with a custom script. The allelic profiles were masked to remove the *INF-* prefix from inferred alleles and to substitute all special classifications or missing values by 0. The core loci were defined with the *ExtractCgMLST* module based on the complete datasets' results and a loci presence threshold of 0.95. Loci below this threshold were considered to be part of the accessory genome. The pairwise Jaccard and allelic distances were computed with a custom script based on the masked allelic profiles. The proportion of classified CDSs and identified loci are based on the total number of CDSs predicted by Pyrodigal and on the total number of loci in each schema, respectively.

2.6.6 Download and analysis of *S. pyogenes emm1* strains

The genome assemblies and metadata for the *S. pyogenes* strains belonging to each lineage were recovered from previous studies [168, 169, 179]. The schema loci containing the lineage-defining SNPs were identified using BLASTp to align the translated CDSs from the MGAS5005 reference genome [180], with RefSeq accession number *GCF_000011765.3*, against the translated schema alleles.

2.6.7 Runtime and peak memory usage measurement

Runtime and peak memory usage were measured with the GNU time command on a desktop computer with an Intel® Core™ i7-4790 CPU, 32GB 1600 MT/s RAM, and a 1TB Samsung SSD 870 QVO. Any analysis that evaluated runtime and peak memory usage used 6 CPU cores to run chewBBACA 3 and chewBBACA 2 and 1 CPU core for pyMLST because the latter cannot use multiple cores.

2.7 Availability and requirements

Project name: chewBBACA 3.
Project home page: `https://github.com/B-UMMI/chewBBACA`
Project documentation: `https://chewbbaca.readthedocs.io/en/latest/index.html`
Operating system(s): Linux and macOS.
Programming language: Python >= 3.8
Other requirements: BLAST+ >= 2.9.0, pyrodigal>=3.0.0, numpy =1.24.3, scipy =1.10.1, biopython>=1.79, plotly>=5.8.0, SPARQLWrapper>=2.0.0, requests>=2.27.1, pandas>=1.5.1
License: GPL-3.0
Any restrictions to use by non-academics: None.

2.8 Declarations

2.8.1 Ethics approval and consent to participate

Not applicable.

2.8.2 Consent for publication

Not applicable.

2.8.3 Availability of data and materials

The datasets, schemas and databases created and used with chewBBACA 3, chewBBACA 2, and pyMLST, and all results generated for each section are available on Zenodo (`https://doi.org/10.5281/zenodo.14637859`) [170]. The supplementary figures and tables are included in the supplementary data.

2.8.4 Competing interests

MR received honoraria for serving on the speakers bureau of Pfizer and Merck Sharp and Dohme and for serving in expert panels of GlaxoSmithKline and Merck Sharp and Dohme. All other authors declare they have no competing interests.

2.8.5 Funding

This work was partly supported by the ISIDORe project (funding from the European Union's Horizon Europe Research & Innovation Programme, Grant Agreement no. 101046133). RM was supported by the Fundação para a Ciência e Tecnologia (FCT) (grant 2020.08493.BD).

2.8.6 Author's contributions

All authors contributed to the design of the tool. RM implemented, tested, and benchmarked the tool. PVC contributed to the implementation of the tool. RM and MR wrote the manuscript. All authors read, revised and approved the final manuscript.

2.9 Supplemental Material

2.9.1 Supplemental Figures

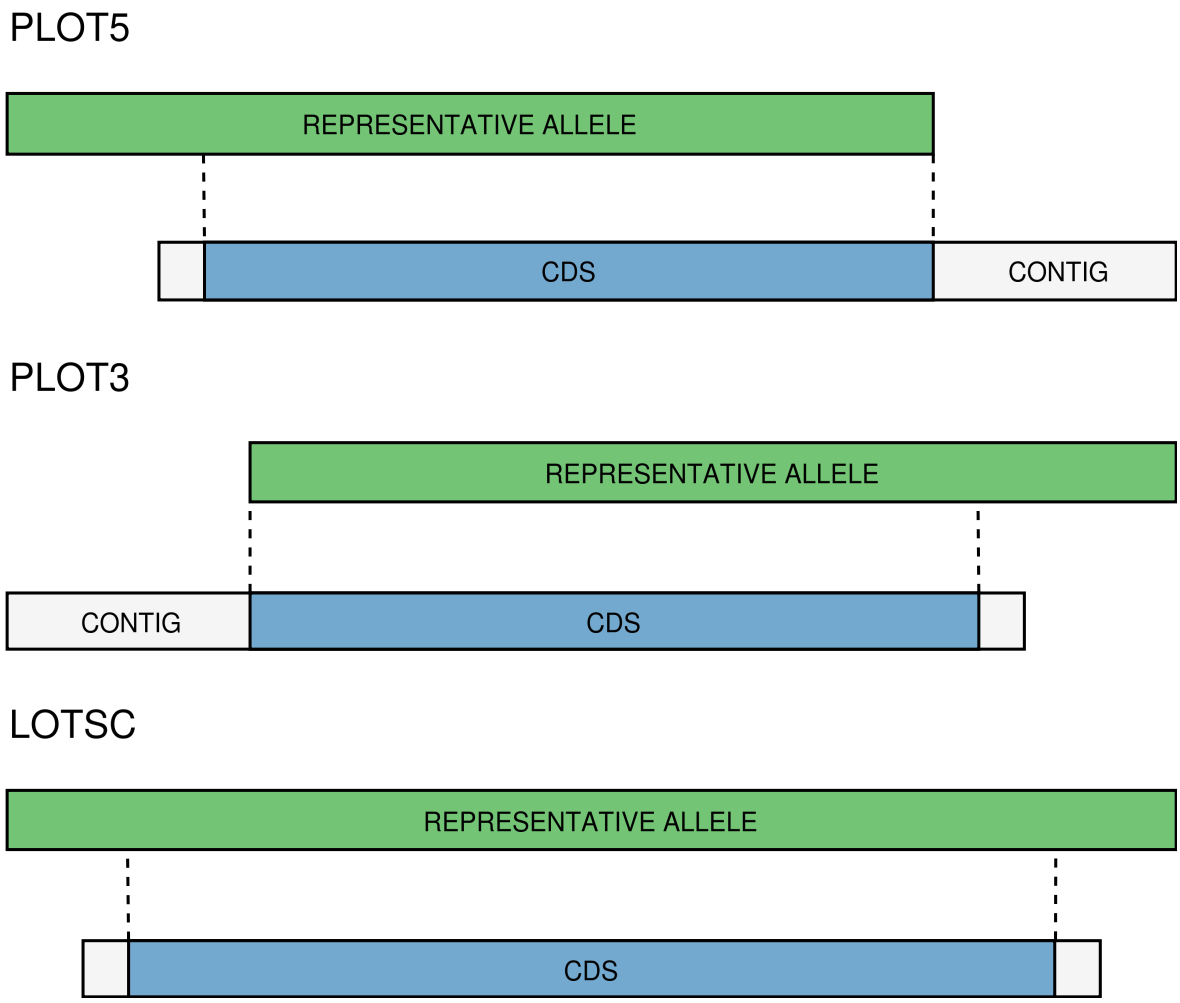


Figure 2.5: PLOT5, PLOT3 and LOTSC classifications. The PLOT3, PLOT5 and LOTSC classifications are related to the position of CDSs in the genomic contigs. PLOT5 and PLOT3 - a CDS is classified as PLOT5 or PLOT3 if it is close to the contig 5'- or 3'-end and if the unaligned portion of the matched representative allele exceeds the contig end. LOTSC - a CDS is classified as LOTSC if the matched representative allele is bigger than the contig containing the CDS.

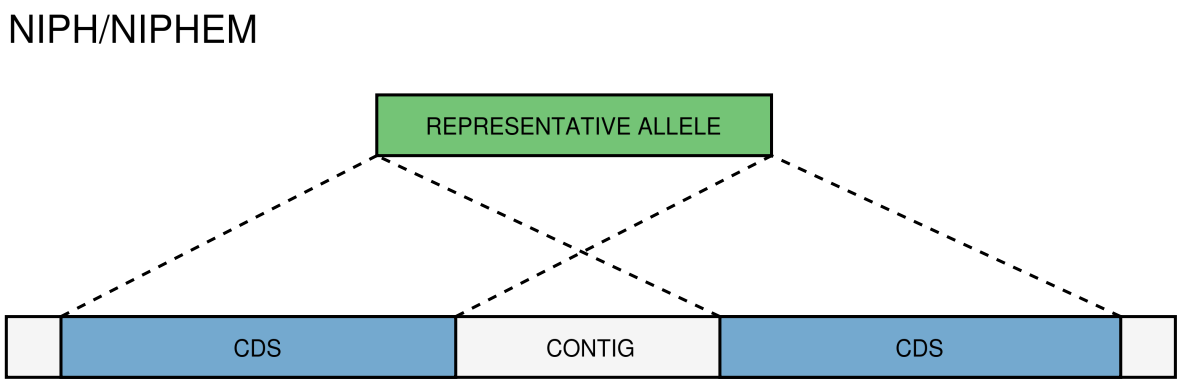


Figure 2.6: NIPH and NIPHEM classifications. The NIPH and NIPHEM classifications are assigned when multiple CDSs from the same genome match the same schema locus. NIPH - assigned when multiple CDSs from the same genome match a single locus. NIPHEM - assigned when multiple CDSs from the same genome are exact matches to alleles of a single locus.

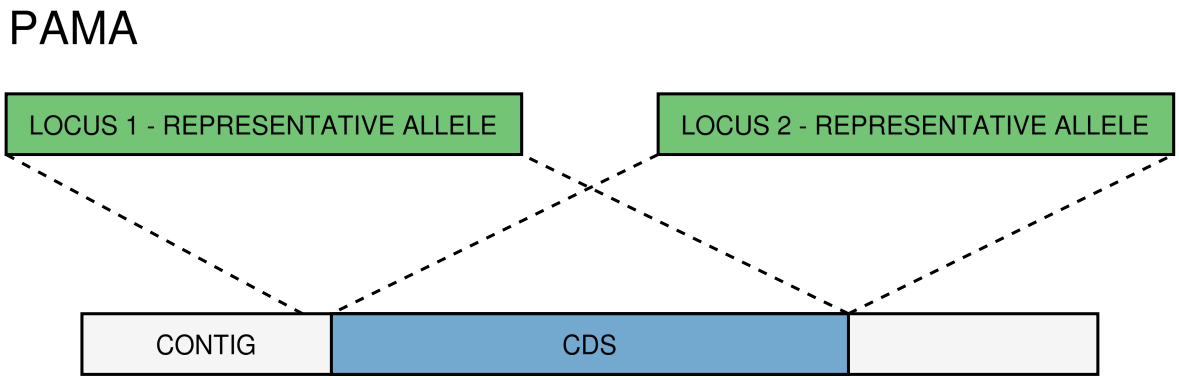


Figure 2.7: PAMA classification. The PAMA classification is assigned when a single CDS from a genome matches multiple schema loci.

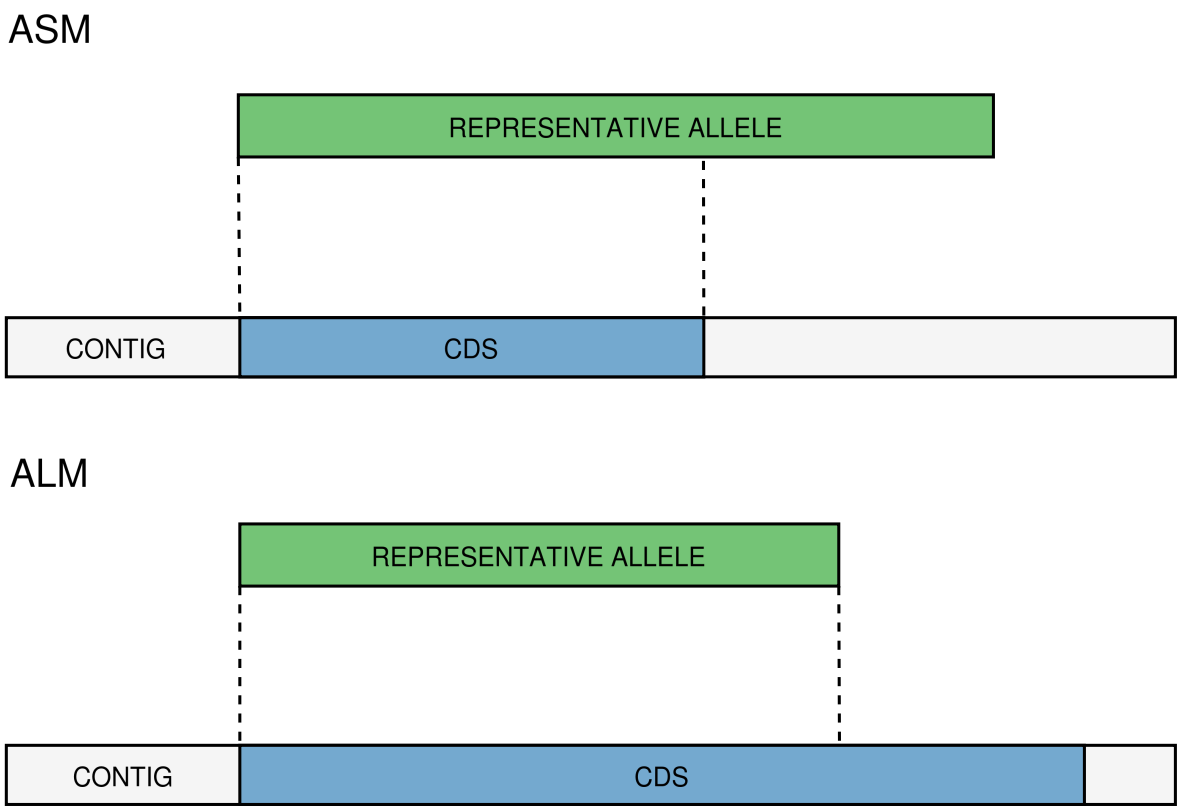


Figure 2.8: Allele Smaller than Mode (ASM) and Allele Larger than Mode (ALM) classifications. The ASM and ALM classifications are assigned when the size of a CDS that matches a schema locus is below or above the locus size variation interval, respectively. The default behaviour is to assign these classifications to alleles that are 20% shorter or longer than the locus allele size mode.

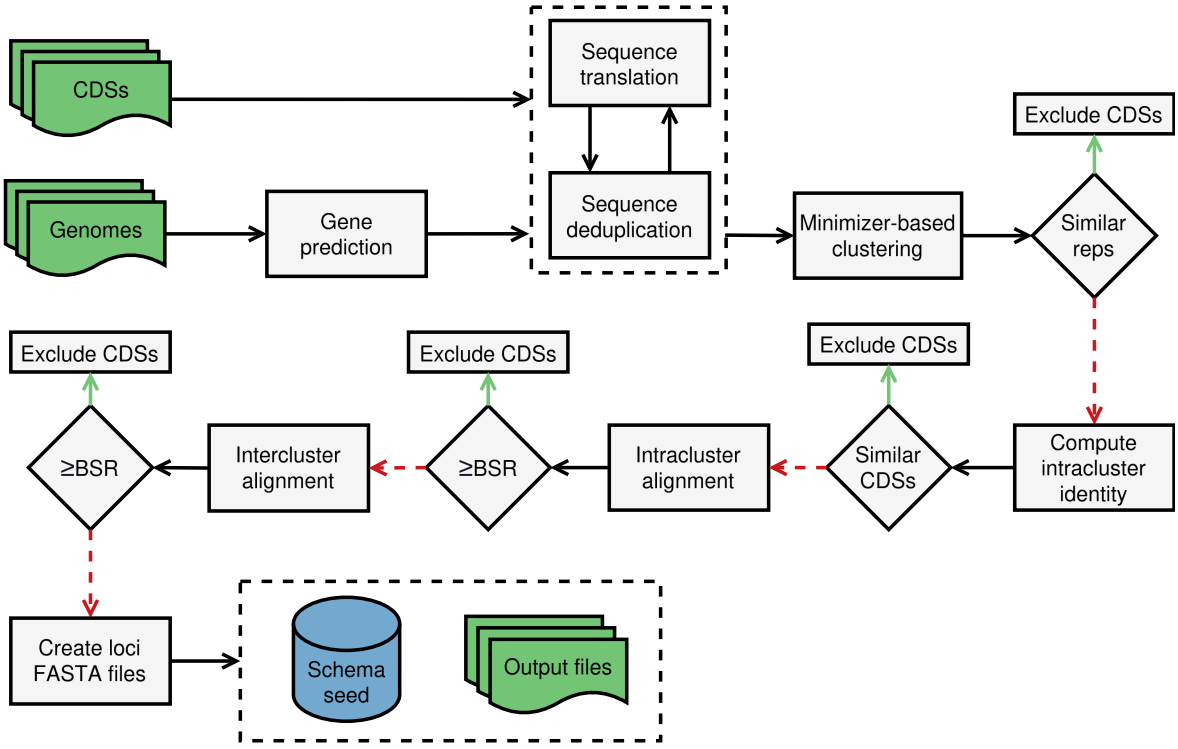


Figure 2.9: Diagram of the *CreateSchema* module. The *CreateSchema* module creates a schema seed based on a set of FASTA files with genome assemblies or CDSs. If genome assemblies are given, the process starts by predicting CDSs for each genome using Pyrodigal. The CDSs identified in the input files are deduplicated and translated, followed by a second deduplication step to determine the set of distinct translated CDSs. The distinct translated CDSs are clustered based on the proportion of minimizers shared with representative CDSs. The largest or one of the largest CDSs is selected as the first representative CDS. New representative CDSs are selected when CDSs share a low proportion (<0.2) of minimizers with any of the chosen representative CDSs. Non-representative CDSs that share a proportion of minimizers ≥ 0.9 with the cluster representative are considered to correspond to the same locus and are excluded from the analysis. The proportion of shared minimizers between non-representative CDSs is determined to exclude CDSs sharing a proportion of minimizers ≥ 0.9 with larger CDSs. Intracluster and intercluster alignment with BLASTp enable identifying and excluding CDSs similar to representative or larger non-representative CDSs based on a BSR ≥ 0.6. Each remaining CDS is considered to be an allele of a distinct locus. The process ends by creating a schema seed, which includes one FASTA file containing a single representative allele per distinct locus identified in the analysis. Green document icons represent input FASTA files and output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents the schema seed created by the *CreateSchema* module.

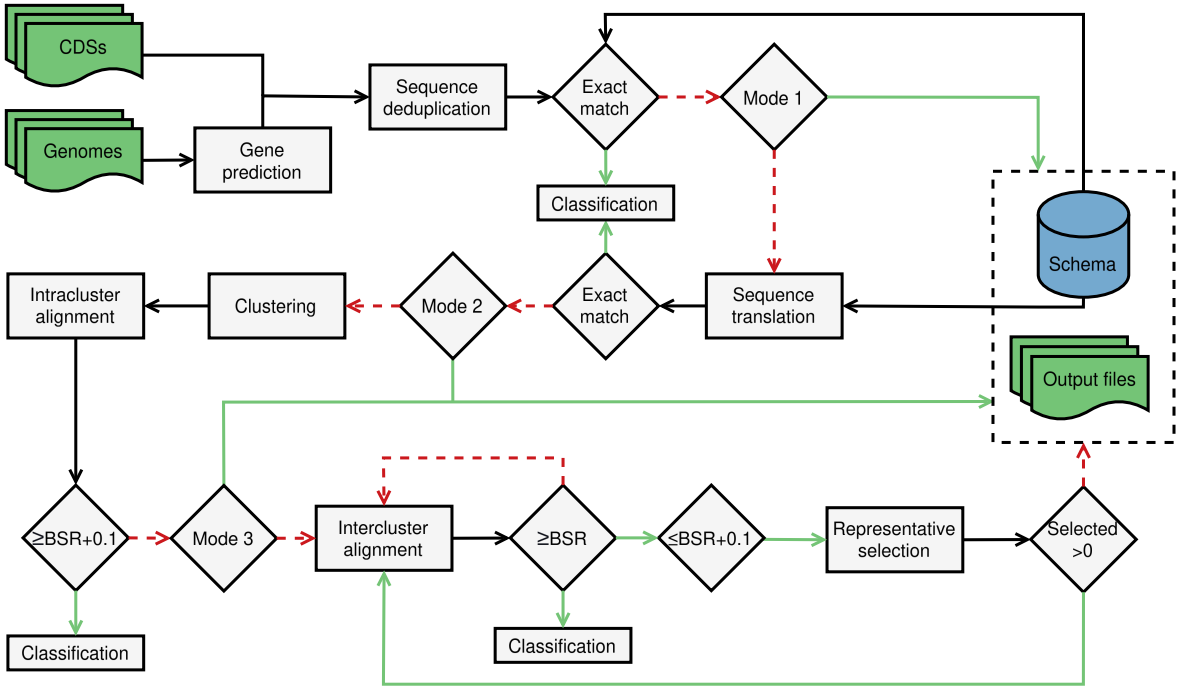


Figure 2.10: Diagram of the *AlleleCall* module. The *AlleleCall* module determines the allelic profiles for strains of interest. The process accepts FASTA files with genome assemblies or CDSs. If genome assemblies are given, the process starts by predicting CDSs for each genome using Pyrodigal. The CDSs identified in the input files are deduplicated and compared against the schema alleles to find and classify exact matches at the DNA level. If the process runs in mode 1, the results are evaluated to write the output files and exit. Otherwise, the CDSs that do not match any schema alleles at the DNA level are translated and matched against the translated schema alleles to find exact matches at the protein level. If the process runs in mode 2, the results are evaluated to write the output files, add new alleles to the schema and exit. Otherwise, the CDSs not classified through exact matching are compared against the schema representative alleles through minimizer-based clustering to identify CDSs that share a proportion of minimizers ≥ 0.2 with the representative alleles. Each cluster's representative allele is aligned against the clustered CDSs with BLASTp to classify CDSs based on the defined BSR value plus 0.1. At this point, if the process runs in mode 3, the results are evaluated to write the output files, add new alleles to the schema and exit. Otherwise, the representative alleles are aligned against the remaining unclassified CDSs to classify them based on the defined BSR value and identify new representative alleles whose BSR is not above the defined BSR value plus 0.1. If the process finds new representative alleles, it aligns them against the unclassified CDSs to find new matches. This process repeats until no new representative alleles are identified. When no new representative alleles are found, the process evaluates the results to create the output files, add new alleles to the schema, and exit. Green document icons represent input FASTA files and output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

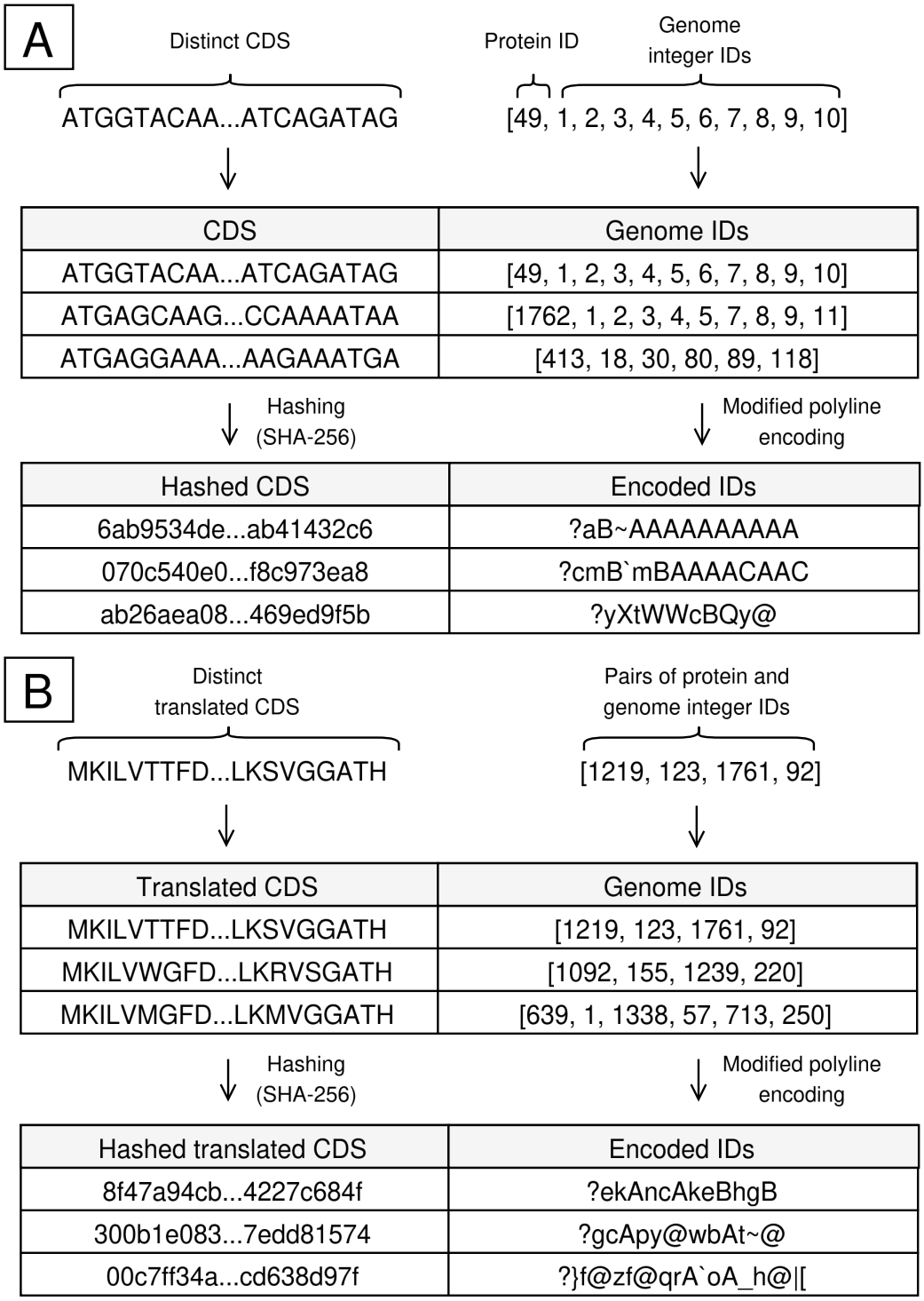


Figure 2.11: Sequence hashing and modified polyline encoding. (A) Each distinct CDS identified in the input genomes is hashed with the SHA-256 algorithm implemented in Python's *hashlib* library. The hash digest is obtained through the *hexdigest* method and mapped to the list of integer identifiers for the genomes containing the CDS encoded with modified polyline encoding. (B) After sequence translation and deduplication, each distinct translated CDS is hashed with the SHA-256 algorithm and the hash digest is mapped against lists with pairs of protein and genome identifiers used to identify each distinct CDS coding for the protein encoded with modified polyline encoding. The modified polyline encoding is applied to reduce the memory used to retain the data in-memory during the process, drastically reducing peak memory usage when processing large datasets. The Python dictionaries created to map the hashes to the lists of identifiers allow quick identification and classification of exact and inexact matches.

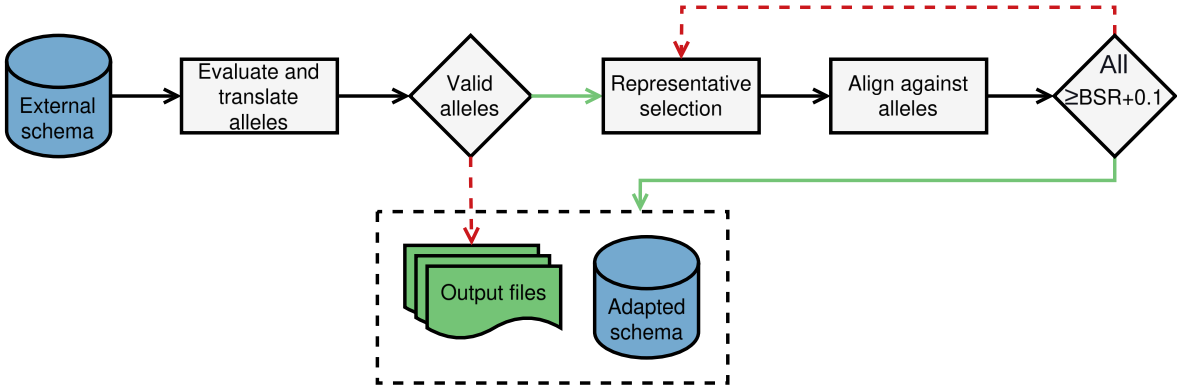


Figure 2.12: Diagram of the *PrepExternalSchema* module. The *PrepExternalSchema* module adapts schemas created with other wg/cgMLST tools or available on external platforms for usage with chewBBACA 3. The process starts by validating and translating the alleles in the external schema. Incomplete (i.e. size not multiple of 3) and invalid (i.e. missing the start or stop codons, or containing in-frame stop codons) alleles, alleles containing ambiguous bases or smaller than the specified minimum length value, are excluded. For each locus that has valid alleles, the process selects the largest or one of the largest alleles as the first representative allele. The representative is aligned against the locus' alleles with BLASTp to compute the BSR for each alignment. If all the BSR values are above the specified BSR plus 0.1, it is considered that the representative allele can adequately capture the diversity of the locus. Otherwise, new representative alleles are selected from those with a BSR above the specified BSR but below that value plus 0.1 to align against the locus' alleles and determine if the set of representative alleles selected captures the locus diversity adequately. Representative selection is repeated until all locus' alleles have a BSR above the specified value plus 0.1 with at least one of the selected representative alleles. The valid and selected representative alleles are written to FASTA files to create a schema compatible with chewBBACA. The list of invalid alleles, the list of loci excluded from the adapted schema due to having no valid alleles, and the number of total alleles and representative alleles per locus in the adapted schema are stored in output files. The green document icons represent output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icons represent schemas.

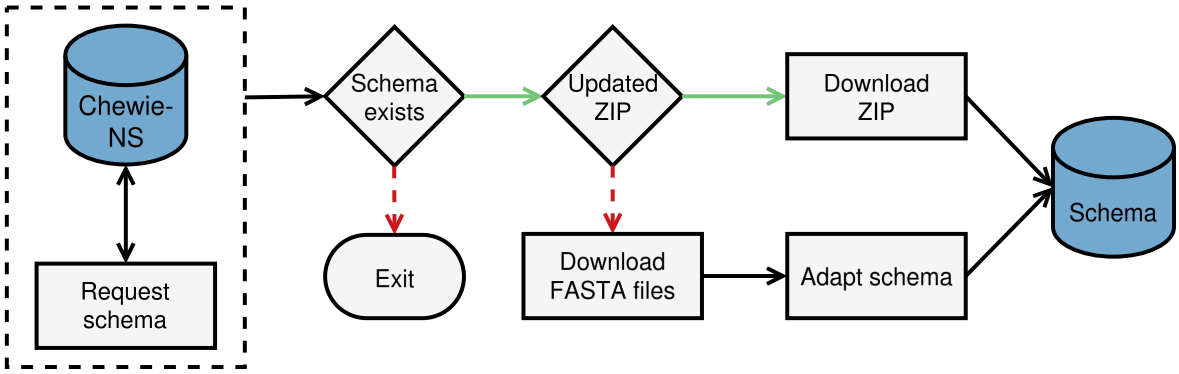


Figure 2.13: Diagram of the *DownloadSchema* module. The *DownloadSchema* module imports schemas from Chewie-NS. The process starts by sending a request with species and schema identifiers to Chewie-NS. If the schema exists, the process checks for a compressed and up-to-date version of the schema to download. If the compressed schema in Chewie-NS is for the latest version of the schema, the compressed schema is downloaded and uncompressed to get a ready-to-use schema. Otherwise, the process will send requests to retrieve the FASTA files with the alleles for all loci and determine the representative alleles with the *PrepExternalSchema* module to create the schema locally. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icons represent schemas.

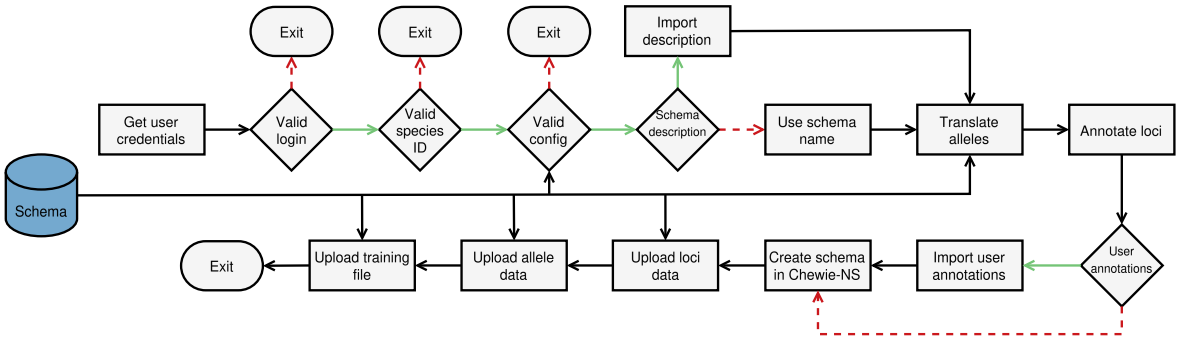


Figure 2.14: Diagram of the *LoadSchema* module. The *LoadSchema* module uploads local schemas to Chewie-NS. The process starts by requesting the user credentials to ensure that the user has contributor privileges. Only contributors are allowed to upload schemas to Chewie-NS. If the user is a contributor, the process checks if the species identifier provided by the user is valid and if the species is listed in Chewie-NS. After this step, the process reads the schema's configuration file to validate the schema parameter values and ensure that there is only a single value associated with each parameter. The initial validation steps are followed by the upload of the schema data to Chewie-NS. The process reads the schema description, if the user provided one, or uses the schema name as description. The alleles are translated and annotation terms for the loci are obtained through UniProt's SPARQL endpoint. If the user provides custom loci annotations, the process reads the file provided by the user and adds the custom annotations to the loci annotation data to send to Chewie-NS. After retrieving loci annotations, the process creates the schema in Chewie-NS by sending the schema's parameter values and the list of file hashes to validate schema files uploaded in subsequent steps. The loci are created and linked to the newly created schema by sending the loci identifiers and annotations to Chewie-NS. The loci FASTA files are compressed and uploaded to Chewie-NS to add the allele sequences to the database and link them to the corresponding loci. The last step in the process uploads the training file in the local schema and associates it to the newly created schema in Chewie-NS. After process completion, Chewie-NS will process the data that was sent to make the schema data and statistics available through the website and the API. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

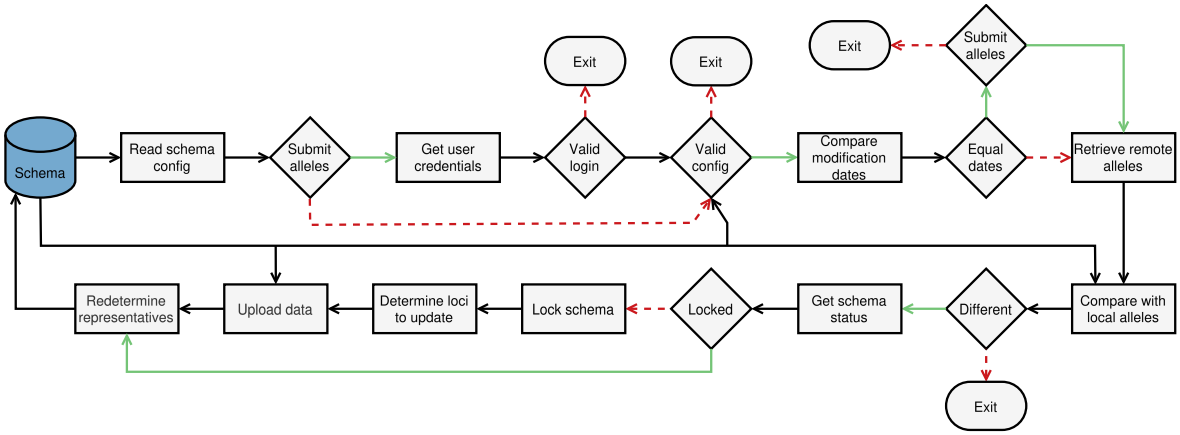


Figure 2.15: Diagram of the *SyncSchema* module. The *SyncSchema* module retrieves new alleles added to remote schemas in Chewie-NS and submits new alleles added to local schemas to update the remote schemas in Chewie-NS. The process starts by reading the schema's configuration file to get the schema's parameter values and ensure the values match the ones listed in Chewie-NS. If the user wants to submit new alleles identified locally (--submit), the process will ask for the user credentials to verify if the user has contributor privileges. Before retrieving or uploading new alleles, the process verifies if the last modification date of the local and remote schemas match. If the dates match and the user does not want to submit new local alleles, the process exits. If the dates do not match or the user wants to submit new local alleles, the process retrieves new alleles added to the remote schema since the last modification date and compares them with the alleles in the local schema. If any alleles are exclusive to the local or remote schema, the process creates updated FASTA files with all the alleles and locks the remote schema to ensure that only the current user can modify the remote schema. The process creates files with the data for the new local alleles and sends them to Chewie-NS, waiting for Chewie-NS to insert the new alleles into the database. After allele insertion in Chewie-NS, the process adapts the updated FASTA files with the *PrepExternalSchema* module to update the local schema and ensure that the local and remote allele identifiers match. If the schema was already locked by another user, the process will skip data upload to Chewie-NS and will update the local schema with new alleles retrieved from Chewie-NS. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

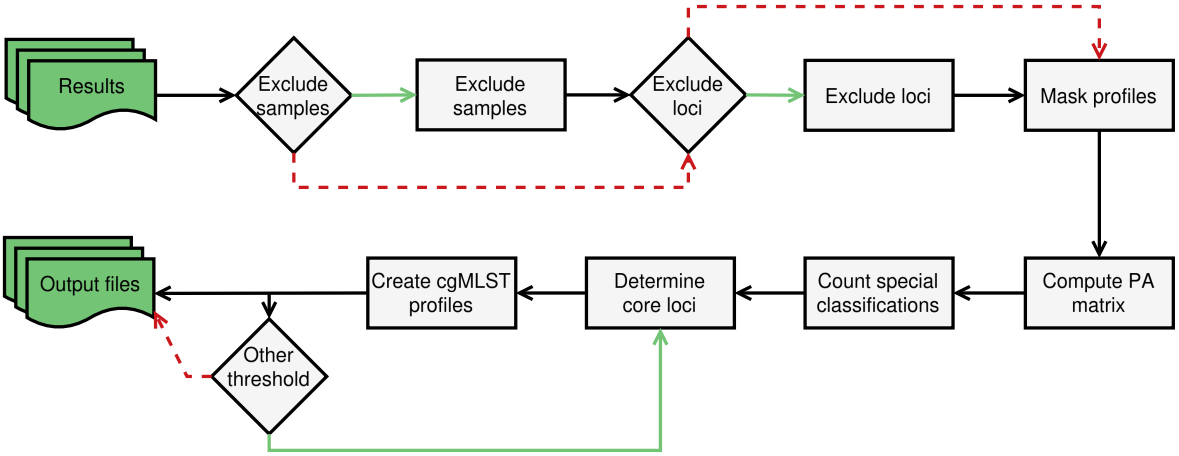


Figure 2.16: Diagram of the *ExtractCgMLST* module. The *ExtractCgMLST* module determines the set of core loci based on the allelic profiles determined by the *AlleleCall* module. The process starts by excluding loci and samples from the analysis based on lists of loci and samples provided by the user. This allows users to filter out low-quality samples and problematic loci that would affect the determination of the core genome. The filtered allelic profiles are masked to remove the INF- prefixes from newly inferred alleles and substitute special classifications by 0. The masked profiles are used to compute a loci presence-absence matrix and count the number of special classifications per sample. The presence-absence matrix is also used to determine the set of core loci based on the default loci presence thresholds of 0.9, 0.95 and 1, or based on threshold values specified by the user. The process creates output files with the list of loci and allelic profiles per threshold and creates an HyperText Markup Language (HTML) file with a scatter plot representing the core genome size variation for each threshold. The green document icons represent input and output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise.

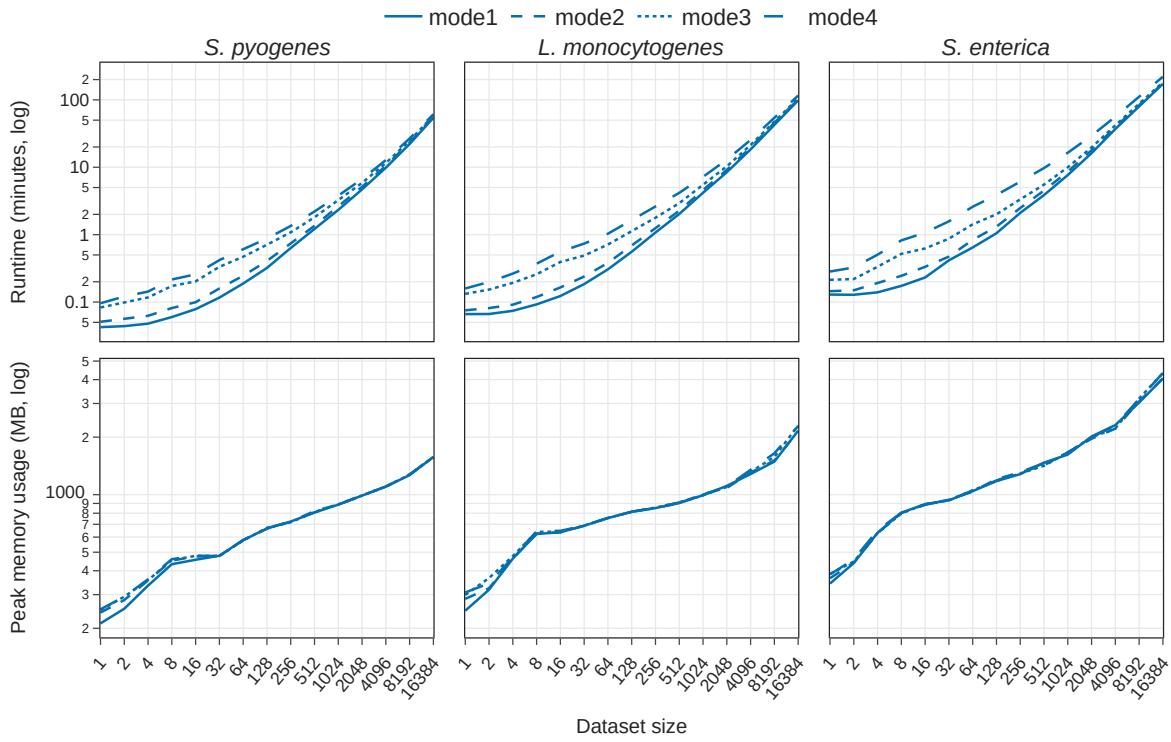


Figure 2.17: Runtime and peak memory usage for the four execution modes available in chewBBACA 3. Runtime and peak memory usage were measured for the allele calling of datasets with 1 to 16384 strains for three bacterial species: *Streptococcus pyogenes*, *Listeria monocytogenes*, and *Salmonella enterica*. The benchmark was performed with five replicates per dataset size, except for the complete dataset (n=16,384 genomes). The values shown are the mean of the replicate values for each dataset. Runtime was measured as the elapsed real time in minutes (logarithmic scale). Peak memory usage was measured as the maximum resident set size in MB (logarithmic scale).

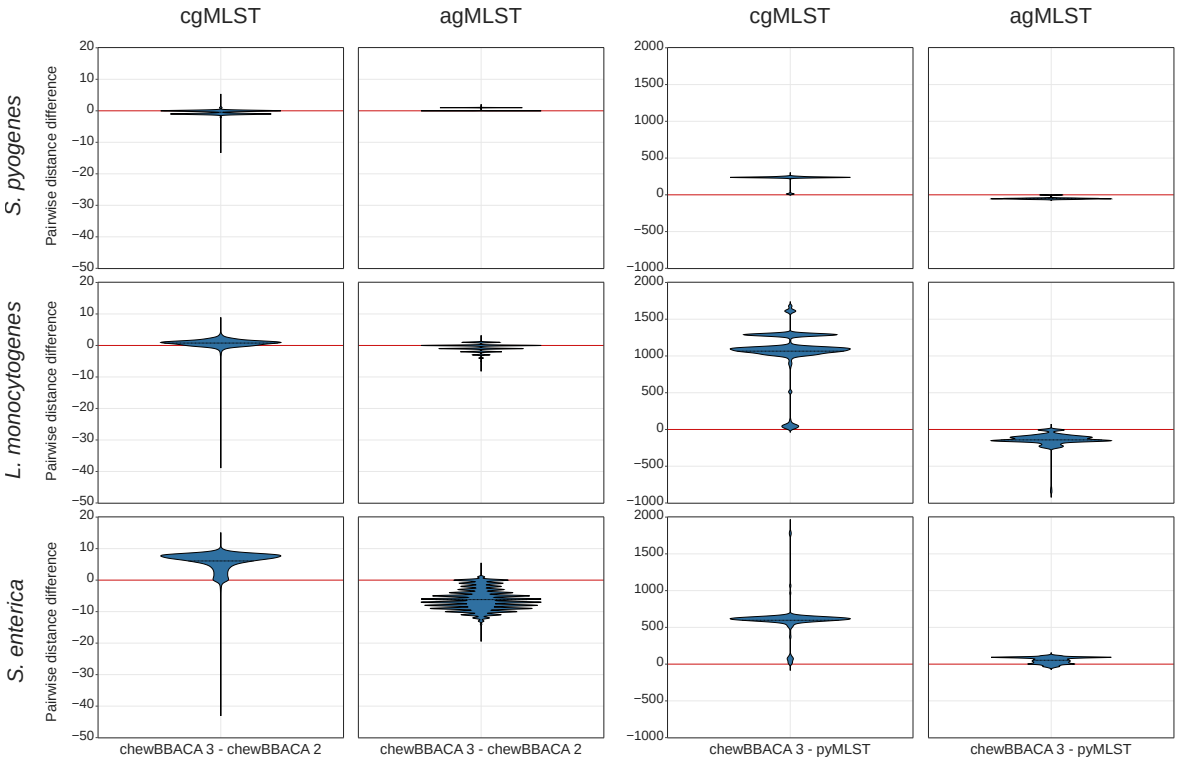


Figure 2.18: Pairwise allelic distances differences. The pairwise distances differences at the core-genome (cgMLST) and accessory-genome (agMLST) levels were computed by subtracting the allelic distance matrices computed based on chewBBACA 2's and pyMLST's results from the allelic distance matrices computed from chewBBACA 3's results for the complete datasets (n=16,384 genomes). A positive value represents a greater difference with chewBBACA 3 and a negative value a smaller difference with chewBBACA 3 than with the comparator. The zero line in each plot is highlighted in red.

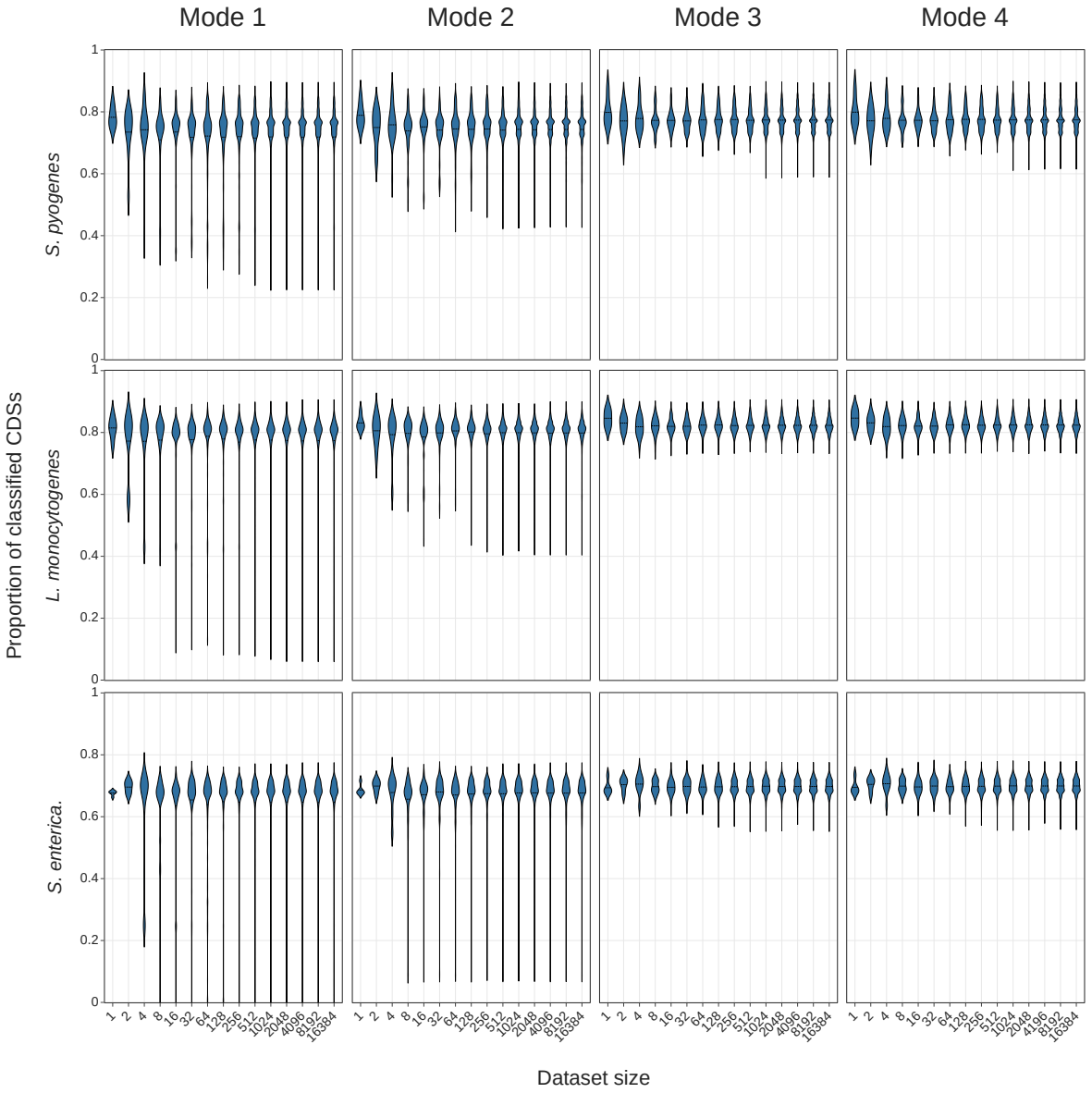


Figure 2.19: Proportion of CDSs classified per execution mode for each species' datasets. The proportion of classified CDSs corresponds to the number of CDSs classified by each execution mode divided by the total number of CDSs predicted for each strain by Pyrodigal. The benchmark was performed with five replicates per dataset size, except for the complete dataset (n=16,384 genomes).

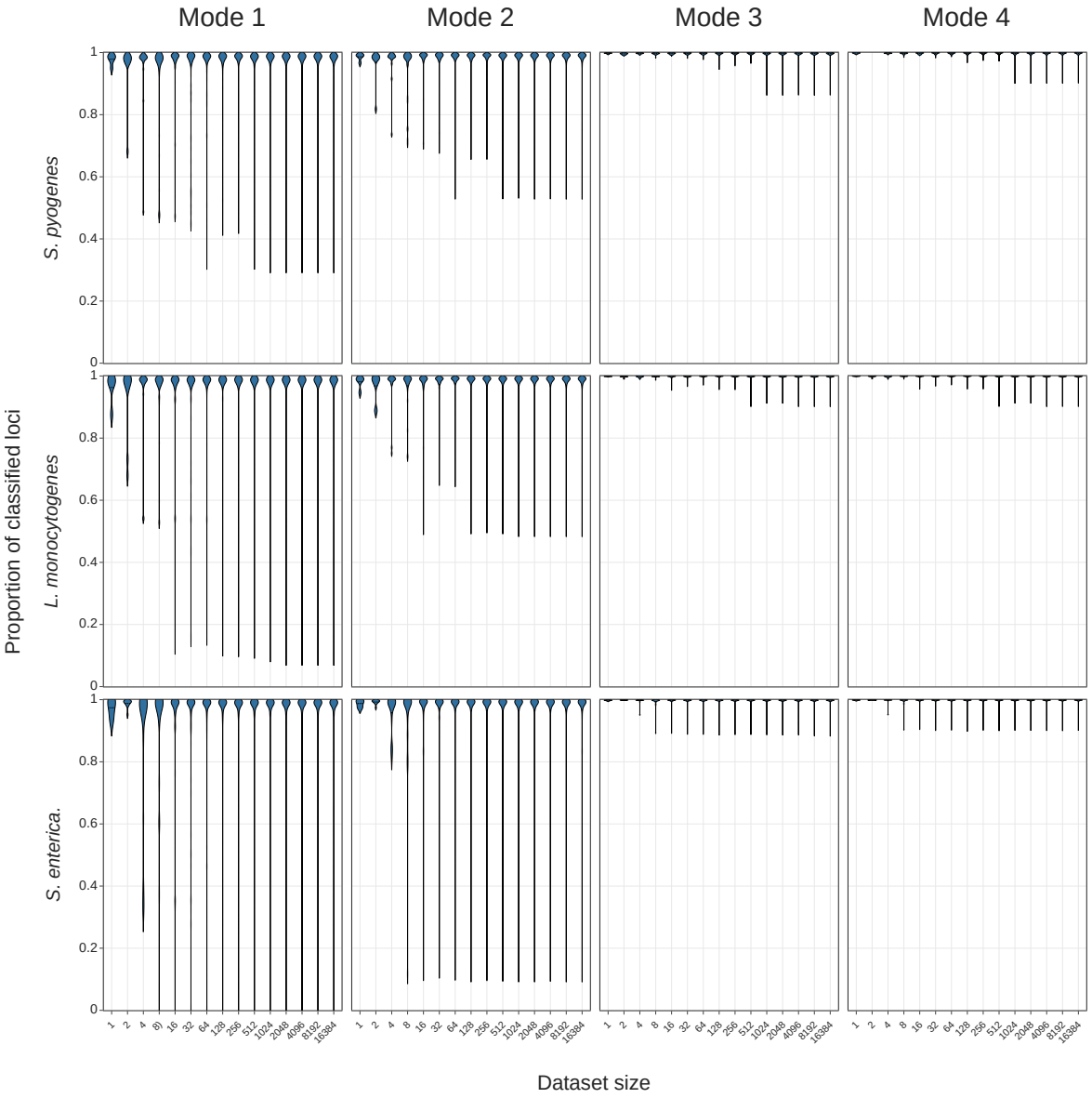


Figure 2.20: Proportion of schema loci classified per execution mode for each species' datasets. The proportion of classified loci corresponds to the number of schema loci identified by each execution mode divided by the total number of schema loci. The benchmark was performed with five replicates per dataset size, except for the complete dataset (n=16,384 genomes).

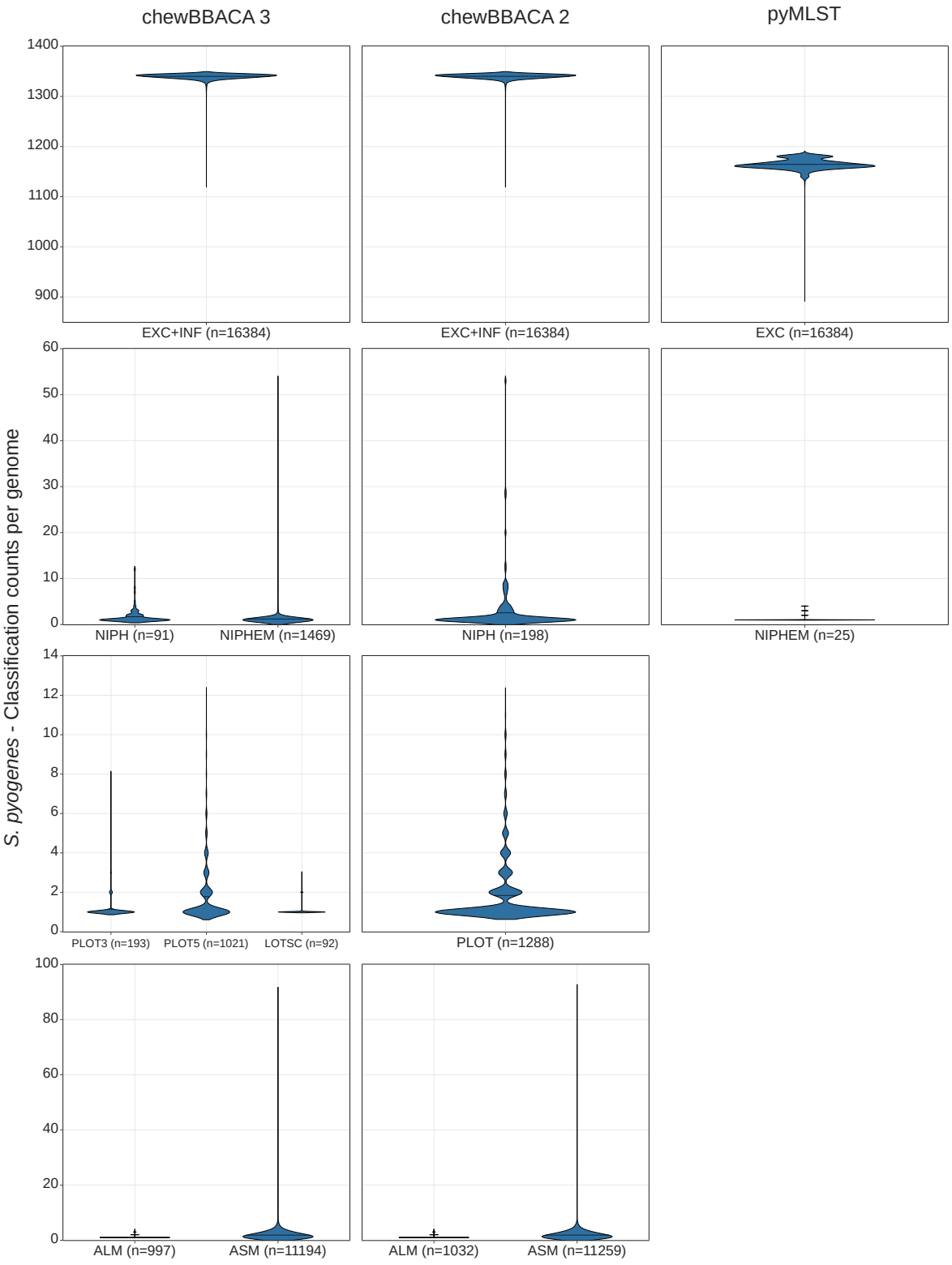


Figure 2.21: Classifications counts for the complete dataset (n=16,384 genomes) of *S. pyogenes* per tool. Each row displays the counts for the special classifications that are equivalent between tools. The x-axis labels show the names of the classifications and the number of genomes with a count above zero inside the parentheses (i.e. genomes with a count of zero for any of the classifications are not included in the plotted values). For pyMLST, the loci with a single matching CDS were converted to Exact Match (EXC) and the loci with multiple matches were converted to NIPHEM. The plot is not shown if the tool does not determine a special classification equivalent to the ones displayed in the row.

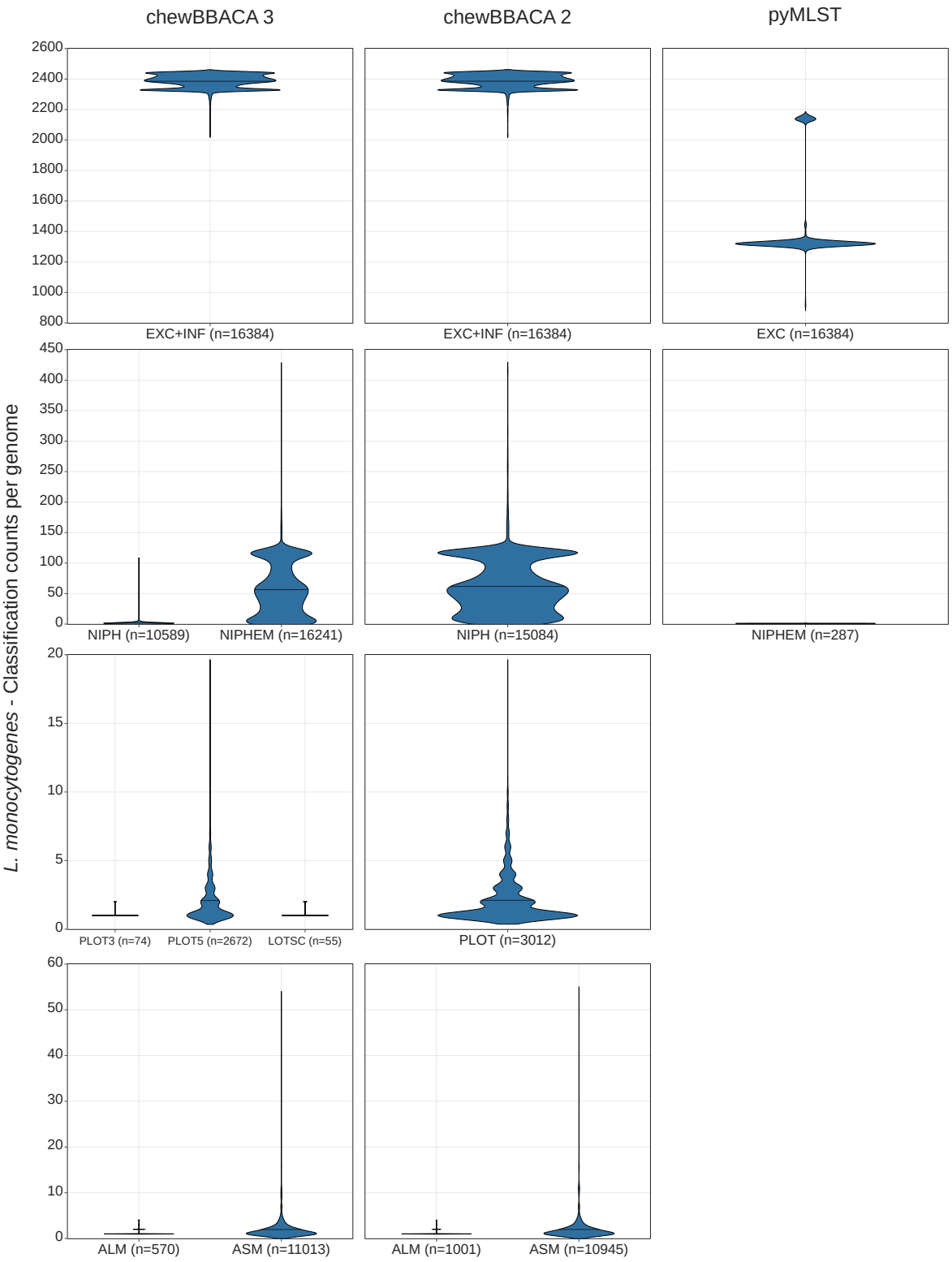


Figure 2.22: Classifications counts for the complete dataset (n=16,384 genomes) of *L. monocytogenes* per tool. Each row displays the counts for the special classifications that are equivalent between tools. The x-axis labels show the names of the classifications and the number of genomes with a count above zero inside the parentheses (i.e. genomes with a count of zero for any of the classifications are not included in the plotted values). For pyMLST, the loci with a single matching CDS were converted to EXC and the loci with multiple matches were converted to NIPHEM. The plot is not shown if the tool does not determine a special classification equivalent to the ones displayed in the row.

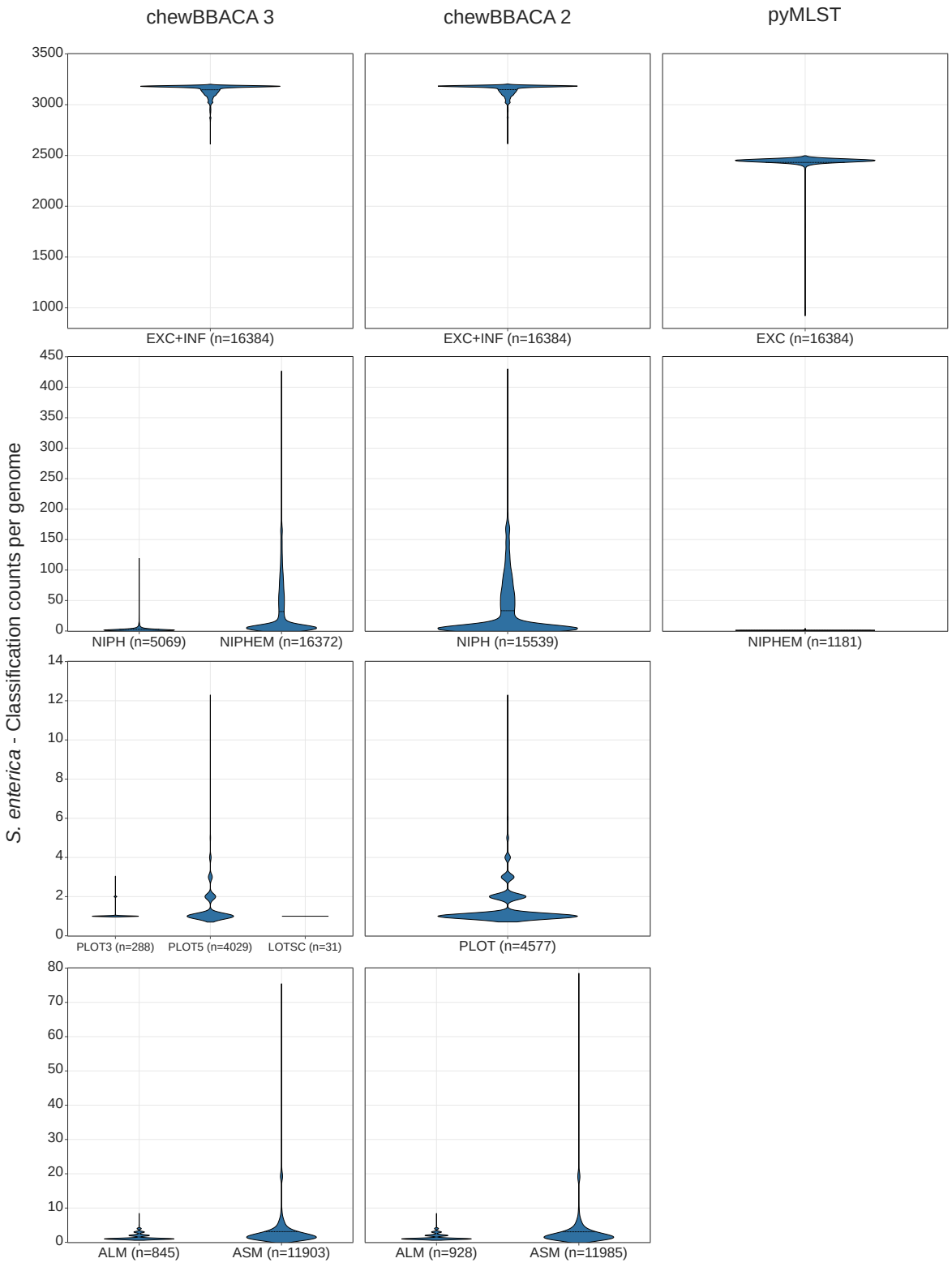


Figure 2.23: Classifications counts for the complete dataset (n=16,384 genomes) of *S. enterica* per tool. Each row displays the counts for the special classifications that are equivalent between tools. The x-axis labels show the names of the classifications and the number of genomes with a count above zero inside the parentheses (i.e. genomes with a count of zero for any of the classifications are not included in the plotted values). For pyMLST, the loci with a single matching CDS were converted to EXC and the loci with multiple matches were converted to NIPHEM. The plot is not shown if the tool does not determine a special classification equivalent to the ones displayed in the row.

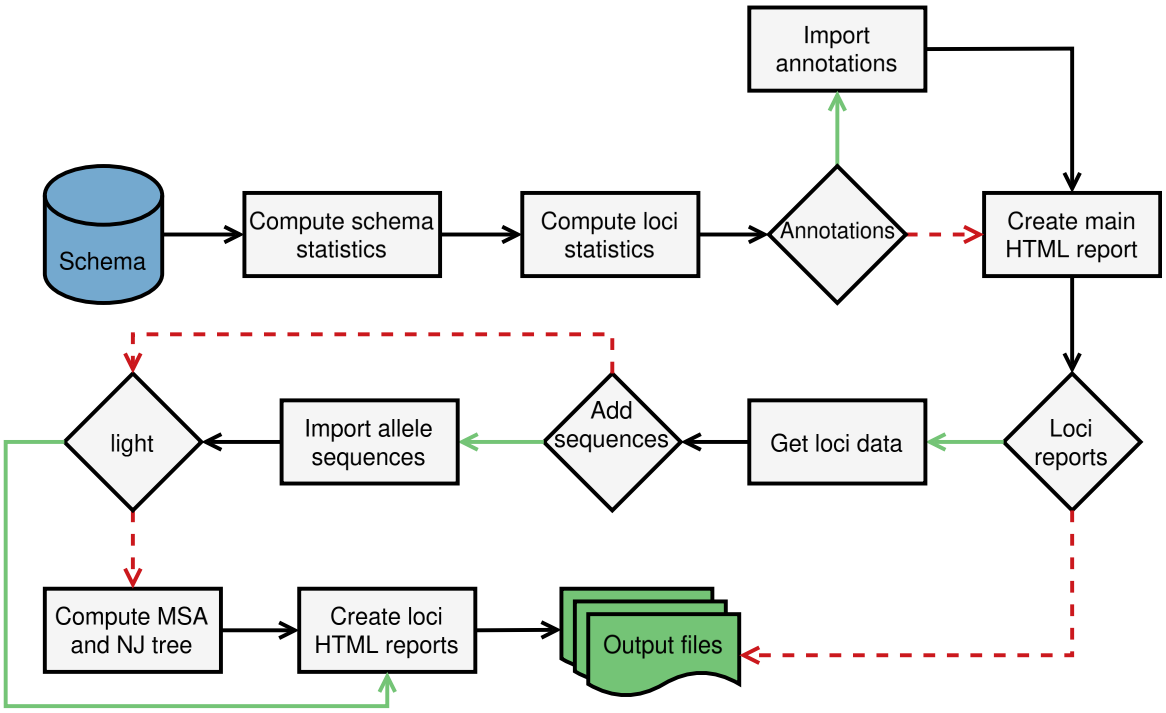


Figure 2.24: Diagram of the *SchemaEvaluator* module. The *SchemaEvaluator* module analyses a schema to create a report that allows users to explore schema structure and loci diversity interactively. The process starts by computing schema statistics, such as the number of loci and alleles, and loci statistics, such as the number of alleles, allele size statistics, and the number of valid and invalid alleles (e.g. alleles that cannot be translated due to being incomplete, containing ambiguous bases, in-frame stop codons, etc.). The schema and loci statistics are included in interactive data tables and charts on the main page of the HTML report. Loci annotations are imported and included in the main page of the report if provided. If the *--loci-reports* option is provided, the process performs a detailed analysis of each locus to add a separate locus page to the HTML report for each locus. Loci data is analyzed in greater detail to get more detailed statistics per locus. If the *--add-sequences* option is provided, the allele DNA sequences are imported and translated to add DNA and protein sequences to code editors on the locus page, which facilitates identifying and manipulating alleles of interest. Additionally, the process computes a MSA for each locus at the protein level with MAFFT to display the MSA and MAFFT's guide tree on interactive components. The MSA and guide tree are not displayed if the *--light* option is provided. The green document icons represent output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

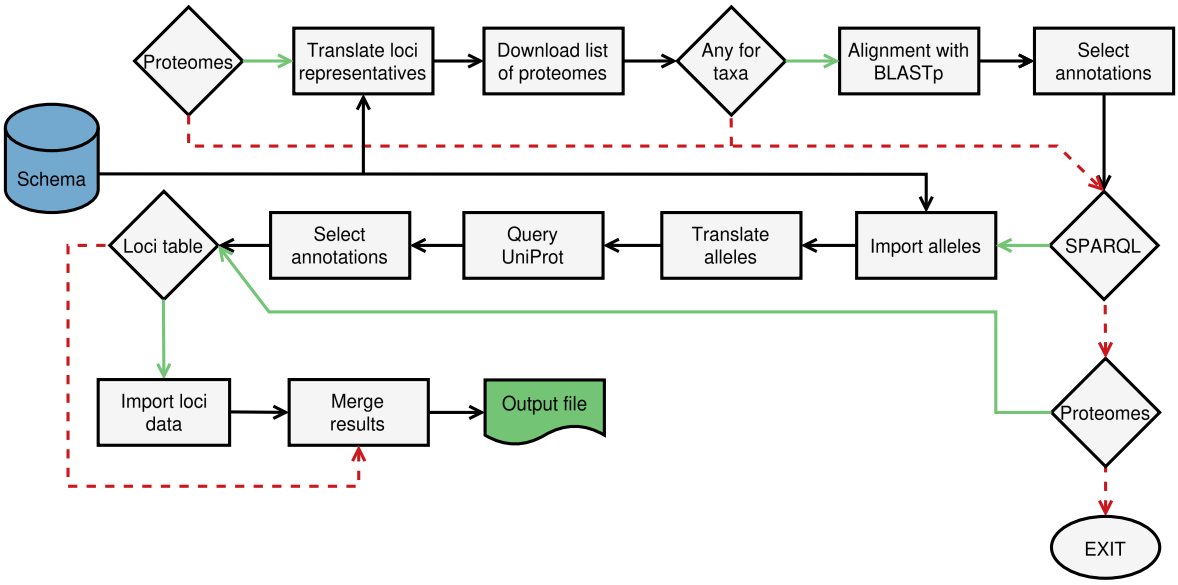


Figure 2.25: Diagram of the *UniprotFinder* module. The *UniprotFinder* module determines annotations for schema loci. The module offers two options to determine annotations: aligning against UniProt's reference proteomes and exact matching through UniProt's SPARQL endpoint. Users must provide at least one valid taxon name to annotate based on the reference proteomes. The process downloads the list of reference proteomes and searches for proteomes for the specified taxa. If there are any proteomes for the specified taxa, they are downloaded, and the loci representative alleles are aligned against the reference proteomes so annotations can be selected based on the BSR. The process searches for annotations through UniProt's SPARQL endpoint by creating queries including the loci alleles and submitting requests to the endpoint. If an allele matches any protein in UniProt, the annotation terms are extracted from the results. The process tries to select the most informative annotation terms. The annotation terms found through both options are merged to create a single annotations table. If the user provides a Tab-Separated Values (TSV) file with additional loci data, such as the file with CDS coordinates created by the *CreateSchema* and *AlleleCall* modules, the process will add the data in that file to the annotations table. The green document icon represents the output file. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

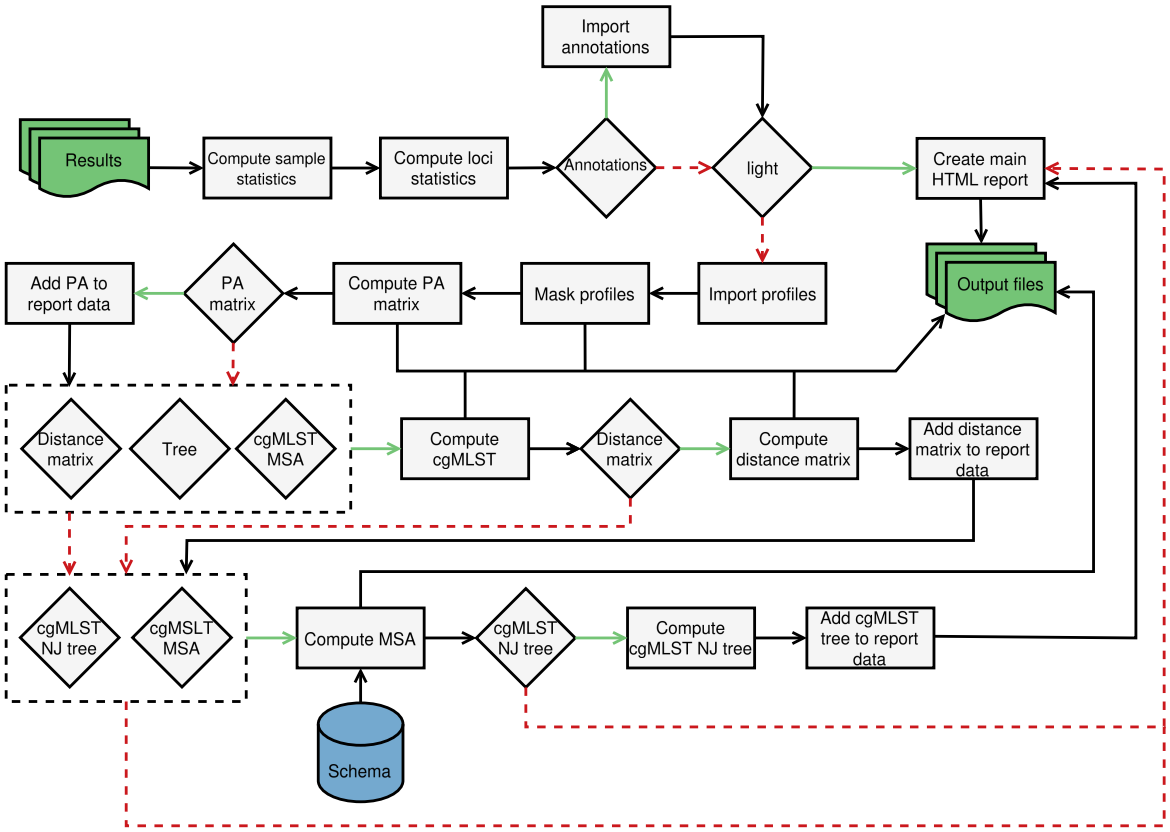


Figure 2.26: Diagram of the *AlleleCallEvaluator* module. The *AlleleCallEvaluator* module analyses allele calling results to create a report that allows users to explore results interactively. The process starts by importing and computing sample and loci statistics based on the allele calling results. The sample and loci statistics are included in interactive data tables and charts on the main page of the HTML report. Loci annotations are imported and included in the main page of the report if provided. If the *--light* option is provided, the process does not add more information to the report. Otherwise, the allelic profiles are imported and masked to remove *INF-* prefixes and substitute special classifications by 0. The masked profiles serve as the basis for computing a presence-absence matrix, enabling the determination of the set of loci that constitute the core genome. The profile data for the core loci are used to compute a matrix of allelic distances. The core loci alleles identified per strain and locus are imported to compute the cgMLST alignment that FastTree uses to compute a NJ tree. The presence-absence and distance matrices and NJ tree data are included in the report to be displayed and explored interactively. The green document icons represent input and output files. Grey rectangle icons represent analysis steps. Diamond icons represent conditional statements, with green arrows used when the condition is met and red dashed arrows otherwise. The blue cylinder icon represents a schema.

2.9.2 Supplemental Tables

Table 2.1: Runtime (in minutes, min) and peak memory usage (in megabytes, MB) values for the creation of the schema seeds with chewBBACA 2 and chewBBACA 3 based on the complete genomes for each species.

	chewBBACA 3										chewBBACA 2						pyMLST	
Species	EXC	INF	PLOT3	PLOT5	LOTSC	NIPH	NIPHEM	ALM	ASM	PAMA	EXC	INF	PLOT	NIPH	ALM	ASM	EXC	NIPHEM
S. pyogenes	21735280	218736	221	1817	98	154	1757	1170	20841	0	21736576	218665	2370	504	1210	20994	19078994	31
L. monocytogenes	38543696	538769	78	5594	59	23357	918118	690	21541	0	38552279	538435	6349	932267	1149	21611	23342559	289
S. enterica	50639542	919227	308	5770	31	14416	523022	1309	37156	31	50663361	916359	6806	517896	1437	36854	39838200	1200

Table 2.2: Number of loci in the schema seeds, number of loci shared between schema seeds based on the BSR, minimizers, and both, and percentage of loci in schema seeds created by chewBBACA 2 that are shared with the schema seeds created by chewBBACA 3.

	Number of loci		Shared loci			
Species	chewBBACA 2	chewBBACA 3	BSR	Minimizers	BSR U Minimizers	Shared loci (%)
S. pyogenes	3365	3688	3342	3040	3343	90.6
L. monocytogenes	5411	5760	5374	4416	5375	93.3
S. enterica	21086	23373	20812	19041	20814	89.1

Table 2.3: Runtime (in minutes, min) and peak memory usage (in megabytes, MB) for the adaptation of the schemas downloaded from cgMLST.org with chewBBACA 2 and chewBBACA 3.

	Runtime (min)		Memory (MB)	
Species	chewBBACA 2	chewBBACA 3	chewBBACA 2	chewBBACA 3
S. pyogenes	630.94	0.47	91.39	128.58
L. monocytogenes	6218.59	2.07	92.05	143.55
S. enterica	17428.95	4.51	93.44	147.23

Table 2.4: Number of loci in the schemas download from cgMLST.org, number of loci in the adapted schemas, and number of loci whose diversity is not completly captured by the selected representative alleles.

		Adapted loci		Loci not totally covered by selected representative alleles	
Species	Number of loci in original schema (cgMLST.org)	chewBBACA 2	chewBBACA 3	chewBBACA 2	chewBBACA 3
S. pyogenes	1095	1095	1095	1	0
L. monocytogenes	1701	1701	1701	15	0
S. enterica	3002	2999	2999	64	0

Table 2.5: Number of loci in the wgMLST and cgMLST schemas, and number of alleles in the cgMLST schemas after performing allele calling with each tool with the complete genomes.

	Number of loci		Number of alleles in cgMLST		
Species	wgMLST	cgMLST	chewBBACA 3	chewBBACA 2	pyMLST
S. pyogenes	3688	1345	71092	71092	60292
L. monocytogenes	5760	2449	87376	87376	52477
S. enterica	23373	3192	341247	341247	251502

Table 2.6: Mean runtime values in minutes for the allele calling of each species' datasets with chewBBACA 3, chewBBACA 2 and pyMLST.

	S. pyogenes			*L. monocytogenes*			*S. enterica*		
Dataset size	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST
1	0.10	0.19	0.17	0.16	0.37	0.20	0.29	0.57	0.51
2	0.12	0.23	0.32	0.20	0.48	0.42	0.33	0.62	1.23
4	0.14	0.27	0.61	0.26	0.62	0.84	0.50	1.32	2.11
8	0.22	0.48	1.24	0.36	0.94	1.72	0.80	2.14	3.90
16	0.26	0.65	2.46	0.56	1.78	3.51	1.07	3.28	7.21
32	0.40	1.30	4.93	0.73	2.77	6.89	1.58	5.32	14.18
64	0.59	2.37	9.86	1.03	4.75	13.73	2.60	11.17	28.47
128	0.87	4.68	20.13	1.64	9.96	27.39	3.88	20.34	56.03
256	1.38	8.52	41.05	2.59	19.69	54.74	6.16	41.01	111.64
512	2.20	16.36	82.09	4.18	36.14	109.26	9.73	77.05	222.02
1024	3.71	29.11	163.95	7.18	69.58	218.93	16.51	146.31	445.61
2048	6.67	52.25	329.27	12.97	134.03	439.27	29.37	276.28	891.29
4096	12.77	94.81	663.53	25.29	255.43	879.07	56.69	527.09	1786.59
8192	26.26	175.87	1339.27	54.45	501.13	1762.43	109.52	1463.99	3578.79
16384	60.63	342.76	2680.91	115.65	2347.44	3531.93	216.80	4227.74	7171.94

Table 2.7: Mean values for the total number of coding sequences (CDSs), distinct number of CDSs, and percentage of total CDSs that are distinct for each species' datasets.

	S. pyogenes			*L. monocytogenes*			*S. enterica*		
Dataset size	Total CDSs	Distinct CDSs	Distinct CDSs (%)	Total CDSs	Distinct CDSs	Distinct CDSs (%)	Total CDSs	Distinct CDSs	Distinct CDSs (%)
1	1684.8	1683.8	99.9	2966.2	2900.4	97.8	4651.8	4593.8	98.8
2	3487.8	3332.8	95.6	6037.8	5381.4	89.2	9141	8745.2	95.7
4	6909.8	5667.4	82.2	12211.2	9292.4	76.1	18188.8	15520.6	85.4
8	13915.6	10406.8	74.9	24325	15508.8	63.8	36792.2	28367.2	77.2
16	27815.4	14623.2	52.6	48774.6	22782	46.7	73943.6	45437.8	61.5
32	55772.8	22583	40.5	97534.4	32150.6	33.0	147210.6	74236.6	50.4
64	111138.4	37902.4	34.1	194415.6	42912.2	22.1	295221.8	121113.2	41.0
128	221802.8	51898.4	23.4	388199.4	63592.6	16.4	589450.4	183784.6	31.2
256	443839.8	75020	16.9	778822.6	91800.6	11.8	1179475.4	269535.4	22.9
512	890029.2	106781.6	12.0	1556165.4	127843.8	8.2	2357310.8	383924.2	16.3
1024	1777989.2	146328.4	8.2	3109809.6	188643	6.1	4707795.8	540833	11.5
2048	3557437.8	197893.8	5.6	6222051.4	280055.4	4.5	9428540.8	759312.4	8.1
4096	7113644	261355.4	3.7	12447727	404021.6	3.3	18853371.4	1068649.6	5.7
8192	14226635.6	339833.2	2.4	24891444.2	585723.2	2.4	37705899	1529198.2	4.1
16394	28451923	430352	1.5	49787476	861124	1.7	75417045	2118041	2.8

Table 2.8: Mean peak memory usage values in megabytes for the allele calling of each species' datasets with chewBBACA 3, chewBBACA 2 and pyMLST.

	S. pyogenes			*L. monocytogenes*			*S. enterica*		
Dataset size	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST
1	245.59	74.80	80.46	300.35	76.85	84.23	383.88	78.50	88.33
2	285.55	75.23	80.35	331.48	78.72	84.79	451.37	80.15	88.30
4	348.74	76.72	80.72	459.23	81.35	84.68	617.49	85.02	89.76
8	444.45	79.88	81.55	616.80	88.08	85.73	784.73	91.39	90.00
16	463.58	89.36	81.96	629.87	101.95	86.06	871.55	107.73	93.26
32	466.22	105.75	84.22	672.21	120.75	87.45	915.02	142.70	95.91
64	564.50	142.06	84.49	736.91	157.54	91.82	1029.75	197.62	96.56
128	649.79	163.51	86.14	794.65	209.10	93.04	1134.74	250.34	101.99
256	700.47	237.88	86.41	830.57	325.06	93.25	1290.67	422.44	104.91
512	789.30	380.45	86.38	888.29	567.43	94.21	1424.32	737.53	104.92
1024	865.61	691.73	86.74	974.11	1073.47	93.98	1609.24	1424.54	105.26
2048	970.07	1301.39	86.77	1084.84	2069.30	94.28	1948.72	2767.81	105.94
4096	1076.68	2517.72	87.01	1300.98	4000.32	94.52	2327.89	5376.79	105.42
8192	1237.61	4931.30	86.98	1575.37	7913.11	94.51	3108.47	10671.73	105.98
16384	1538.04	9792.28	87.39	2287.09	15749.59	94.20	4235.34	21251.72	106.15

Table 2.9: Mean runtime values in minutes for the allele calling of each species' datasets with chewBBACA 3's four execution modes.

	S. pyogenes				*L. monocytogenes*				*S. enterica*			
Dataset size	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	0.04	0.05	0.08	0.10	0.07	0.08	0.13	0.16	0.13	0.14	0.21	0.28
2	0.04	0.06	0.10	0.12	0.07	0.08	0.15	0.20	0.13	0.15	0.22	0.33
4	0.05	0.06	0.12	0.14	0.07	0.09	0.19	0.26	0.14	0.19	0.33	0.51
8	0.06	0.08	0.17	0.22	0.09	0.12	0.26	0.37	0.17	0.24	0.52	0.82
16	0.08	0.10	0.20	0.26	0.12	0.16	0.39	0.56	0.23	0.33	0.62	1.07
32	0.12	0.16	0.33	0.42	0.18	0.24	0.49	0.74	0.42	0.47	0.87	1.58
64	0.19	0.25	0.47	0.61	0.31	0.38	0.71	1.03	0.65	0.84	1.42	2.60
128	0.32	0.41	0.71	0.88	0.56	0.69	1.13	1.65	1.05	1.31	2.01	3.88
256	0.63	0.74	1.08	1.35	1.08	1.25	1.78	2.58	2.10	2.48	3.34	6.11
512	1.19	1.35	1.81	2.21	2.03	2.26	2.93	4.17	3.87	4.52	5.54	9.74
1024	2.33	2.63	3.21	3.79	4.18	4.60	5.45	7.22	7.67	8.62	9.99	16.43
2048	4.70	5.12	5.86	6.69	8.45	9.13	10.32	12.98	16.21	17.89	19.71	29.40
4096	9.84	10.52	11.55	12.77	18.41	19.70	21.39	25.44	36.97	39.53	41.94	56.25
8192	22.14	23.32	24.81	26.69	43.50	45.82	48.56	54.71	78.66	81.90	86.39	109.49
16384	54.21	56.25	58.64	61.20	98.39	101.03	105.28	115.84	172.90	175.64	182.55	220.60

Table 2.10: Mean peak memory usage values in megabytes (MB) for the allele calling of each species' datasets with chewBBACA 3's four execution modes.

	S. pyogenes				*L. monocytogenes*				*S. enterica*			
Dataset size	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4	Mode 1	Mode 2	Mode 3	Mode 4
1	207.17	236.03	243.21	245.41	241.32	278.33	291.99	300.58	334.90	356.94	376.05	373.74
2	247.25	274.51	285.88	285.58	310.95	316.36	357.78	335.50	428.30	431.12	436.29	448.07
4	326.66	347.68	348.49	352.61	452.29	452.32	466.07	459.63	617.17	614.39	614.29	635.11
8	422.72	441.80	444.70	449.01	608.20	611.59	623.93	627.72	786.50	782.67	778.62	786.11
16	445.85	464.16	467.19	467.81	621.31	620.10	625.42	632.11	871.49	867.79	871.27	864.63
32	468.18	467.03	466.80	466.77	672.06	671.52	672.82	671.89	912.87	918.69	912.17	914.32
64	564.73	564.28	562.36	565.82	736.79	738.39	737.35	737.71	1018.99	1013.50	1029.77	1027.37
128	649.72	648.26	655.04	651.98	794.47	795.35	794.25	795.16	1151.16	1178.55	1150.94	1158.56
256	706.43	704.90	700.72	706.41	831.13	831.47	835.84	837	1251.01	1277.62	1261.15	1281.28
512	787.97	801.64	792.56	789.06	882.62	882.87	889.19	888.90	1436.54	1392.310156	1383.53	1428.23
1024	866.24	866.88	868.71	866.87	969.94	982.68	969.36	976.03	1583.510156	1628.719531	1619.13	1620.02
2048	970.48	970.21	970.40	970.10	1085.65	1059.37	1071.66	1077.11	1969.12	1915.98	1950.73	1945.92
4096	1077.019531	1076.55	1076.57	1076.98	1252.169531	1274.26	1283.17	1316.99	2258.61	2163.53	2170.379688	2251.50
8192	1236.29	1244.03	1242.92	1246.09	1455.23	1473.02	1538.639844	1608.01	2962.67	3009.10	3110.929688	3100.90
16384	1537.65	1536.51	1538.26	1540.00	2110.769531	2096.86	2249.089844	2232.05	3958.66	3963.51	4222.44	4232.04

Table 2.11: Number of loci in each species' cgMLST schemas, number of core and accessory loci determined based on each tool's results and number of core and accessory loci determined based on chewBBACA 2's and pyMLST's results that are shared with the sets of core and accessory loci determined based on chewBBACA 3's results.

		core loci			accessory loci			core loci shared with chewBBACA 3		accessory loci shared with chewBBACA 3	
Species	Schema loci	chewBBACA 2	chewBBACA 3	pyMLST	chewBBACA 2	chewBBACA 3	pyMLST	chewBBACA 2	pyMLST	chewBBACA 2	pyMLST
S. pyogenes	1345	1322	1321	1075	23	24	270	1321	1073	23	22
L. monocytogenes	2449	2313	2314	1022	136	135	1427	2312	983	134	96
S. enterica	3192	3001	3009	2365	191	183	827	3000	2255	182	72

Table 2.12: Total number of coding sequences (CDSs) predicted by Pyrodigal for each species' complete dataset, total number of CDSs classified by each tool, percentage of the total CDSs classified by each tool and average number of CDSs classified per strain.

		Total CDSs classified			Percentage of total CDSs classified			Average number of CDSs classified per strain		
Species	Total number of CDSs (n=16384)	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST	chewBBACA 3	chewBBACA 2	pyMLST
S. pyogenes	28,451,923	21,982,083	21,980,537	19,092,545	77.26	77.26	67.10	1,341.68	1,341.59	1,165.32
L. monocytogenes	49,787,476	41,011,719	40,100,999	23,454,353	82.37	80.54	47.11	2,503.16	2,447.57	1,431.54
S. enterica	75,417,045	52,680,767	52,142,713	39,868,671	69.85	69.14	52.86	3,215.38	3,182.54	2,433.39

Table 2.13: Special classification counts for each species' complete dataset per tool.

	chewBBACA 3									chewBBACA 2					pyMLST	
Species	EXC	INF	PLOT3	PLOT5	LOTSC	NIPH	NIPHEM	ALM	ASM	PAMA	EXC	INF	PLOT	NIPH	ALM	ASM	EXC	NIPHEM
S. pyogenes	21735280	218736	221	1817	98	154	1757	1170	20841	0	21736576	218665	2370	504	1210	20994	19078994	31
L. monocytogenes	38543696	538769	78	5594	59	23357	918118	690	21541	0	38552279	538435	6349	932267	1149	21611	23342559	289
S. enterica	50639542	919227	308	5770	31	14416	523022	1309	37156	31	50663361	916359	6806	517896	1437	36854	39838200	1200

Chapter 3

Chewie Nomenclature Server (Chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas

This chapter is a reproduction of the following publication:

R. Mamede, P. Vila-Cerqueira, M. Silva, J. A. Carriço, M. Ramirez, Chewie Nomenclature Server (Chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas, Nucleic Acids Research, Volume 49, Database Issue, January 2021, D660-D666, DOI: `https://doi.org/10.1093/nar/gkaa889`

The adoption of wg/cgMLST has been largely driven by the availability of web platforms that provide powerful analytic capabilities for wg/cgMLST and centralize data analysis, ensuring the standardization of data analysis steps and the comparability of the results. These platforms store genomic and gene sequences, as well as associated metadata, and allow users to identify loci and alleles present in submitted strains based on comparisons against wg/cgMLST schemas with well-defined allelic nomenclatures. These centralized services greatly promote interoperability and provide a richer context by allowing users to compare their strains against strains submitted by other users from diverse geographic locations. However, by requiring users to submit their data and centralizing data analysis, the services provided by these platforms may not be adequate for users operating under strict data privacy policies and may offer limited scalability, especially as data availability increases and when a timely analysis is desirable, such as in an outbreak context.

This chapter presents Chewie-NS, a deployable wg/cgMLST platform for easy sharing of wg/cgMLST schemas that allows users to perform local and private analysis under a common allelic nomenclature. Chewie-NS was designed to store wg/cgMLST schemas for any bacterial species and to provide a simple user interface for users to intuitively explore the diversity of loci included in each schema. Schema and loci data can be easily browsed and downloaded through the website or API. Furthermore, integration with chewBBACA 3,

presented in **Chapter 2**, allows users to quickly download schemas and start performing local and private analyzes based on the common allelic nomenclature managed by Chewie-NS. By decentralizing data analysis, users operating under strict data privacy policies can still use Chewie-NS' services, and data processing is not limited by the resources available to the remote server.

As first co-author, I was involved in the implementation of the backend component and of the modules that allow for the integration with chewBBACA 3. My contribution to the Frontend component consisted of the creation of Python scripts that generate pre-computed data to be displayed in the tables and plots of the website. The development of Chewie-NS was in part guided by a proof of concept that had been previously implemented by members of the lab where the work was carried out. I focused mainly on defining and creating the endpoints of the API and on making sure that the API provided the functionalities necessary for the integration with chewBBACA and the Frontend component. I developed the API and the modules for integration with chewBBACA 3 simultaneously to ensure that they were fully compatible. In addition, I had to fine-tune and optimize several API endpoints, restructure the database used to store schema data and create custom Python scripts to guarantee that Chewie-NS provided the desired functionalities at scale.

Chewie Nomenclature Server (Chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas

Rafael Mamede[1,2], Pedro Vila-Cerqueira[1], Mickael Silva[1], João A. Carriço[1], Mário Ramirez[1],

[1] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal;

[2] Gulbenkian Institute for Molecular Medicine.

3.1 Abstract

Chewie-NS (`https://chewbbaca.online/`) allows users to share genome-based gene-by-gene typing schemas and to maintain a common nomenclature, simplifying the comparison of results. The combination between local analyses and a public repository of allelic data strikes a balance between potential confidentiality issues and the need to compare results. The possibility of deploying private instances of Chewie-NS facilitates the creation of nomenclature servers with a restricted user base to allow compliance with the strictest data policies. Chewie-NS allows users to easily share their own schemas and to explore publicly available schemas, including informative statistics on schemas and loci presented in interactive charts and tables. Users can retrieve all the information necessary to run a schema locally or all the alleles identified at a particular locus. The integration with the chewBBACA suite enables users to directly upload new schemas to Chewie-NS, download existing schemas and synchronize local and remote schemas from chewBBACA command line version, allowing an easier integration into high-throughput analysis pipelines. The same Representational State Transfer (REST) API linking Chewie-NS and the chewBBACA suite supports the interaction of other interfaces or pipelines with the databases available at Chewie-NS, facilitating the reusability of the stored data.

3.2 Introduction

The importance of distinguishing strains within the same microbial species has been proven critical for identifying chains of transmission and understanding pathogen evolution, as recently illustrated by the SARS-CoV-2 pandemic [181, 182]. The advent and widespread adoption of high-throughput sequencing allowed leveraging genomic information for this purpose [181, 182]. One of the most common approaches in bacterial typing is GbG methods, which extend the concept of MLST to include all genes present in the core genome of a given species (cgMLST) or, trying to cover a significant fraction of a species' pan-genome, in whole

genome (wgMLST) [125]. Current software approaches implementing these wg/cgMLST typing methods suffer from standardization issues when comparing results between different tools and between different laboratories or users [146].

We have previously developed a suite, chewBBACA [143], allowing the creation of GbG schemas and performing allele calls on assembled draft genomes. Since chewBBACA was designed to perform local analysis to address concerns over data privacy and scalability, it has the drawback that small adjustments in parameters may lead to inconsistencies between runs. Moreover, the software allows users to create their own wg/cgMLST schemas but currently no tool is available for the easy sharing of schemas, which potentially hampers long-term and multinational studies, as well as the reusability of already published schemas [183, 184].

There are well-established websites for performing GbG analyses, such as PubMLST (`https://pubmlst.org/`) [129] and EnteroBase (`https://enterobase.warwick.ac.uk/`) [151], that centralize analysis and hosting of public and private schemas. chewBBACA does not depend on a web server and by enabling local analyses and schema creation allows for scalable and private analyses of genomes, but the existing implementation lacked an easy way to share schemas and the associated allelic information, which is possible in a centralized solution.

In order to allow users to share GbG typing schemas and for a common allelic nomenclature to be maintained [185], we developed Chewie-NS, a nomenclature server based on the TypOn ontology [186] offering a web interface that also integrates directly with local instances of chewBBACA and can be programmatically accessed by external resources. Chewie-NS aims to complement the private local analysis of strains by also allowing the simple communication of results while providing an interface for users to easily explore the allelic diversity within species. The importance of the latter is becoming increasingly clear with the recognition that bacterial phenotypes can be profoundly altered by allelic variants [187, 188]. Other publicly available web services require submission of raw data, something that may raise privacy and ownership concerns, while our approach of enabling local analyses is more flexible and scalable, and respects data privacy concerns. Current wg/cgMLST typing methods suffer from standardization difficulties or issues that manifest not only when trying to reconcile results from different tools, but also when the same tool is run at different times with small adjustments in parameter values or database modifications that may lead to inconsistencies. This is an even more complex problem than it was for classical MLST methods [189]. Having a repository of schemas, their associated parameters and the allelic diversity identified will allow the consistent use of gene-by-gene typing schemas by different groups and to build upon the results of different studies to monitor microbial populations and study outbreaks.

Chewie-NS is available at `https://chewbbaca.online` and its source code is available at `https://github.com/B-UMMI/Chewie-NS`. Detailed documentation, including a descriptive tutorial on how to deploy and use the server, can be found at `https://chewie-ns.`

readthedocs.io. Additionally, a tutorial version of the server aiming at familiarizing users with the integration between the chewBBACA suite and Chewie-NS, which allows users to perform mock submissions of schemas and synchronizations without the need to register and with a much reduced database, is available at https://tutorial.chewbbaca.online/.

3.3 Database Creation

3.3.1 Backend

The architecture of Chewie-NS is shown schematically in 3.1. The backend component of Chewie-NS makes use of the Virtuoso triple store (v. 7.2.6) (https://virtuoso.openlinksw.com/). This database management system allows the integration of a Resource Description Framework (RDF) to implement the TypOn ontology [186] structure to store schema data. Additionally, a PostgreSQL database (v. 10) (https://www.postgresql.org/) was adopted for user management. These databases are accessible through a Python 3 REST API developed in the Flask (v. 1.1.0) (https://flask.palletsprojects.com/en/1.1.x/) web development microframework, which allows requests through defined endpoints and facilitates programmatic access to the nomenclature server. Requests and Hypertext Transfer Protocol Secure (HTTPS) connections are handled by a web server, NGINX (v. 1.17) (https://www.nginx.com/), that communicates with Gunicorn (v. 20.0.4) (https://gunicorn.org/), a Web Server Gateway Interface (WSGI) application server capable of running multiple processes of the web application and distributing incoming requests to ensure scalability and load balancing. A queueing system was implemented to manage all tasks with possible concurrent user access through Redis (v. 5.0.6) (https://redis.io/) and Celery (v. 4.4.0rc2) (https://docs.celeryproject.org/en/stable/getting-started/introduction.html).

3.3.2 Frontend

The UI for Chewie-NS was built with the JavaScript frameworks React (v. 16.12.0) (https://reactjs.org/) and Material-UI (v. 4.9.14) (https://material-ui.com/). The UI provides a list of available schemas and displays relevant schema and locus statistics in a responsive and interactive manner. Access to daily updated compressed files of the schemas for download and local use is also provided. All interactive charts were rendered with the graph visualization library Plotly.js (v. 1.52.1) (https://plotly.com/javascript/) through its React component, react-plotly (v.2.4.0) (https://plotly.com/javascript/react/).

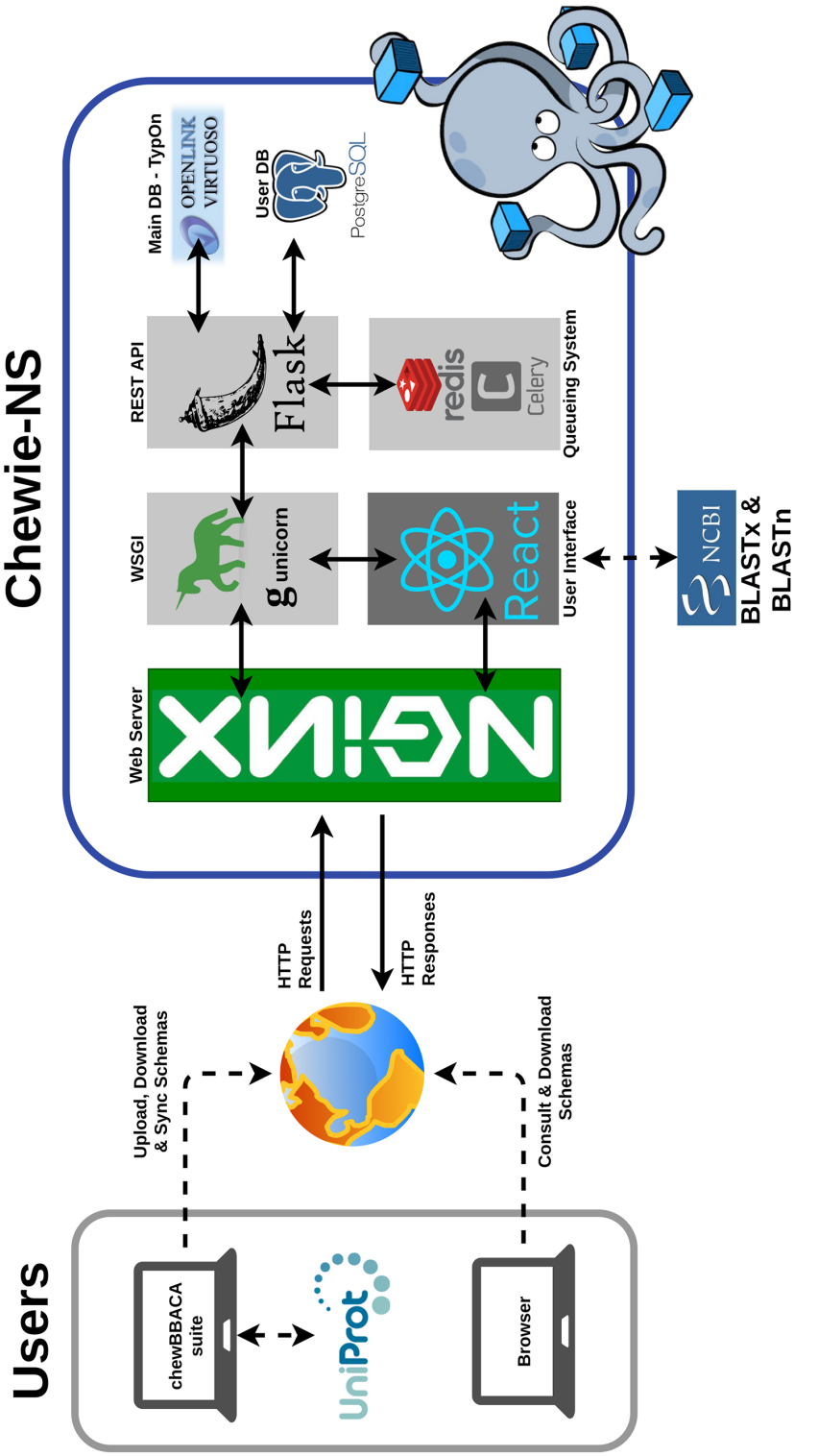


Figure 3.1: The Chewie-NS service: global overview of the technologies used and API connectivity.

3.3.3 Chewie-NS usage

Local installation. Deployment of local instances can be easily achieved through Docker Compose (`https://www.docker.com/`), available at `https://github.com/B-UMMI/Chewie-NS`. The use of a container orchestrator (`https://docs.docker.com/compose/`) supports the easy deployment of local instances independently of the hardware available, allowing the creation of private trusted databases if public access is not possible. Instructions on how to achieve this can be found at `https://github.com/B-UMMI/Chewie-NS`. This can be particularly important for national public health institutions in the context of restrictive or ambiguous data sharing laws because it allows stricter user access control.

Application programming interface. A RESTful API also referred to as a RESTful web service or REST API, i.e. based on representational state transfer (REST), is available. The user can interact with Chewie-NS's API through the web interface, by clicking on the 'API' button on the menu. This will open a page with Swagger UI (`https://swagger.io/tools/swagger-ui/`), a user-friendly tool for the user to interact directly with the REST API. Programmatic access is also possible through command line applications such as curl or tools such as Postman (`https://www.postman.com/`). Chewie-NS's REST API allows interaction with the PostgreSQL database to manage user registrations on local instances. Through the API, users are also able to query the Virtuoso database to download compressed schemas, search for specific alleles and query data about specific species, loci or alleles.

3.3.4 Web interface

Schemas overview. A table summarizes the species and number of schemas available for each species in Chewie-NS. Selecting a species leads to another table (3.2) with a list of relevant information about each available schema, namely the schema internal identifier, the user provided schema name, the username of the creator, the number of loci in the schema, the number of alleles, the software and version used to create the schema, the date of creation, the date of the last modification, the BLAST [190] score ratio selected, the translation table used, the minimum locus length and size threshold. In the table, there is a link to download the compressed file of the schema and the training file used to create it, both necessary to use the schema locally with the chewBBACA suite. Each table entry has also a link to a page containing more details about the schema. Below this table, an interactive bar chart displays the number of alleles per locus for each schema. The user can zoom in on the chart to obtain a better view of a given set of loci and can click on a bar to go to a page with more details on that particular locus.

Schema details. The schema evaluation and annotation page contains a description of the schema provided by the schema creator. During the schema upload, this information can be provided in a file using markdown, a simple plain-text-formatting syntax that allows the easy

Figure 3.2: The schemas overview page of Chewie-NS.

integration of hyperlinks, tables and images, allowing for a rich use of data for the description of the schema. Below this table are four charts in different tabs. Two charts (3.3) display characteristics of the schema: the distribution of the number of alleles per locus and of locus size. Two interactive charts represent for each locus its size summary statistics versus the number of alleles, and another a box plot of the size distribution of each locus. In all charts, the user can zoom in on particular regions for more detailed inspection and, on the latter two, clicking on the chart element opens a page with more information on that particular locus. Below the charts is a table of all the loci in the schema, including relevant information for each locus. This table, as all other tables of Chewie-NS, is searchable, facilitating finding loci with particular characteristics (3.4). Similarly to other tables, the table can also be exported in comma-separated values format.

Locus details. Already in the schema evaluation and annotation page is shown most of the information of each locus. This includes the internal locus identification and label, the automated annotation created by chewBBACA including a link to the relevant UniProt [188] page, a user locus name and user custom annotation (supporting markdown syntax), number of alleles and allele size information. The possibility of schema creators offering their own annotation allows for domain-specific information to be added to the schema, including potentially richer complementary data and links to relevant external resources. Two charts are offered, one summarizing the size distribution of the alleles (frequency of binned sizes) and the other representing the sizes of each allele. Direct links to perform Basic Local Alignment Search Tool (BLAST) searches using allele 1 of that locus (Nucleotide BLAST (BLASTn) and translated nucleotide BLAST (BLASTx)) are available at the bottom of the page, allowing the user to check for similarities that could offer further insights into the origin or likely

function of the protein potentially encoded by the locus. A multi-fasta file containing the alleles of the locus can be downloaded by the users from this page. On a different page, a simple query feature allows finding exact matches to specific sequences stored in any of Chewie-NS's databases, returning the loci and schemas where the sequence is found.

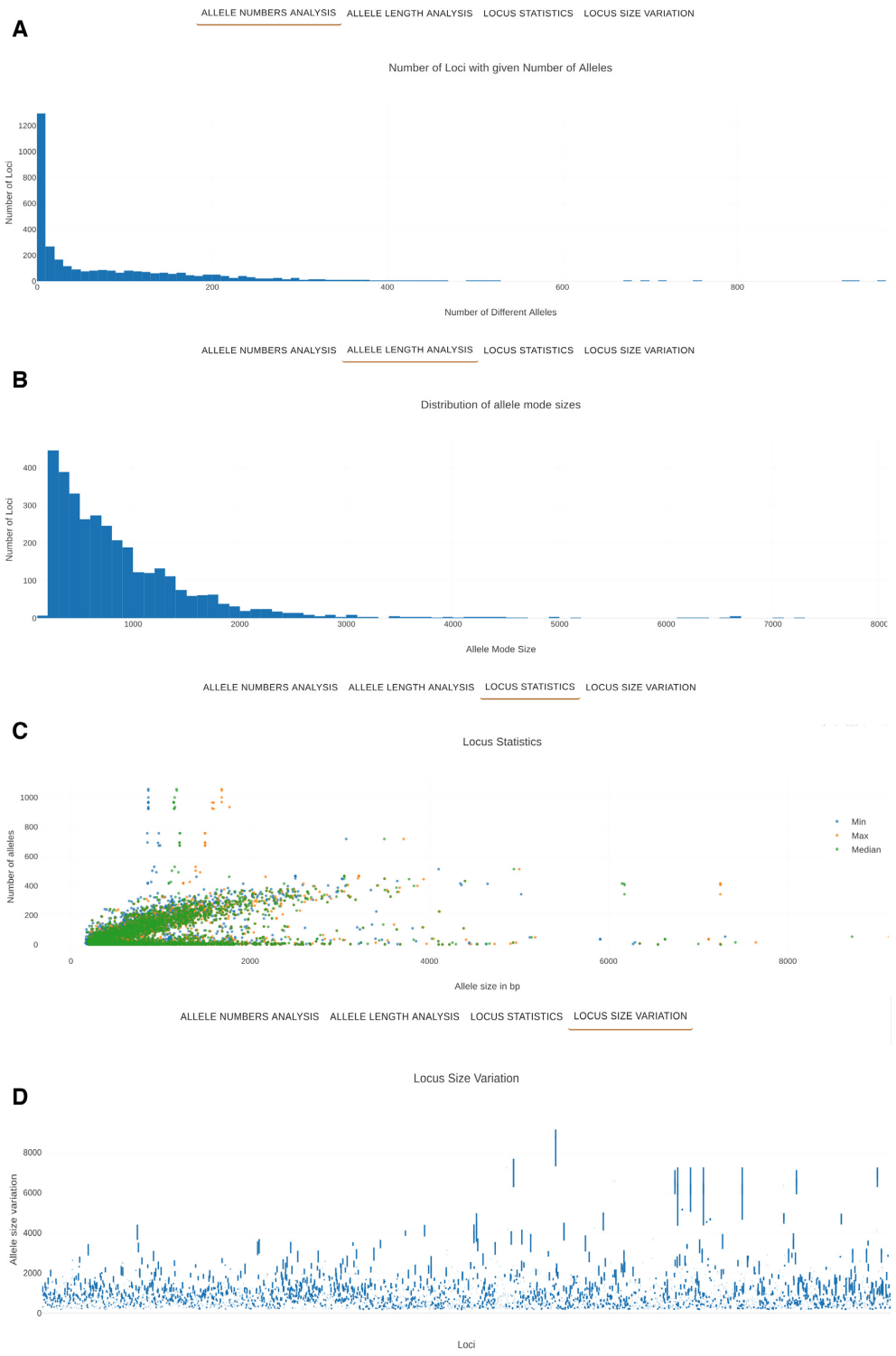


Figure 3.3: Summary charts displaying relevant information on a given schema. **(A)** Distribution of loci by number of alleles. **(B)** Distribution of loci by allele mode size. **(C)** Representation of summary statistics (minimum allele size in blue, maximum allele size in orange and median allele size in green) for each locus. **(D)** Box plots of loci size distribution; the loci in the x-axis are ordered by locus ID.

3. CHEWIE NOMENCLATURE SERVER (CHEWIE-NS): A DEPLOYABLE NOMENCLATURE SERVER FOR EASY SHARING OF CORE AND WHOLE GENOME MLST SCHEMAS

Integration with the chewBBACA suite and use of the API. By taking advantage of Chewie-NS's API, chewBBACA is capable of handling not only the schema creation, but also its upload, synchronization and download. Users of chewBBACA registered in Chewie-NS and with contributor privileges will be able to automatically upload a novel schema, making it available in Chewie-NS. Any authorized registered user can also contribute novel alleles identified in local analyses to Chewie-NS, contributing to the incremental development of the schema. This involves only allele information without the need to share a complete allelic profile with Chewie-NS. On the other hand, one does not have to be registered to download any of the data stored in Chewie-NS,including the compressed schemas or the novel alleles submitted to the Chewie-NS database and that were not present on the compressed file to update the local schema. Detailed instructions on the chewBBACA commands to achieve this can be found at `https://chewie-ns.readthedocs.io/en/latest/user/chewbbaca.html`.

If a user creates a schema through a software other than chewBBACA, the API can still be used to submit this new schema to Chewie-NS or to add novel alleles to non-chewBBACA schemas. A user would have to register with Chewie-NS and would have to take on the responsibility for making the correct API calls for schema submission. Furthermore, it would be up to each user submitting new alleles to guarantee the consistency of these novel alleles with the schema originally deposited. These functions are handled transparently by chewBBACA in its interaction with Chewie-NS. Anyone can use the API to download any of the schemas deposited in Chewie-NS to be run locally with chewBBACA or any other allele-calling algorithm. These multiple ways in which a user can interact with Chewie-NS allows tailoring the sharing of information to user preferences or to restrictions imposed on particular users.

In order to facilitate familiarization of the interaction between chewBBACA and Chewie-NS, a tutorial website was created (`https://tutorial.chewbbaca.online/`), together with step-by-step instructions on how to perform mock operations with a small size schema (`https://chewie-ns.readthedocs.io/en/latest/user/tutorial.html`). This allows users to perform submissions and synchronization of schemas without the need for registering. The schemas submitted to the tutorial site are not permanent and are removed automatically 48 h after creation.

For reproducibility and traceability purposes, a feature to retrieve database snapshots at specific dates is available, allowing a user to be able to recover the exact schema, as it was available on a given date. Full documentation of the schemas allows for traceability, which is critical in public health applications. These various options will continue to allow data privacy, while striving for a common nomenclature. The detailed parameterization associated with each schema created with chewBBACA and the consistency checks implemented in Chewie-NS mean that no human curation is necessary after the schema creation step, contributing to the rapid update of the database and exchange of information. However, although Chewie-NS can be used to store and retrieve information of schemas not created with chewBBACA, it

does not currently automatically guarantee the consistency of newly submitted alleles since each allele-calling algorithm will have specific parametrization requirements. Nevertheless, these can be implemented in the future as other allele-calling algorithms make use of the Chewie-NS platform.

Annotations

speB speI speJ speQ speR

Uniprot Label	Uniprot URI	User locus name	Custom Annotation	Locus ID	Locus Label	Total Number of Alleles	Alleles Mode	Size Range (bp)
Strepotococcal cysteine protease (Streptopain) / Streptococcal pyrogenic exotoxin B (SpeB)	http://purl.uniprot.org/uniparc/UPI000181F3F3	speB	Cysteine protease exotoxin SpeB (https://doi.org/10.1111/j.1365-2958.2011.07709.x; https://doi.org/10.1515/BC.2011.208)	86156	wgMLST-00086156	204	1197	966-1254
Exotoxin J	http://purl.uniprot.org/uniprot/Q7BAE3	speJ	Streptococcal pyrogenic exotoxin J (https://doi.org/10.1016/j.molmed.2013.10.004)	86338	wgMLST-00086338	56	699	636-711
N/A	N/A	speI	Streptococcal pyrogenic exotoxin I (https://doi.org/10.1016/j.molmed.2013.10.004)	87087	wgMLST-00087087	10	780	780-780
exotoxin	http://purl.uniprot.org/uniparc/UPI000DA4198B	speR	Streptococcal pyrogenic exotoxin SpeR (https://doi.org/10.1016/j.jinf.2019.02.005)	87813	wgMLST-00087813	8	675	675-675
Pyrogenic exotoxin K	http://purl.uniprot.org/uniparc/UPI0002D8ABCA	speQ	Streptococcal pyrogenic exotoxin SpeQ (https://doi.org/10.1016/j.jinf.2019.02.005)	88901	wgMLST-00088901	43	780	657-822

Figure 3.4: Schema loci table: search functionality. The figure presents the results of filtering for *speB*, *speI*, *speJ*, *speQ* and *speR* in the *user locus name* field of the *Streptococcus pyogenes* schema 1. Note that the user provided annotations complement and correct some of the annotations retrieved from UniProt.

3.4 Discussion

Chewie-NS accomplishes four important goals. First, it stores all the information required to define a chewBBACA schema, facilitating accessibility of schemas so that different schemas can be easily compared and evaluated. Second, it maintains a public compendium of the variability in each locus. Since chewBBACA loci are open-reading frames (ORFs), this will allow monitoring the variability of the proteins potentially encoded by these loci. This is important when studying microbial pathogens because small variations in sequence can lead to dramatic changes in virulence [191] or antimicrobial resistance [187]. On the other hand, allelic diversity can also be indicative of stabilizing or diversifying selective pressures, which in turn can be leveraged to obtain insights into pathogen evolution or interaction with the host [192]. Allelic diversity is also important in reverse vaccinology [193] and to monitor the continued potential efficacy of some available vaccines [194]. Third, through its integration with chewBBACA, it offers a simplified way for the user to control the flow of information between the local instance and Chewie-NS. This is important to keep the local instance of chewBBACA updated with the current common nomenclature stored in Chewie-NS databases and to contribute new alleles to the common databases, but it also allows for limited sharing of data to comply with any regulations the user may be

operating under. Finally, the proposed workflow hopes to stimulate and facilitate data sharing between users using the same schemas, allowing for a faster detection of strain similarity, therefore contributing for genomic epidemiology studies and also faster outbreak detection and investigation by expediting strain comparison between different laboratories or institutions. Although the current integration of chewBBACA with Chewie-NS facilitates the user interaction with Chewie-NS when using this allele-calling software, the API can be exploited by other allele-calling software to also interact directly with Chewie-NS, allowing an easier submission of schemas and of new alleles to schemas not created with chewBBACA.

The definition of a wg/cgMLST schema involves not only the choice of the target loci and of what constitutes a locus [for instance, an open-reading frame (ORF) as defined by Prodigal [159] in the case of chewBBACA, or a fragment of DNA between two primers in traditional MLST], but also the algorithm and parameters of the allele-calling software. If one uses the same set of target loci defined in the same way, but a different allele-calling software, one cannot guarantee that the alleles called for a given isolate would be the same as with another allele-calling software, a problem that can potentially become all the more acute with the identification of novel alleles. Even when using the same allele-calling software, if the parameters used are different, the alleles identified in a given isolate may also be different. Moreover, the addition of novel alleles to a schema that do not conform with the parameters defined initially may have hard to anticipate consequences on the subsequent allele-calling processes. Upstream of this, we would like to stress the importance of the use of shared assembly pipelines, to ensure that the deposited allele sequences are determined based on standardized procedures, as it has been shown that different assemblers can result in different variants in the assembly [195] and that this variability can introduce artificial allelic variability, even when using the same schema and allele-calling software.

The possibility of setting up local instances of Chewie-NS in a simplified way using Docker Compose facilitates creating private services that can cater to trusted groups of users and allow the implementation of Chewie-NS in institutions operating under strict privacy rules. In a public health context, this can also be used to deploy services allowing an easier communication between different agencies operating under distinct mandates.

The databases currently available in the public instance of Chewie-NS (`https://chewbbaca.online/`) include schemas developed within the INNUENDO project [184] for *Salmonella enterica*, *Campylobacter jejuni*, *Escherichia coli* and *Yersinia enterocolitica*, a schema developed for *Arcobacter butzleri* [183], an adaptation of a schema generated using the Ridom SeqSphere+ software for *Acinetobacter baumannii* [196] and in-house developed schemas for *Streptococcus agalactiae* and *Streptococcus pyogenes*. However, we expect that users of chewBBACA and of other allele-calling software will increasingly contribute schemas for these as well as additional species to be deposited in Chewie-NS.

3.5 Data Availability

Chewie-NS is freely accessible at `https://chewbbaca.online/`. Its source code is hosted at `https://github.com/B-UMMI/Chewie-NS` together with instructions on how to deploy it locally using Docker Compose and the documentation can be found at `https://chewie-ns.readthedocs.io/`. A tutorial version of the server, which allows users to perform mock submissions of schemas and synchronizations with a much reduced database, can be accessed at `https://tutorial.chewbbaca.online/`.

3.6 Acknowledgements

The authors would like to thank Catarina Inês Mendes for fruitful discussions and for her multiple design suggestions and Ana Correia for her guidance in solving issues with the Chewie-NS implementation.

3.7 Funding

Fundos Europeus Estruturais e de Investimento (FEEI) and Fundação para a Ciência e a Tecnologia (FCT) [LISBOA-01-0145-FEDER-016417]; FCT and FEDER [01/SAICT/2016 no. 022153]; FCT [PTDC/CCI-BIO/29676/2017]. Funding for open access charge: FEEI and FCT [LISBOA-01-0145-FEDER-016417].
Conflict of interest statement. None declared.

Chapter 4

Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of *Streptococcus pyogenes*

This chapter is a reproduction of the following publication:

A. Friães, R. Mamede, M. Ferreira, J. Melo-Cristino, M. Ramirez, Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of *Streptococcus pyogenes*, Journal of Clinical Microbiology, Volume 60, Issue 6, May 2022. DOI: `https://doi.org/10.1128/jcm.00315-22`

The supplementary material referred to throughout the text can be consulted in the last section of this chapter, **Section 4.7**.

Schemas are the centerpiece of wg/cgMLST. A wg/cgMLST schema captures the diversity of a set of loci, generally identified in a single bacterial species, and, through allele calling, allows the determination of the allelic profiles of strains of interest. Creating a schema for wg/cgMLST is not necessarily complex. Creating a well-curated wg/cgMLST schema that captures the diversity of thousands of loci of a bacterial species is a different matter altogether. Well-curated means that all loci or nearly all loci in the schema are *bona fide* loci, not spurious loci, such as frameshifted, truncated, or gene fusions, which are common issues plaguing wg/cgMLST schemas. Limiting the set of loci included in the schemas to the core loci to define a cgMLST schema is a way to minimize these problems. In theory, cgMLST schemas should include loci present in all or nearly all strains of a dataset representative of the diversity of a species, corresponding to loci with more conserved sequences due to their important functions. Although cgMLST schemas may provide robust results, in many cases they do not capture a considerable fraction of the diversity of bacterial species, limiting their applicability, such as for high-resolution typing of closely related strains in outbreak contexts. wgMLST schemas incorporate loci from the most variable fraction of the genome, usually termed the accessory genome, and provide greater resolution than cgMLST schemas. However, creating

a wgMLST schema is a laborious process, as incomplete genome assemblies, errors derived from genome misassembly, and the dynamic nature of the accessory genome can hinder identification of *bona fide* alleles and blur loci definitions.

This chapter describes the creation of a wgMLST schema for high-resolution typing of *Streptococcus pyogenes*. *S. pyogenes* is a major human pathogen that was included in the WHO BPPL released in 2024 due to reports of increased resistance to macrolides. Following the creation of a schema seed based on a set of high-quality complete genomes with an early version of chewBBACA 3, presented in **Chapter 2**, the schema was populated with alleles identified in several datasets representing the known diversity of *S. pyogenes*. The schema was then extensively annotated through automatic methods, followed by curation by the first first co-author, an expert in the biology of *S. pyogenes*, to validate the annotations and identify spurious loci. The schema was refined to remove or substitute spurious loci by valid alleles. The refinement resulted in a final wg/cgMLST schema with 3,044 loci, which captures far more of the diversity of *S. pyogenes* than the computed cgMLST schemas, which include only between 763 and 1321 loci. Using a wgMLST schema instead of hard-defined cgMLST schemas allows adjusting cgMLST analyses dynamically for each dataset, identifying more loci than conventional cgMLST and achieving performance comparable to SNP-based methods.

I participated in all steps related to schema development and dataset analyses to evaluate schema performance. Throughout the schema development process, I was responsible for selecting the genomes used for schema creation, annotating the schema loci based on multiple sources, and refining the schema based on the suggestions of the first first co-author. When necessary, I implemented custom scripts to annotate, refine, and evaluate schema performance.

Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of *Streptococcus pyogenes*

Ana Friães[1,*], Rafael Mamede[1,*], Mariana Ferreira[1], José Melo-Cristino[1], Mário Ramirez[1],

[1]Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal

[*]Contributed equally

4.1 Abstract

Streptococcus pyogenes is a major human pathogen with high genetic diversity, largely created by recombination and horizontal gene transfer, making it difficult to use SNP-based genome-wide analyses for surveillance. Using a Gene-by-Gene approach on 208 complete genomes of *S. pyogenes*, a novel wgMLST schema was developed, comprising 3,044 target loci. The schema was used for cgMLST analyses of previously published data sets and 265 newly sequenced draft genomes with other molecular and phenotypic typing data. Clustering based on cgMLST data supported the genetic heterogeneity of many *emm* types and correlated poorly with PFGE macrorestriction profiling, superantigen gene profiling, and MLST sequence type, highlighting the limitations of older typing methods. While 763 loci were present in all isolates of a data set representative of *S. pyogenes* genetic diversity, the proposed schema allows scalable cgMLST analysis, which can include more loci for an increased resolution when typing closely related isolates. The cgMLST and PopPUNK clusters were broadly consistent in this diverse population. The cgMLST analyses presented results comparable to those of SNP-based methods in the identification of two recently emerged sublineages of *emm1* and *emm89* and the clarification of the genetic relatedness among isolates recovered in outbreak contexts. The schema was thoroughly annotated and made publicly available on the Chewie-NS online platform (`https://chewbbaca.online/species/1/schemas/1`), providing a framework for high-resolution typing and analyzing the genetic variability of loci of particular biological interest.

4.2 Introduction

Streptococcus pyogenes (Lancefield group A *Streptococcus* (GAS)) remains a significant cause of global morbidity and ranks among the top 10 infectious causes of death [197]. In 2018, the WHO highlighted the importance of developing a GAS vaccine and set out priority

activities to reach this goal, including a better characterization of the epidemiology of GAS infections and the identification of appropriate candidate antigens [198].

In recent decades, sequence-based typing of the hypervariable region of the *emm* gene, encoding the M protein, was the most frequently used method to identify GAS lineages [199]. However, complementary methods have long been used, which, together with *emm* typing, allow finer discrimination of the circulating strains, including serotyping of the major backbone pilus protein (T antigen), PFGE macrorestriction profiling, MLST, and profiling of superantigen (SAg)-coding genes [200–202].

WGS analysis allowed the identification of emerging intra-*emm* clones with increased fitness or virulence that were otherwise indistinguishable by other typing methods. Such is the case of *emm*89 clade 3, which emerged during the 2000s and quickly outcompeted other emm89 lineages [203–205]. Isolates from this lineage lack the genes encoding hyaluronic acid capsule biosynthesis and carry a high-expression promoter in the operon encoding streptolysin O and NAD-glycohydrolase (Pnga-3) [204, 206]. More recently, an *emm*1 lineage ($M1_{UK}$), differing from the contemporary globally disseminated *emm*1 lineage ($M1_{global}$) by 27 SNPs, was identified in the United Kingdom [168] and subsequently reported in The Netherlands, the United States, and Canada [207–209].

WGS has been decisive in clarifying the molecular and evolutionary mechanisms underlying the success of long-term-circulating lineages [210, 211] and has proven useful in the identification of outbreak-related cases [212, 213]. Genomic data have the additional potential benefit of providing information on the variability of candidate vaccine antigens and genes involved in antimicrobial resistance [214, 215], further supporting the use of HTS in GAS surveillance.

Most genome-wide analyses performed on GAS have been based on the comparison of SNPs between isolates. This usually involves mapping short-read sequence data or aligning de novo-assembled sequences to a selected reference genome [168, 204–206, 210–214]. However, the choice of an appropriate reference is challenging when simultaneously comparing diverse lineages [125, 216], such as in population-based studies of *S. pyogenes* infection isolates. SNP-based phylogenetic analysis also requires the removal of regions of recombination, which are an important source of diversity in GAS [125, 214, 216, 217]. These limitations can be largely overcome by the use of GbG approaches like wgMLST or cgMLST [218], which do not require comparison to a reference genome and which intrinsically dampen the effect of recombination [125, 216, 219]. MST-like downstream analyses further facilitate the use of wg/cgMLST. Additionally, wg/cgMLST schemas can be curated and maintained in centralized databases, providing a standardized nomenclature and ensuring reproducibility and comparison of results across laboratories [125, 196, 218, 220]. Indeed, similarly to SNP-based approaches, cgMLST schemas have been successfully used for both outbreak identification and population-based surveillance of multiple pathogens [196, 216, 219–223]. However, it is important to remember that wg/cgMLST is not designed to interrogate noncoding regions

of the genome and therefore would have been unable to detect the polymorphisms in the *nga-ifs-slo* promoter present now in the $M1_{global}$ lineage [168].

The aims of this study were to define a publicly available annotated wgMLST schema for *S. pyogenes* and evaluate its suitability for high-resolution typing and documenting the variability of loci encoding proteins of biological relevance.

4.3 Materials and Methods

Bacterial strains and data sets. A collection of 265 nonduplicate GAS strains isolated from pharyngitis, skin and soft tissue infections, and normally sterile sites in Portugal between 2001 and 2009 was selected for HTS and comparison of cgMLST with other typing methods (see supplemental Data Set 1 in reference [224]). These isolates were previously characterized regarding *emm* type, T type, PFGE profile, SAg gene profile, and antimicrobial resistance [201, 225–228] and represent four *emm* types: *emm*1, *emm*3, *emm*4 (including erythromycin-resistant and -susceptible isolates), and *emm*89 (including isolates carrying P*nga*-1, P*nga*-2, and P*nga*-3) [203].

In order to evaluate the performance of the proposed wgMLST schema in more diverse collections, outbreak recognition, and the identification of recently emerged intra-*emm* lineages of interest, publicly available data sets from three previous publications were also included [168, 213, 214]. Data Set 2 comprises 2,006 assemblies from a collection of isolates previously selected to represent the genetic, geographic, temporal, and clinical diversity of GAS [214]. Data Set 3 consists of 119 isolates associated with 21 outbreaks recorded in England from 2010 to 2015 and 170 contemporaneous sporadic isolates with the same *emm* types [213]. Data Set 4 comprises 135 assemblies from noninvasive *emm*1 isolates recovered in the United Kingdom from 2009 to 2016 [168] and the MGAS5005 complete genome that was used as a reference. The United Kingdom assemblies include 123 isolates carrying 27 SNPs characteristic of the recently emerged $M1_{UK}$ lineage and 5 intermediate isolates carrying 13 or 23 of those SNPs [168]. Data Set 5 includes all the *emm*89 assemblies included in Data Sets 1 to 4 (n = 194) and the 7 complete genomes of *emm*89 that were used to create the schema.

The majority of the strains included in Data Sets 2 to 4 [224] were retrieved from collections of publicly available genome assemblies [175, 229]. For strains for which it was not possible to retrieve a public genome assembly, the raw sequencing data were downloaded from the ENA and subsequently assembled. All assemblies were filtered according to assembly quality, *emm* type, and multilocus ST criteria, as detailed below.

High-throughput sequencing. Genomic DNA was extracted from cultures of GAS grown overnight in Todd-Hewitt broth (Oxoid, Basingstoke, UK) using the PureLink genomic DNA

minikit (Invitrogen, Carlsbad, CA, USA). The initial bacterial lysis step was carried out in the presence of 45 U of mutanolysin (Sigma-Aldrich, St. Louis, MO, USA) and 86 μg of hyaluronidase (Sigma-Aldrich, St. Louis, MO, USA). WGS libraries were generated using the Nextera DNA library preparation kit (Illumina, San Diego, CA, USA). The libraries were sequenced in an Illumina MiSeq or NextSeq instrument.

Sequencing data analysis. Raw sequence reads were assembled with INNUca v4.2.2 [230], with the following parameters: *-s Streptococcus pyogenes*, *-g 2*, *–estimatedMinimumCoverage 10*, *–trueCoverageProceed*, and *–fastQCproceed*. Samples that failed any of the quality control steps related to sequence quality or assembly coverage were excluded from the data sets. Assemblies are available as supplemental material [224].

In silico ST prediction was performed using MLST v2.19.0 [176] with default parameters and the PubMLST database updated on 11 March 2021. Genome assemblies with partial matches to any of the MLST genes or for which it was not possible to identify at least one of the MLST genes were excluded, except for ST293, ST403, ST404, and ST688, which lack the *yqiL* gene, and ST1087, which lacks the *xpt* gene. Strains with a predicted ST that was inconsistent with the classification reported in the original study were also excluded.

The *emm* type was determined using emmTyper v0.2.0 [231] with verbose mode and the CDC M-type-specific sequence databases updated on 11 March 2021. Genome assemblies without an identified *emm* type, with matches only for alleles flagged in the CDC database as possible *emm*-like genes, or with a predicted *emm* type that was inconsistent with the classification reported in the original study were excluded from the data sets. Assemblies classified with multiple *emm* types were also excluded (multiple subtypes of the same *emm* type were accepted), except for *emm*34/*emm*230 (*emm*34 corresponds to the *enn* gene, and the *emm* type is 230), *emm*13L/*emm*13 (these two types correspond to the same sequence), and other cases that were inspected in Geneious v8.1.9 to validate matches to the *emm* gene.

Variant calling to determine the set of SNPs in each assembled genome from Data Set 4 [224] was performed with Snippy v4.6.0 [232] with default parameters and the complete genome of strain MGAS5005 (RefSeq accession no. GCF_000011765.3) as the reference strain.

The *Pnga* variant was determined with SeqTyper v2.3 [233] with default parameters. A Fasta file with the sequences for all variants was given as the input to the blast module, followed by variant calling with the assembly module.

Schema creation, annotation, and curation. The complete genomes available in the NCBI RefSeq database as of 20 July 2020 were downloaded to select a set of 208 genome assemblies (see Table S1 in the supplemental material [224]) for schema creation with chewBBACA v2.7.0 [143]. Assemblies with a status of suppressed in the NCBI database were excluded, except for accession no. GCF_001535505.1, GCF_001547815.1, GCF_000013525.1, and GCF_900636425.1, whose status was changed to suppressed after the schema creation

and allele calling processes. Loci originating from these four genomes were inspected to ensure their validity. This initial schema seed, composed of 3,318 distinct loci, was populated through the inclusion of allelic variants from all assemblies included in the data sets and sourced from public databases [175, 229]. For schema annotation, the chewBBACA *UniprotFinder* process and custom scripts [234] were used to create a file with locus coordinates and annotation terms selected from the UniProt database, prioritizing the selection of terms from Swiss-Prot over terms from TrEMBL, and from matches against the translated coding sequences in the GenBank files of the genomes used for schema creation. Some product and gene names were further complemented with relevant literature references. The annotated schema was thoroughly curated to identify and remove spurious loci such as gene fusions, truncated genes, and paralogous loci. These loci were identified based on the retrieved annotations, the inspection of the genomic context, and the list of paralogous loci reported by the chewBBACA *AlleleCall* process and a custom script evaluating interlocus similarity [234]. Due to the minimum sequence length parameter enforced during schema creation, the *sagA* gene, present in the streptolysin S-encoding operon, was not in the initial schema. Given the importance of this gene for GAS pathogenesis and the potential interest in its variability, a locus was added representing the *sagA* gene. The full list of changes applied to the schema is available in Table S2 [224].

The schema was uploaded to Chewie-NS [152], where a more detailed description of schema creation, annotation, and curation can be found (`https://chewbbaca.online/species/1/schemas/1`).

cgMLST analysis. Allelic profiles of the core loci (shared by 100% of the isolates under analysis [cgMLST-100]) were used to create MSTs with the goeBURST algorithm in the desktop or online version of PHYLOViZ [235, 236]. Groups of isolates linked by up to n different loci in the MST were determined using the desktop version. The genes present in 95% (cgMLST-95), 99% (cgMLST-99), and 100% (cgMLST-100) of the isolates were identified with the chewBBACA *ExtractCgMLST* process for Data Set 2. The lists of genes for each gene presence threshold are available as supplemental material [224]. Intracluster and intercluster pairwise distances were determined using custom scripts [234].

PFGE cluster definition. Previously generated SmaI/Cfr9I macrorestriction PFGE patterns [175, 201, 225, 228] were used to create a unweighted pair group method with arithmetic means (UPGMA) dendrogram with BioNumerics software (Applied Maths, Sint-Martens-Latem, Belgium). The Dice similarity coefficient was used, with optimization and position tolerance settings of 1.0 and 1.5, respectively. PFGE clusters were defined based on $\geq$ 80% relatedness on the dendrogram [200].

Statistical analysis. The results of cgMLST-100 and other typing methods were compared using Simpson's index of diversity (SID), the adjusted Wallace (AW) coefficient, and the adjusted Rand (AR) coefficient [200, 237], calculated with an online tool (`http://www.comparingpartitions.info/`). For comparison with other typing methods, groups were

defined by cutting MSTs at a suitable allelic difference to have an SID similar to that of the method to which it was being compared.

Data availability. The annotated wgMLST schema and a detailed description of its development are publicly available in Chewie-NS [152] at `https://chewbbaca.online/species/1/schemas/1`. The genome assemblies and allele calling results for each data set, a static version of the wgMLST schema, the list of loci in each subschema, the pairwise distances computed for Data Sets 2 and 3, and Tables S1, S2, S5, and S9 can be found in the supplemental material [224]. Raw sequencing data and sample metadata for the 265 isolates included in Data Set 1 have been deposited in the ENA under project accession number PRJEB49967. The custom scripts used for schema annotation, curation, and result analyses are part of the *Schema Refinery* repository [234].

Table 4.1: Simpson's index of diversity and 95% confidence intervals for the typing methods used to characterize 265 *S. pyogenes* isolates recovered in Portugal.

Typing method	No. of partitions	SID ($CI_{95\%}$)[c]
emm type	4	0.742 (0.727-0.756)
ST	15	0.826 (0.800-0.852)
T-type[a]	6	0.744 (0.720-0.768)
SAg profile	19	0.835 (0.813-0.857)
PFGE	16	0.792 (0.766-0.817)
MST_{1000}[b]	5	0.743 (0.728-0.758)
MST_{45}[b]	15	0.807 (0.779-0.835)
cgMLST-100	245	0.999 (0.998-1.000)

[a] The SID for T type was calculated for the subset of 248 isolates with a defined T type (17 isolates were nontypeable).

[b] Groups of isolates linked by up to *n* different loci in the MST (MST_n).

[c] SID, Simpson's index of diversity; $CI_{95\%}$, 95% confidence interval.

4.4 Results

Development of the wgMLST schema for *S. pyogenes*. The final annotated wgMLST schema comprises 3,044 loci with 371,549 alleles. Out of these, 1,096 (36%) loci presented low variability, presenting 1 to 19 DNA alleles (see Fig. 4.7 in the supplemental material).

These correspond essentially to genes that were identified in a minority of assemblies, mostly associated with prophages and other mobile genetic elements. The exception is *sagA*, encoding the streptolysin S precursor peptide, which presented only 13 alleles despite being ubiquitous among *S. pyogenes* isolates. The short length of this locus (162 bp) may be partly responsible for the limited number of alleles.

On the opposite extreme, among the 10 most variable loci (>750 alleles) are genes encoding well-known surface-exposed virulence factors but also transcriptional regulators known to play a major role in GAS pathogenesis and virulence, namely, CovS, RopB, and Mga. For these loci, the diversity of DNA alleles also results in a very large number of protein variants (range, 535 to 1,695) (Fig. 4.8).

Specific analyses can be performed through the identification and creation of subschemas for smaller sets of biologically relevant loci, such as genes encoding virulence factors and transcriptional regulators, for which subschemas are provided as supplemental material [224].

Comparison of cgMLST with other typing methods. To compare cgMLST analysis with conventional typing methods, a collection of 265 infection isolates with previous information on *emm* type, ST, T type, PFGE profile, PCR profile of 11 SAg genes, and antimicrobial resistance was used (see Data Set 1 in [224]). This data set includes isolates of *emm* types 1, 3, 4, and 89; 15 distinct STs; 6 T types (17 isolates were nontypeable); 19 SAg profiles; and 16 PFGE clusters (Table 4.1).

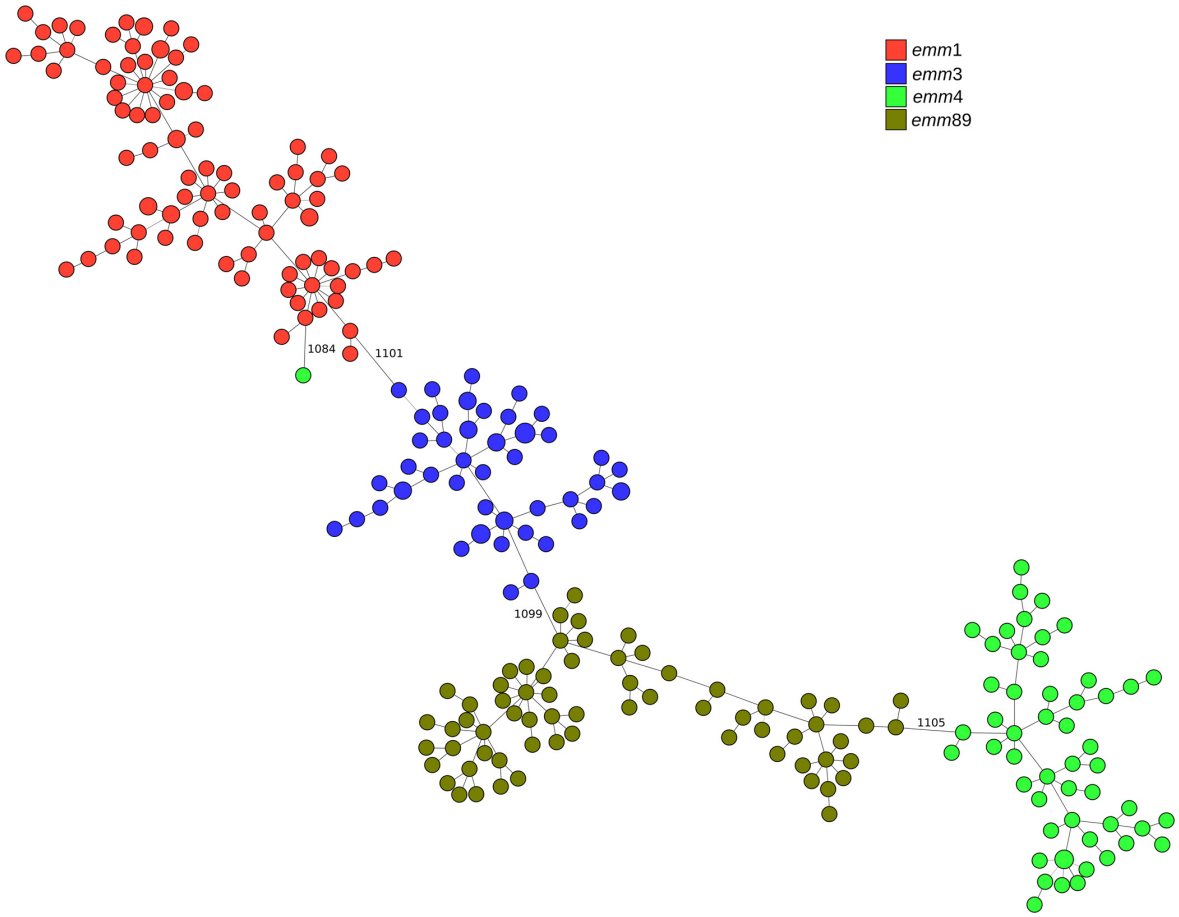


Figure 4.1: Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST-100 profiles of 265 *S. pyogenes* isolates recovered in Portugal (see Data Set 1 in [224]). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to *emm* type. Link distances of $\geq 1,000$ allelic differences are labeled (from a total of 1,230 compared loci).

Allele calling using the wgMLST schema followed by cgMLST-100 analysis generated 245 different profiles representing 1,230 loci. The resulting MST separated the isolates according to *emm* type, except for one *emm*4 isolate that did not cluster with the others (Fig. 4.1). The minimum distance between clusters of different *emm* types varied between 1,084 and 1,105 allelic differences, while those among isolates of the same *emm* type were ≤ 28 for *emm*1, ≤ 100 for *emm*3, ≤ 157 for *emm*4 (excluding the distantly related isolate), and ≤ 225 for *emm*89. Clustering of isolates at a cutoff of 1,000 differences created four groups separating the four *emm* types and one singleton, corresponding to the *emm*4 isolate, resulting in high concordance between the MST groups linked by up to 1,000 different loci and *emm* types (Table 4.1; see also Tables 4.3 and 4.4 in the supplemental material).

A lower congruence was obtained between the distributions of isolates in the MST and the remaining typing methods (ST, PFGE, SAg profiling, and T type) (Fig. 4.9 to 4.12). The AW and AR values between T types and MST groups linked by up to 1,000 different loci were only slightly lower than those of *emm* types (Tables 4.3 and 4.4), but T type had a lower typeability since 17 isolates were nontypeable. Although MST groups linked by up to 45 different loci resulted in a number of partitions and SIDs comparable to those of ST, SAg

profiling, and PFGE (Table 4.1), the AW coefficient between MST groups linked by up to 45 different loci and these typing methods was lower than that between MST groups linked by up to 1,000 different loci and *emm* type (Table 4.4). This means that MST groups linked by up to 45 different loci could not confidently predict the ST, PFGE cluster, or SAg profile, or the converse, which was also reflected in lower AR values (<0.900) (Table 4.3).

The use of a wgMLST schema instead of a universally defined cgMLST-100 set of loci allows scalable analysis in which higher resolution can be obtained by including larger numbers of common loci when analyzing closely related isolates. As an example, the cgMLST-100 obtained exclusively for the *emm*4 isolates grouped into the same MST group linked by up to 1,000 different loci (n = 54) comprises 52 profiles of 1,382 cgMLST-100 loci. The *emm*4 isolates presenting the M phenotype of macrolide resistance (erythromycin resistant and clindamycin susceptible) shared ST39 and an SAg profile with most susceptible isolates (see Data Set 1 in [224]), rendering these two methods unable to differentiate macrolide-resistant isolates. One PFGE cluster was associated with macrolide resistance [238], although it also included two susceptible isolates. Similarly, one of the MST groups linked by up to 33 different loci comprised exclusively all but two of the macrolide-resistant isolates (Fig. 4.13). Not surprisingly, the set of 46 loci that were present universally and exclusively in the subset of erythromycin-resistant isolates (list available in the supplemental material in reference [224]) represents mostly phage-related genes, including *mef*(A) and *msr*(D), the genes most commonly associated with the M phenotype in GAS [239, 240].

Performance of the wgMLST schema on a large and genetically diverse data set. The genetic structure of the GAS population is known to vary temporally and geographically, with an associated impact on the disease spectrum and incidence [241, 242]. To evaluate the performance of the proposed wgMLST schema on the analysis of genetically diverse data sets, we used a large collection of isolates previously selected to represent the genetic, geographic, temporal, and clinical diversity of GAS [214]. A total of 2,006 assemblies were included in the data set, comprising 140 *emm* types and 443 STs and organized into 292 phylogroups defined by POPulation Partitioning Using Nucleotide Kmers (PopPUNK) [142, 214] (see Data Set 2 in [224]).

We defined 1,321-locus cgMLST-95, 1,204-locus cgMLST-99, and 763-locus cgMLST-100 schemas (available in the supplemental material in [224]).

Allele call results identified 1,700 cgMLST-100 profiles. The resulting MST indicates that many *emm* types include diverse genetic lineages, with 12 of the 19 most prevalent *emm* types (>30 isolates) comprising isolates distributed in multiple tree regions (Fig. 4.2). Accordingly, 50 of the 67 *emm* types comprising $\geq$ 10 isolates included assemblies that differed in >50% of the 763 cgMLST-100 loci (up to 708 differences [93%] in *emm*4) (Fig. 4.3A). In 31 of these *emm* types, the mean intra-*emm* allelic difference was larger than the smaller difference from another *emm* type (Table 4.11). This is possibly due to the diversity of geographic and temporal origins of the isolates in this data set and is in line with a previous report of genetic

heterogeneity within *emm* types [214]. It is also reflected in a low congruence between *emm* types and MST groups linked by up to 450 different loci despite similar SID values (Tables 4.5 to 4.7).

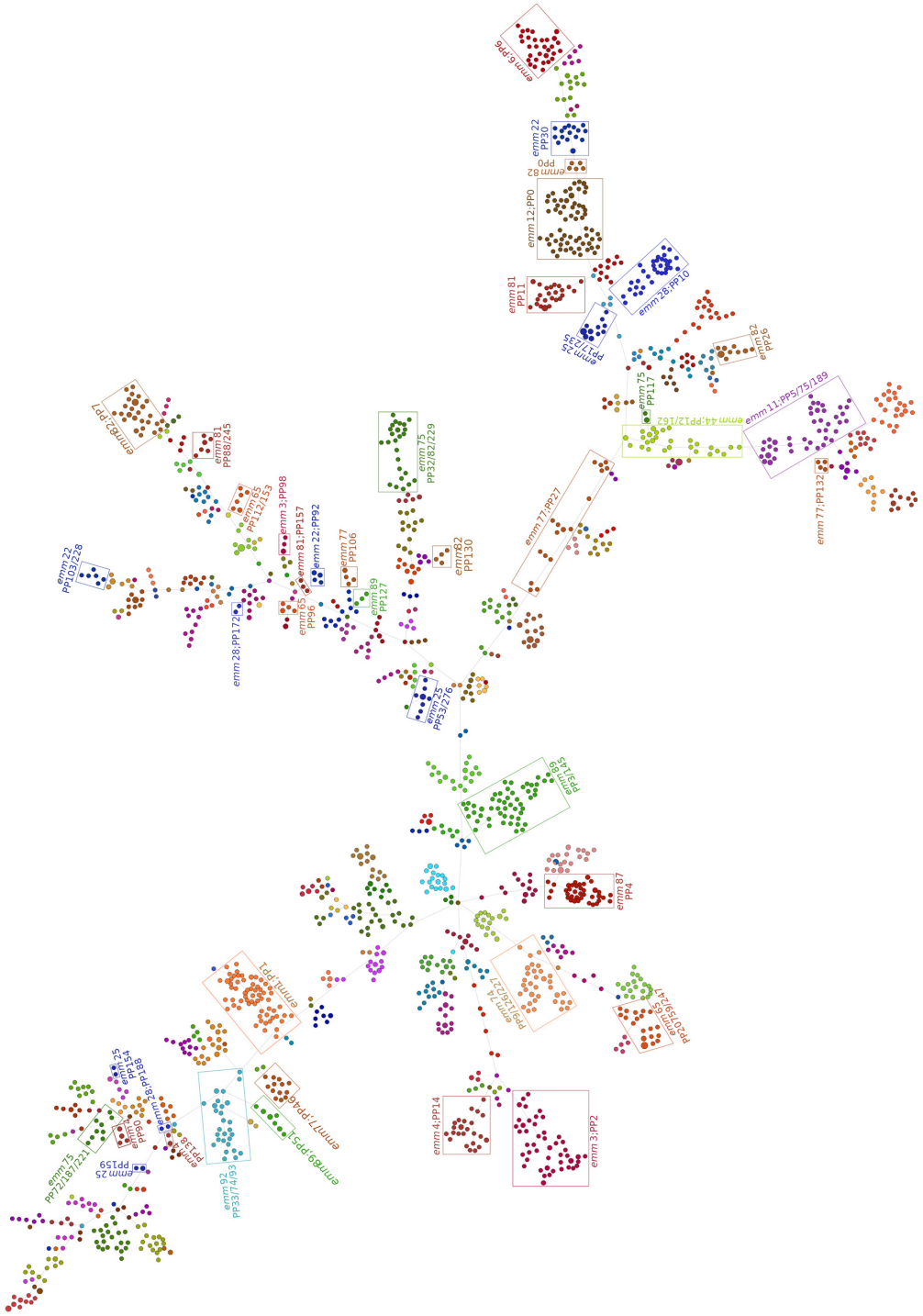


Figure 4.2: Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 2,006 genetically diverse *S. pyogenes* isolates recovered worldwide [214] (see Data Set 2 in reference [224]). The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Nodes are colored according to *emm* type. Groups of clustered *emm* types represented by >30 isolates are highlighted inside rectangles and labeled with the respective *emm* types and PopPUNK phylogroup numbers (for simplicity, isolated nodes of *emm* types 4, 22, 44, 65, 75, 77, 81, and 92 are not highlighted). A total of 763 core loci were compared.

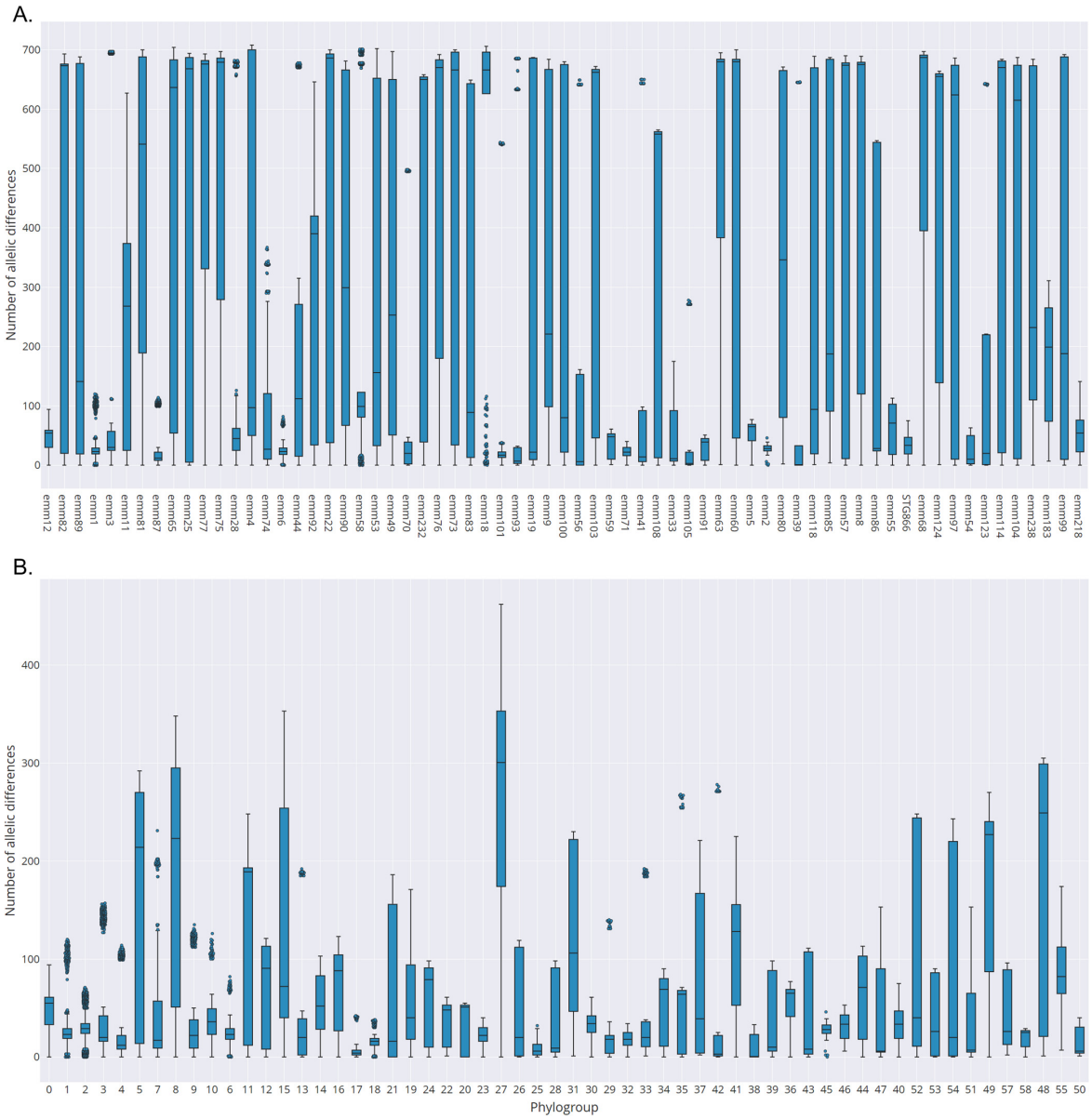


Figure 4.3: Box-and-whisker plots for the pairwise distances of the assemblies from Data Set 2 [214, 224] included in each *emm* type with $\geq$ 10 isolates (A) or in each PopPUNK phylogroup with $\geq$ 10 isolates (B). The distances were calculated based on the allele call results for the 763 cgMLST-100 loci of the 2,006 assemblies (interactive versions of these plots are available as supplemental material in reference [224]).

The overall congruence between STs and MST groups linked by up to 50 different loci was poor although slightly higher than that observed for the less diverse Data Set 1 (AR coefficients of 0.810 and 0.709, respectively) (Table 4.6). In contrast, there was good congruence, with high AW and AR values, between PopPUNK phylogroups and MST groups linked by up to 200 different loci (Tables 4.6 and 4.7). Still, PopPUNK phylogroups can be rather diverse, including multiple STs and isolates differing in up to 61% of the core 763 loci (phylogroup 27) (Fig. 4.3B; Table S9 [224]), highlighting the advantage of using multiple methods for analyzing the evolution of GAS lineages.

Performance of the wgMLST schema in an outbreak context. To evaluate the potential contribution of the proposed wgMLST schema for outbreak recognition, we used a previously

published data set comprising isolates from 21 outbreaks in England and contemporaneous nonrelated isolates with the same *emm* types [213]. A total of 119 outbreak isolates and 170 sporadic isolates were included (see Data Set 3 in [224]). Allele calling for the 119 outbreak isolates identified 58 profiles of 1,263 cgMLST-100 loci. In agreement with the SNP-based clustering presented previously [213], the MST clustered the isolates according to *emm* type, with a minimum distance of 1,079 allelic differences between isolates of different *emm* types, while isolates of different subtypes or outbreaks of the same *emm* type were more closely related (Fig. 4.4). Individual MSTs were created for *emm* types 1, 5, 11, 28, 75, 89, and 94, including outbreak and sporadic isolates (Fig. 4.14 to 4.20). Since these MSTs included only isolates sharing the same *emm* type, they comprised larger sets of cgMLST-100 loci (1,384 to 1,547 loci), potentially allowing higher resolution in the discrimination of outbreak isolates. Ten isolates with epidemiological links could be excluded from the respective outbreaks because they did not cluster with isolates of the same outbreak or differed by too many loci (Table 4.8 and Fig. 4.14 and 4.17 to 4.19). These isolates also matched the outbreak exclusion criteria based on SNP analysis [213]. Except for these 10 excluded isolates, outbreak isolates linked in the MSTs shared >99.5% of their core genome (maximum link distance of 6 allelic differences), and the mean distance within a given outbreak was much lower than the mean distance among sporadic isolates of the same *emm* type (Table 4.2). However, in *emm* types 1 and 5, there were sporadic isolates with cgMLST profiles very similar to those of outbreak 1 (OB1) and OB19, respectively (0 to 2 allelic differences), indicating that these outbreak strains were also present in the community (Table 4.2; Fig. 4.14 and 4.15).

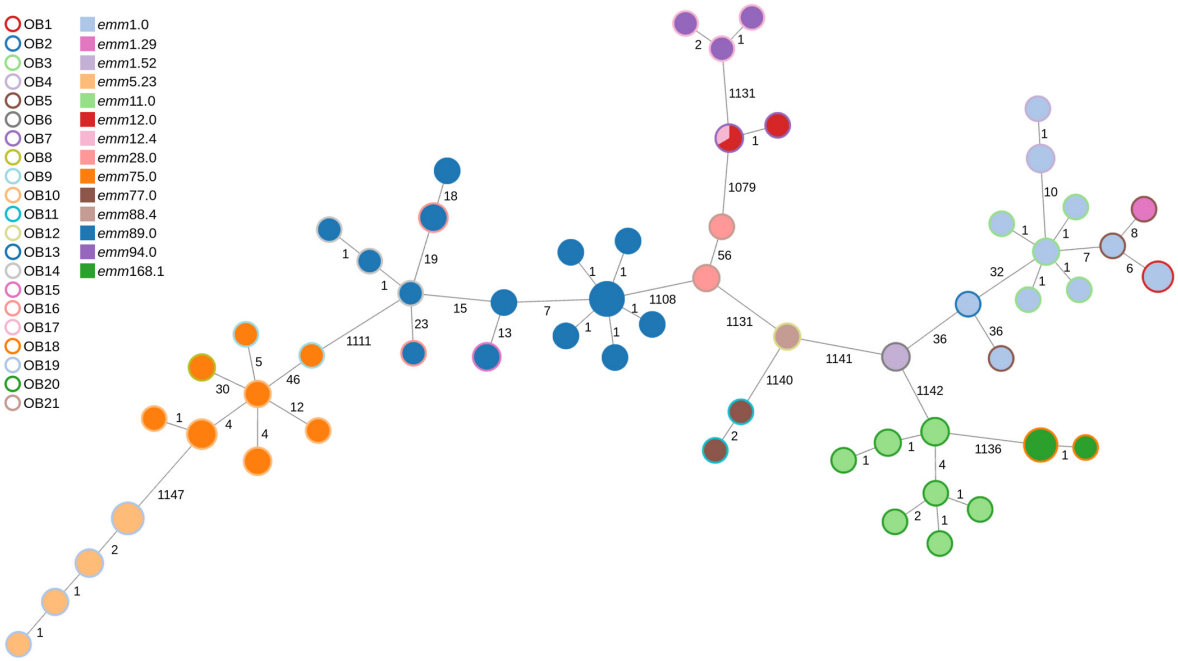


Figure 4.4: Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST-100 profiles of 119 outbreak *S. pyogenes* isolates recovered in the United Kingdom [213] (see Data Set 3 in reference [224]). The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to the *emm* type, and the outer ring is colored according to the outbreak number. Link distances are labeled as the number of allelic differences between nodes (from a total of 1,263 compared loci).

Table 4.2: Distances (numbers of allelic differences) among outbreak isolates and between each outbreak and sporadic isolates of the same *emm* type determined by cgMLST-100 analysis for each *emm* type, using a collection of isolates from the United Kingdom[a]

emm type	No. loci in cgMLST	Subset (no. isolates)	Mean distance within subset (range)	Mean distance to sporadic isolates (range)
1	1488	OB1 (6)	0.6 (0-1)	17.9 (2-59)
		OB3 (6)	1.7 (1-2)	18.4 (8-53)
		OB4 (4)	0.5 (0-1)	24.8 (15-60)
		OB6 (3)	0.7 (0-1)	50.1 (46-55)
		Sporadic (30)	25.7 (3-63)	NA
5	1485	OB19 (14)	1.6 (0-4)	124.1 (0-174)
		Sporadic (27)	112.0 (0-175)	NA
11	1384	OB20 (10)	3.8 (0-8)	89.8 (35-592)
		Sporadic (26)	118.6 (1-597)	NA
28	1510	OB21 (2)	0	44 (12-67)
		Sporadic (11)	51.0 (0-74)	NA
75	1547	OB8 (2)	0	20.1 (14-65)
		OB10 (11)	5.4 (0-11)	19.6 (12-70)
		Sporadic (39)	19.5 (0-76)	NA
89	1392	OB13 (17)	0.93 (0-2)	28.4 (11-42)
		OB14 (3)	1.3 (1-2)	29.0 (16-41)
		OB15 (3)	0 (0-0)	32.5 (16-47)
		OB16 (4)	0.5 (0-1)	33.1 (14-45)
		Sporadic (31)	31.7 (0-50)	NA
94	1506	OB17 (3)	2.7 (1-4)	29.6 (10-48)
		Sporadic (6)	31.4 (2-50)	NA

[a] See reference [213] and Data Set 3 in reference [224]. Ten outbreak isolates were excluded according to the results of both cgMLST-100 and SNP analyses [213]. NA, not applicable.

Performance of the wgMLST schema in the identification of recently emerged lineages. We tested if the proposed wgMLST schema has enough discriminatory power to identify two recently emerged intra-*emm* lineages that were originally identified by whole-genome SNP analysis, namely, $M1_{UK}$ and *emm*89 clade 3 [168, 204, 205]. Allele calling was performed for the 135 assemblies from noninvasive *emm*1 isolates [168] together with the complete genome of strain MGAS5005, a reference representative of the $M1_{global}$ lineage (see Data Set 4 in [224]). The graph in Fig. 4.5 represents the resulting MST with all links of up to 19 differences depicted. All $M1_{UK}$ isolates were tightly clustered, together with an intermediate isolate ($M1_{inter}$) carrying 23 of the 27 SNPs characteristic of the $M1_{UK}$ lineage [168]. The MST links within this cluster ranged between 0 and 13 differences, while the closest links to the $M1_{inter}$ cluster (13 SNPs) and an $M1_{global}$ isolate were 20 and 31 differences, respectively. $M1_{global}$ isolates presented higher genomic diversity, with MST links of up to 49 differences. Allele calling was performed for all *emm*89 assemblies included in the four data sets described above and all the complete *emm*89 genomes used to create the schema (n = 201) (see Data Set 5 in [224]). In addition, the P*nga* variant was determined for all isolates. The absence of the *hasA* gene of the capsule locus was confirmed in all P*nga*-3 isolates, while all other isolates carried this gene, except for two ST568 isolates that have an internal nonsense codon in *hasA*. The graph depicting all links of up to 55 differences (Fig. 4.6) showed limited diversity in the isolates carrying P*nga*-3, which clustered closely, with MST links with 0 to 27 differences, while the shortest link to a P*nga*-2 isolate was 57 differences. The P*nga*-2 and, especially, P*nga*-1 isolates were more diverse, presenting fewer links with up to 55 differences and comprising multiple sublineages associated with different STs (Fig. 4.6; Fig. 4.21). Both the wider geographic range and collection time span may contribute to this higher diversity. As previously reported [203], MLST was not suitable for discriminating P*nga*-3 isolates from those carrying P*nga*-2 since ST101 was prevalent among

both lineages (Fig. 4.21). Analysis of the *emm*89 isolates from Data Set 1 showed that P*nga*-3 isolates and most P*nga*-2 isolates were also grouped into the same PFGE cluster, and some of them shared the same SAg profile, while the T serotype B3264 was ubiquitous, except for the single P*nga*-1 isolate (T11) and one P*nga*-3 isolate that was nontypeable (Fig. 4.22 to 4.24). PopPUNK clustering also could not discriminate P*nga*-3 isolates, which were clustered with isolates carrying P*nga*-1 and P*nga*-2 in phylogroup 3 (see Data Set 5 in [224]).

4.5 Discussion

The reduced costs of HTS have facilitated a wider application of whole-genome data to the epidemiological surveillance of multiple pathogens. This leads to a requirement for standardized analysis pipelines producing reproducible and portable results that can be easily compared across laboratories and with those of previously used typing methods [243]. Here, we propose a wgMLST schema for *S. pyogenes*, consisting of 3,044 loci. Hard-defined cgMLST schemas comprising the subsets of loci present in 95% (1,321 loci), 99% (1,204 loci), and 100% (763 loci) of the assemblies of a collection representing the genetic diversity of *S. pyogenes* [214] are also presented.

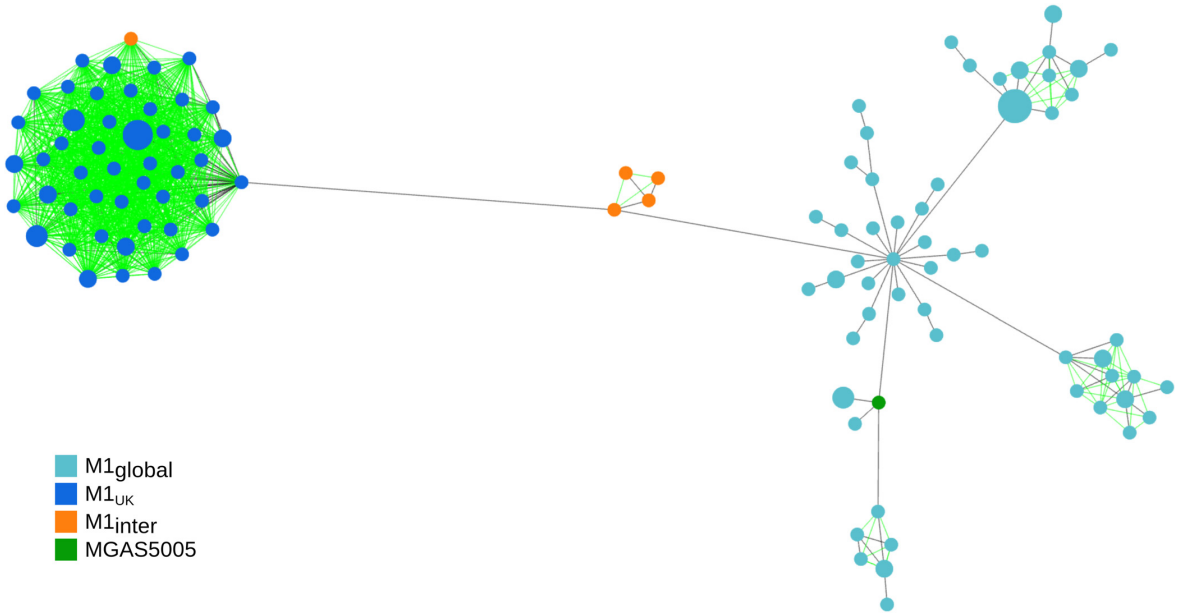


Figure 4.5: Graph representation of the relationships between the cgMLST-100 profiles of 135 noninvasive *emm*1 isolates recovered in the United Kingdom [168] and reference strain MGAS5005 (see Data Set 4 in reference [224]), depicting all links with $\leq$ 19 allelic differences (from a total of 1,404 compared loci). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to the M1 lineage, with MGAS5005 (reference genome for the $M1_{global}$ lineage) in green. Links that would not be present in the standard MST are shown in green. Links shown in black represent the MST links and may represent distances with >19 allelic differences.

However, the use of a wgMLST schema from which the cgMLST loci are selected according to the specific data set under analysis has the advantage of allowing the inclusion of larger

subsets of loci and, hence, increased resolution when comparing closely related isolates [125]. This can be particularly important to track the emergence of intra-*emm*-type sublineages or identify outbreak-related isolates. The application of the schema proposed here to previously published data sets and analysis of the resulting MSTs showed a performance comparable to that of SNP-based methods in distinguishing recently emerged intra-*emm*-type sublineages as well as in identifying clusters of epidemiologically and genetically related isolates associated with local, short-term outbreaks [168, 204, 213]. Analyses based on wg/cgMLST build upon the strengths of GbG approaches, which do not require a reference genome or the removal of regions of recombination [125, 216, 219]. This is particularly important when analyzing collections of genetically diverse lineages, such as in long-term surveillance studies, particularly in organisms where mobile genetic elements and recombination play major roles in genomic plasticity and evolution, such as *S. pyogenes* [214, 217]. Moreover, GbG approaches constitute a framework that has been widely used in surveillance, which can facilitate the transition to wg/cgMLST by reference laboratories involved in surveillance activities. Comparison of cgMLST-based clustering with other typing methods used for *S. pyogenes* revealed poor concordance, although in temporally and geographically restricted data sets, the groups defined by *emm* typing were also supported by cgMLST. By including a much higher number of loci, cgMLST was expected to present a higher discriminatory power than the traditional seven-gene MLST schema and to further discriminate isolates sharing the same ST [216, 220]. However, such a simplistic expectation was not universally borne out by the data, which highlights the limitations of the seven-gene MLST schema to correctly identify GAS lineages based on broader genomic information. It is worth noting that from the seven genes included in traditional MLST, two (*gtr* and *yqiL*) were excluded from the wgMLST schema because they shared alleles with paralogous genes, and one (*xpt*) was absent in at least one GAS lineage and therefore was not always included in the cgMLST analysis. In contrast, a good correlation was found between cgMLST clustering and PopPUNK [142, 214], another whole-genome-based clustering method. However, the flexibility of wg/cgMLST allows increased resolution by lowering the number of allelic differences used to define clusters and a dynamic cgMLST definition, providing further discrimination within PopPUNK clusters. The proposed wgMLST schema is publicly available on the Chewie-NS platform [152], where multiple statistics regarding the whole schema and individual loci can be visualized (`https://chewbbaca.online/species/1/schemas/1`).

The close integration with the chewBBACA suite [143] facilitates its use in surveillance and epidemiological studies and the maintenance of a common nomenclature across different studies. By virtue of the comprehensive annotation, the database can be used to obtain relevant data for basic research, such as the variability of genes of interest (virulence factors, antimicrobial resistance genes, candidate vaccine antigens, and transcriptional regulators, etc.) [214, 215] in addition to its use for typing purposes.

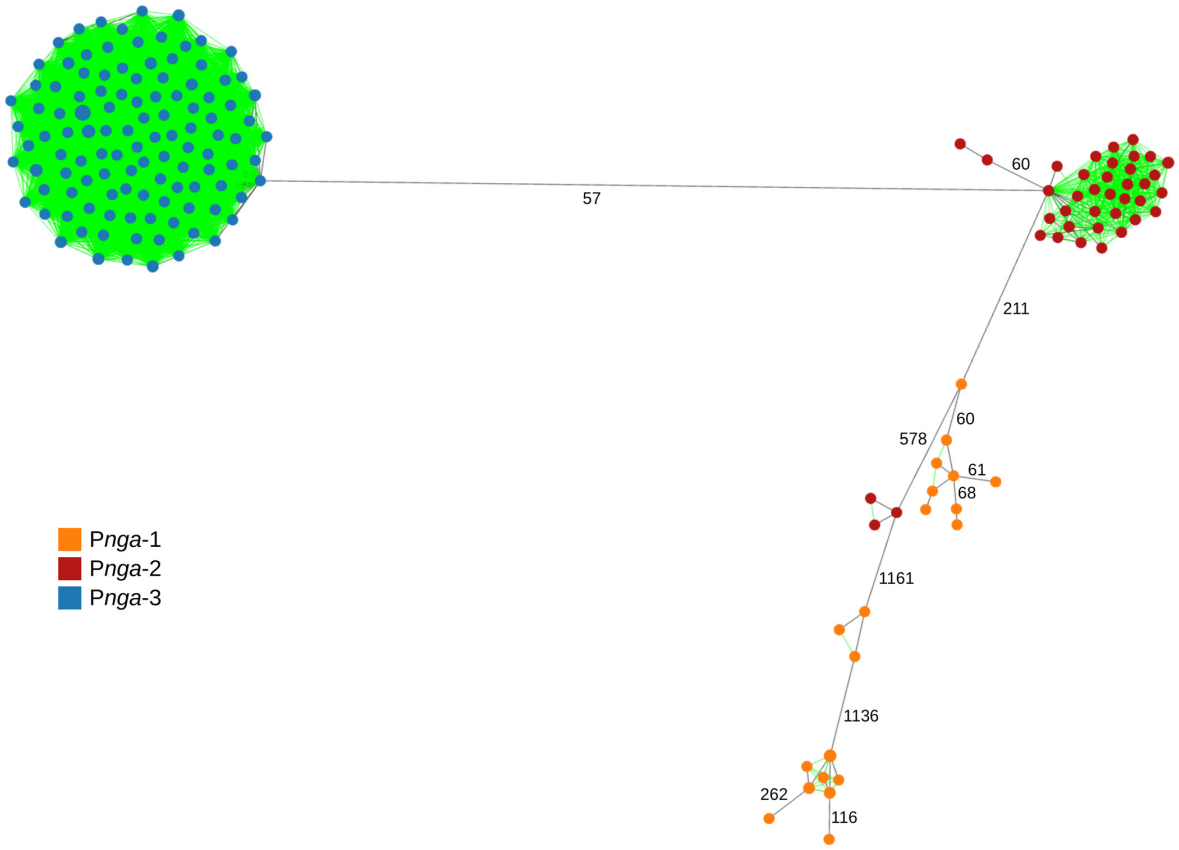


Figure 4.6: Graph representation of the relationships between the cgMLST-100 profiles of 201 *emm*89 isolates (see Data Set 5 in reference [224]) depicting all links with $\leq$ 55 allelic differences (from a total of 1,279 compared loci). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to the variant of the *nga* promoter (P*nga*). Links that would not be present in the standard MST are shown in green. Links shown in black represent the MST links and may represent distances with >55 allelic differences (labeled links).

4.6 Acknowledgments

R.M. was supported by the FCT (grant 2020.08493.BD). Partial support was received from the ONEIDA project (LISBOA-01-0145-FEDER-016417), cofunded by the FEEI (Fundos Europeus Estruturais e de Investimento) from the Programa Operacional Regional de Lisboa, Portugal, 2020 (POR Lisboa 2020), and by national funds from the FCT and the LISBOA-01-0145-FEDER-007391 project, cofunded by FEDER through POR Lisboa 2020 and the FCT.

4.7 Supplemental Material

4.7.1 Supplemental Figures

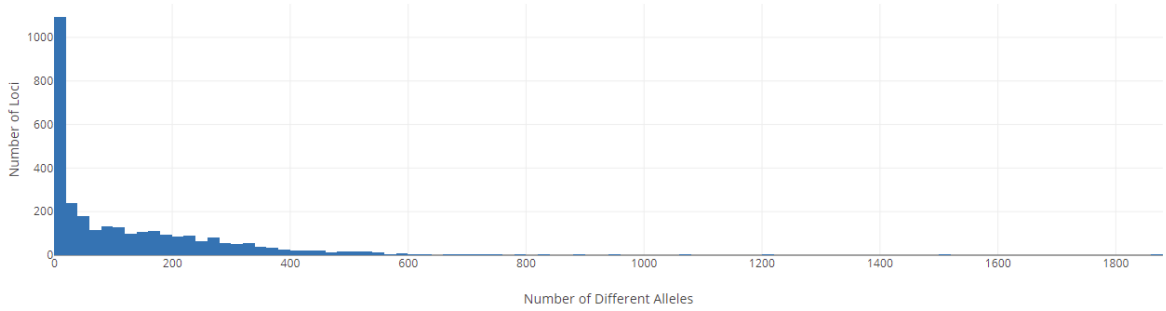


Figure 4.7: Number of loci with given number of alleles in the wgMLST schema of *S. pyogenes* (`https://chewbbaca.online/species/1/schemas/1`).

Figure 4.8: Number of DNA alleles and protein variants of the 10 loci with the largest number of distinct alleles in the *S. pyogenes* wgMLST schema.

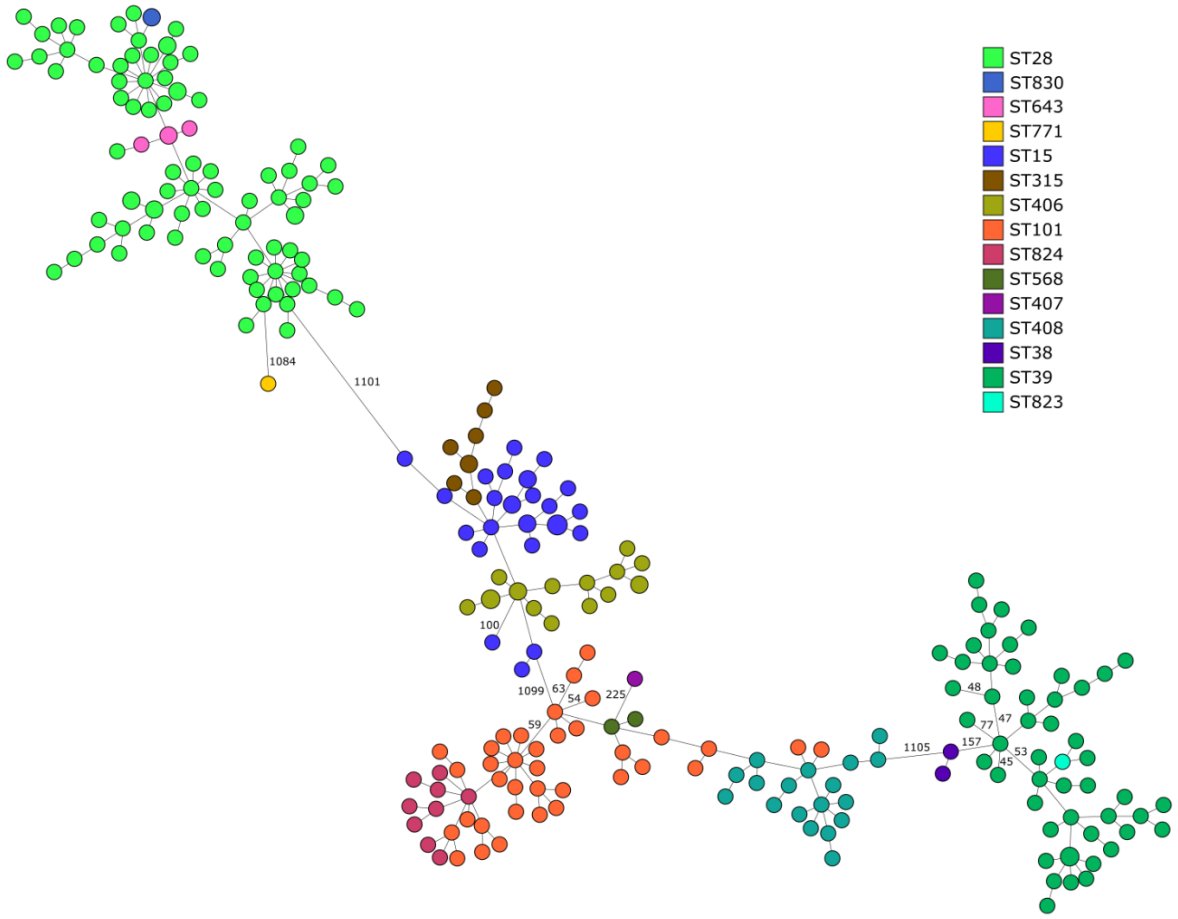


Figure 4.9: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]]. Nodes are colored according to ST. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Link distances $\geq$ 45 differences are labeled (from a total of 1,230 compared loci).

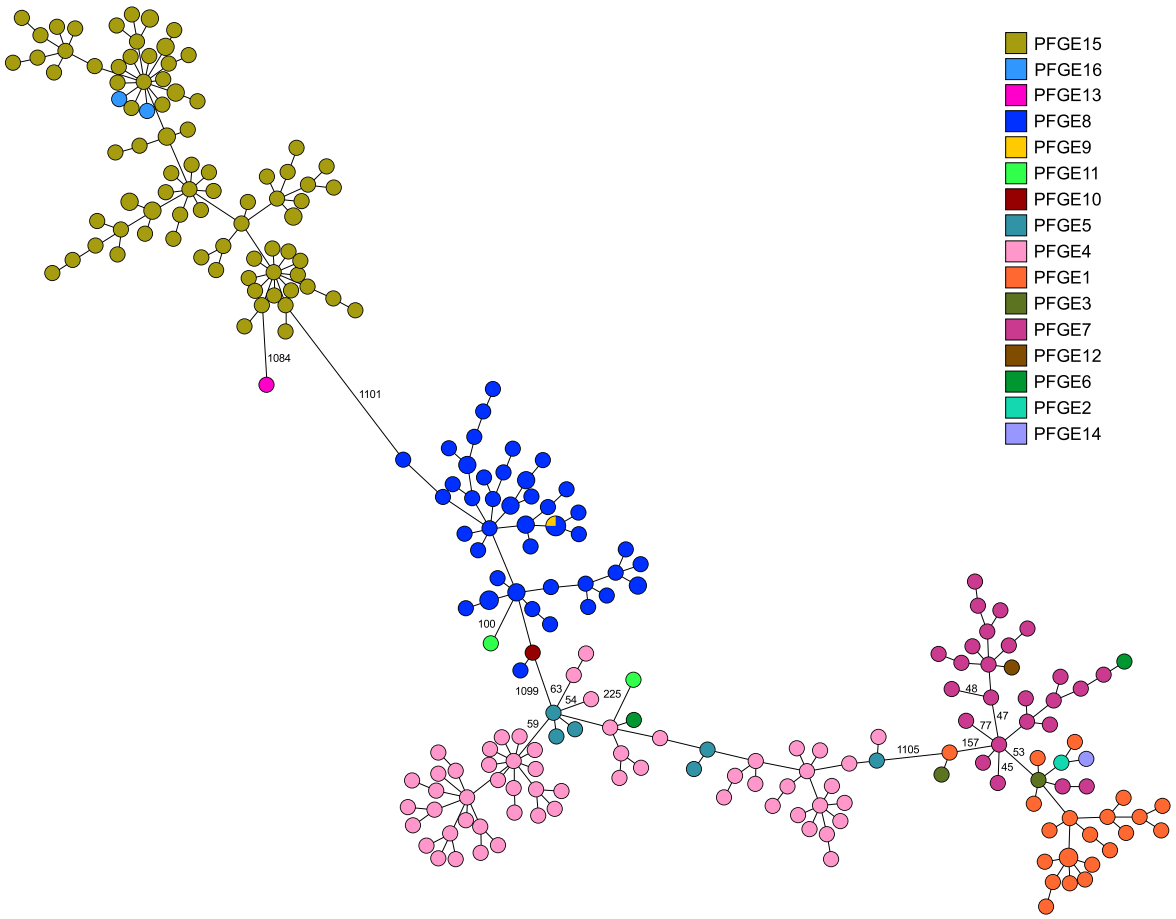


Figure 4.10: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]]. Nodes are colored according to PFGE cluster. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Link distances ≥ 45 differences are labeled (from a total of 1,230 compared loci).

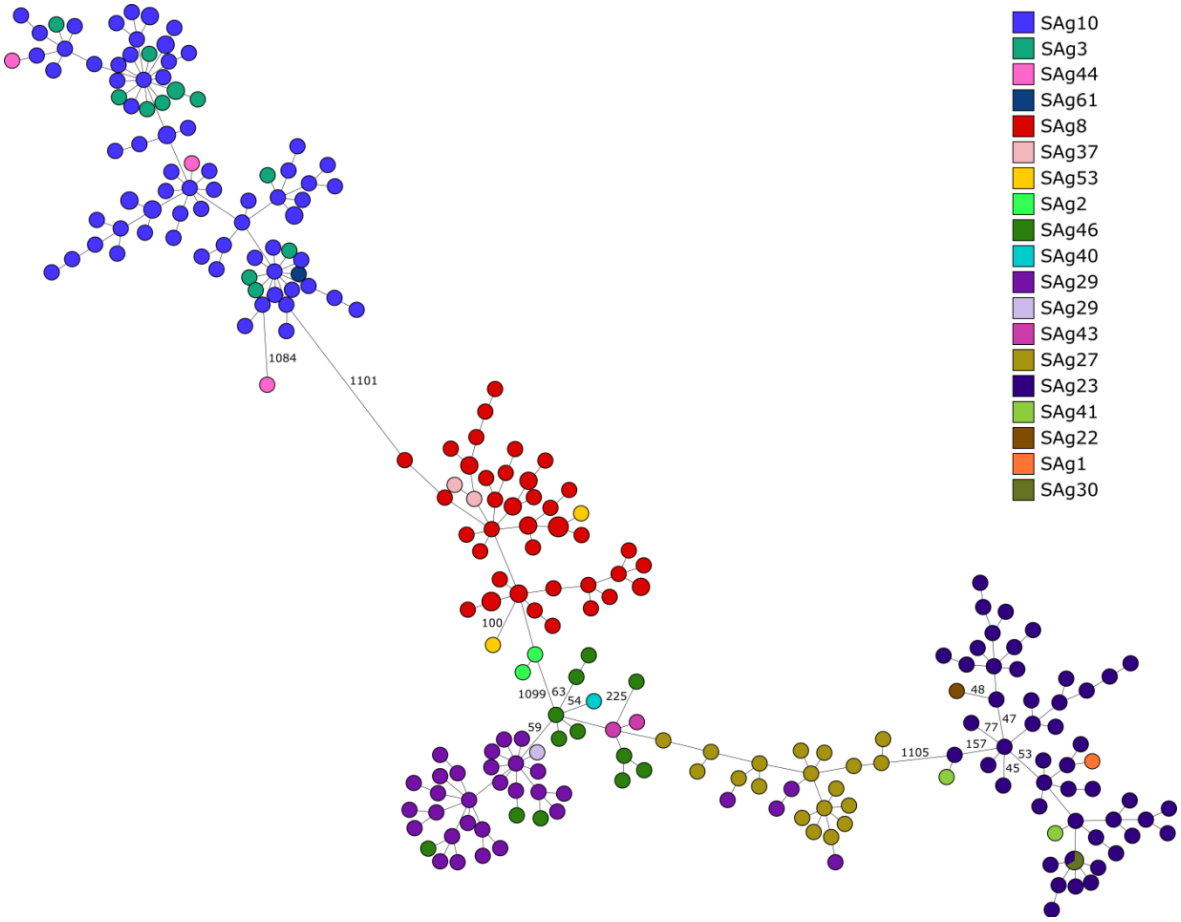


Figure 4.11: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]]. Nodes are colored according to SAg profile. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Link distances $\geq$ 45 differences are labeled (from a total of 1,230 compared loci).

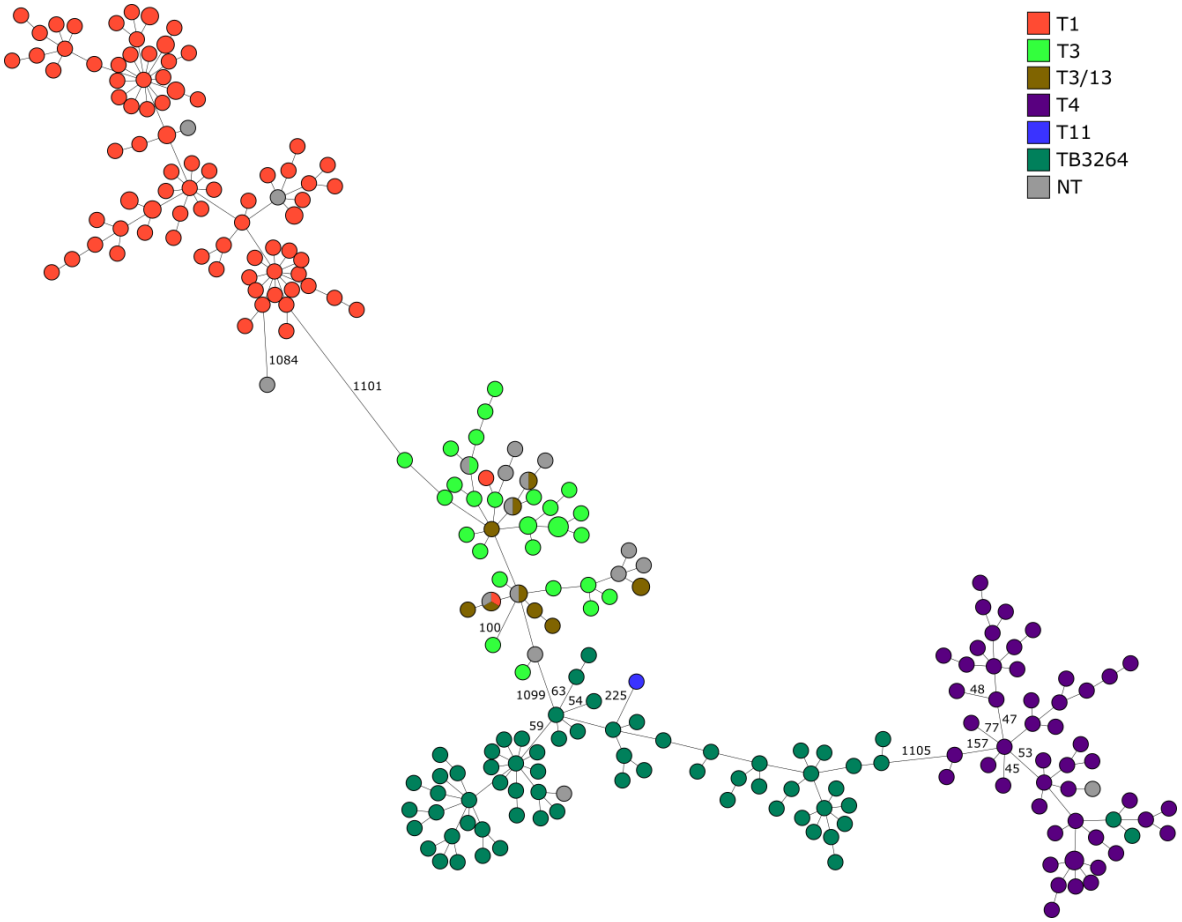


Figure 4.12: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]]. Nodes are colored according to T-type. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Link distances $\geq$ 45 differences are labeled (from a total of 1,230 compared loci). NT, non-typeable.

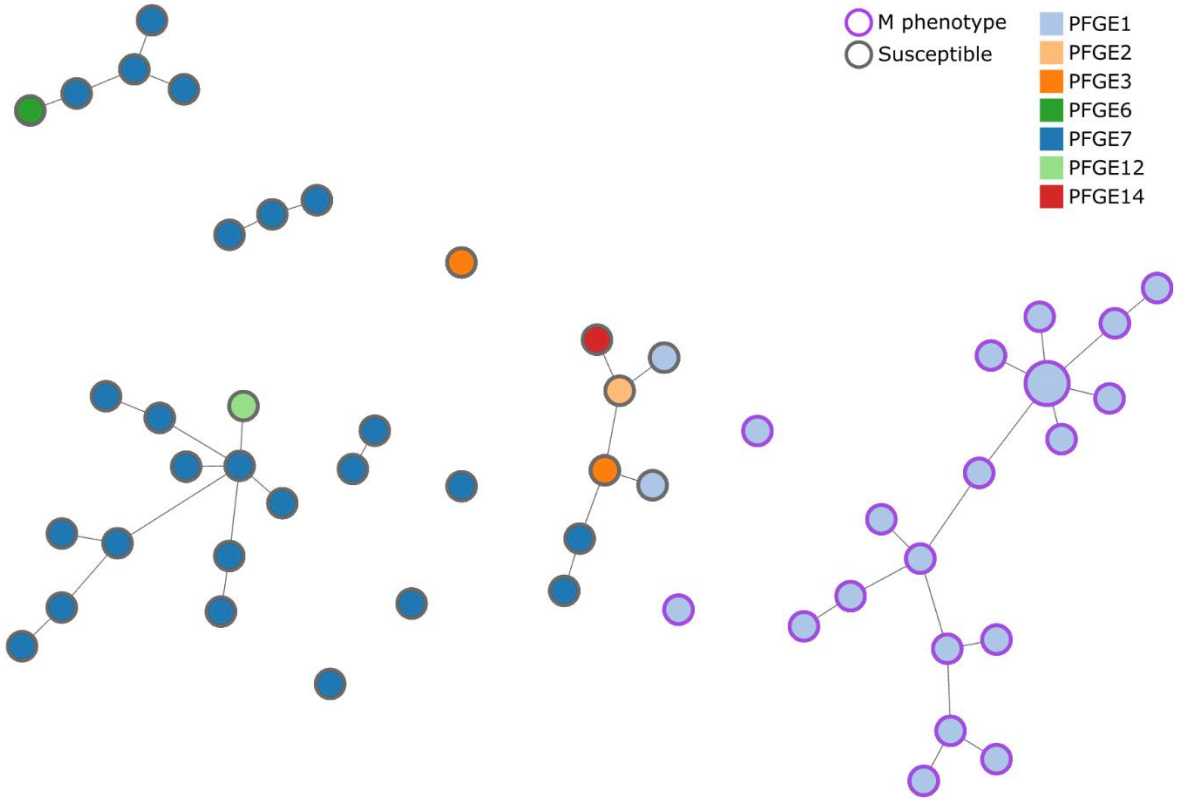


Figure 4.13: Representation of the MST groups defined at 33 allelic differences cutoff (from a total of 1,382 compared loci) for 54 *emm*4 isolates recovered in Portugal [Dataset 1 [224]]. The nodes are colored according to PFGE cluster and the outer ring according to macrolide resistance phenotype. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. M phenotype, erythromycin-resistant and clindamycin-susceptible.

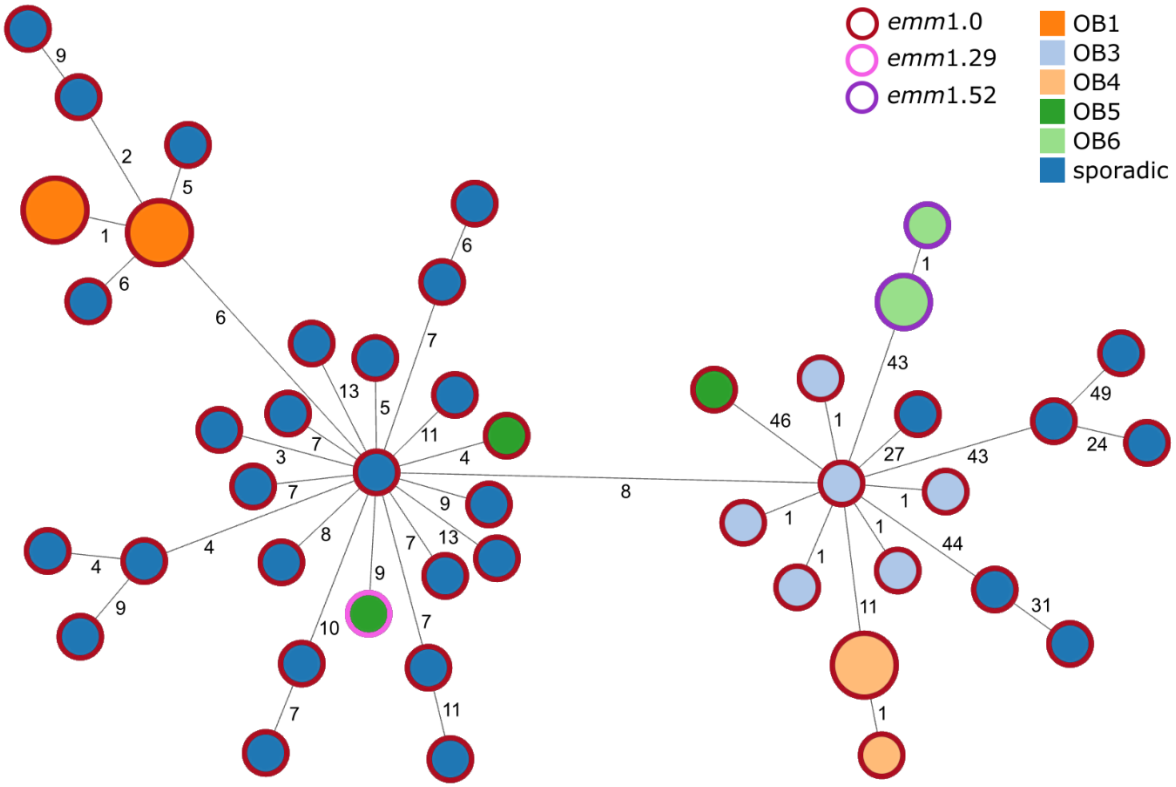


Figure 4.14: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 22 outbreak and 30 sporadic *emm*1 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to outbreak number and the outer ring according to *emm* subtype. Outbreak 2 was not included because after assembly and the exclusion criteria were applied only one representative isolate was kept, which was excluded from the outbreak by SNP analysis [213]. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,488 compared loci). All isolates belong to ST28.

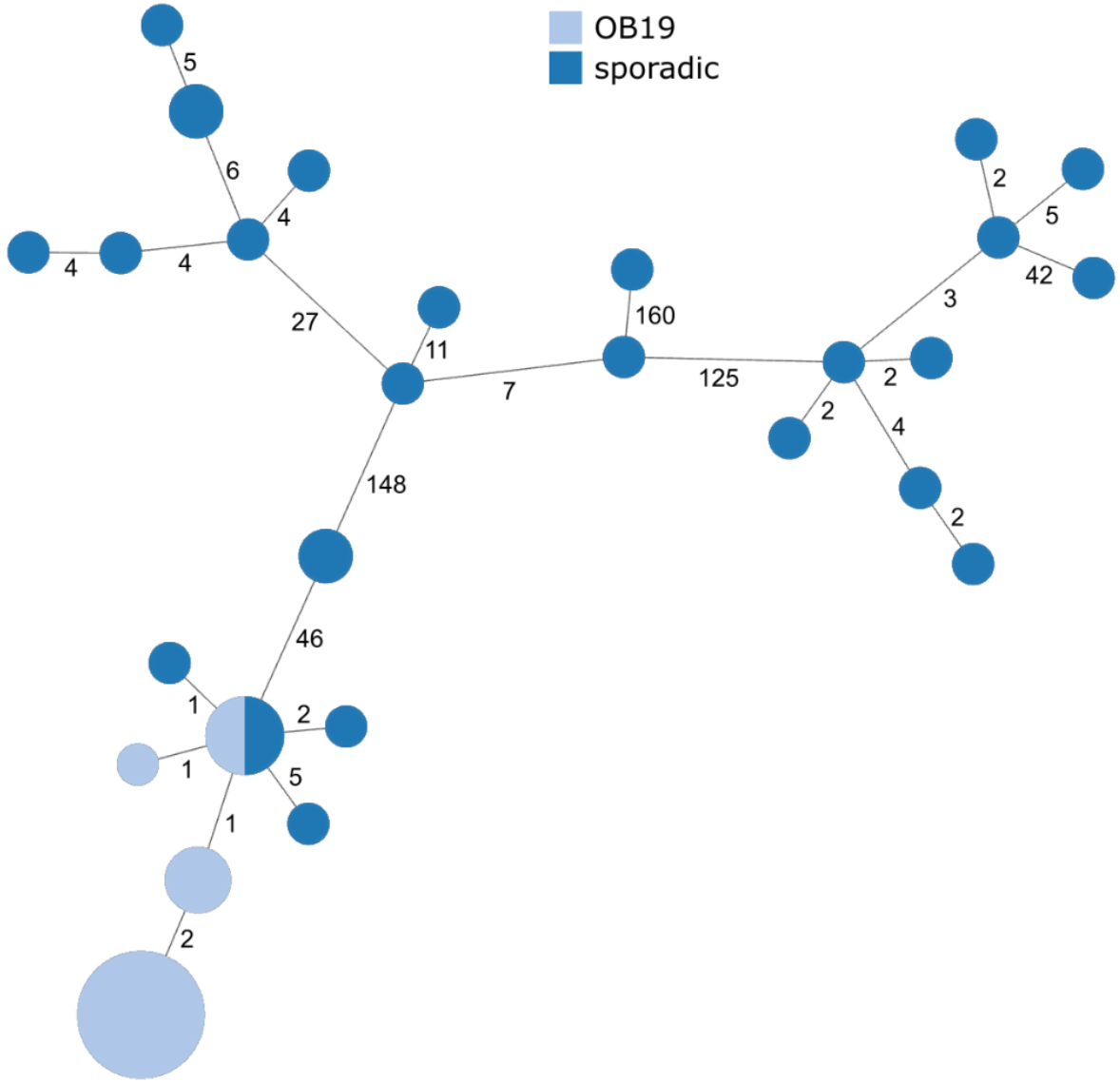


Figure 4.15: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 14 outbreak and 27 sporadic *emm*5 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Nodes are colored according to outbreak number. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,485 compared loci). All isolates belong to ST99.

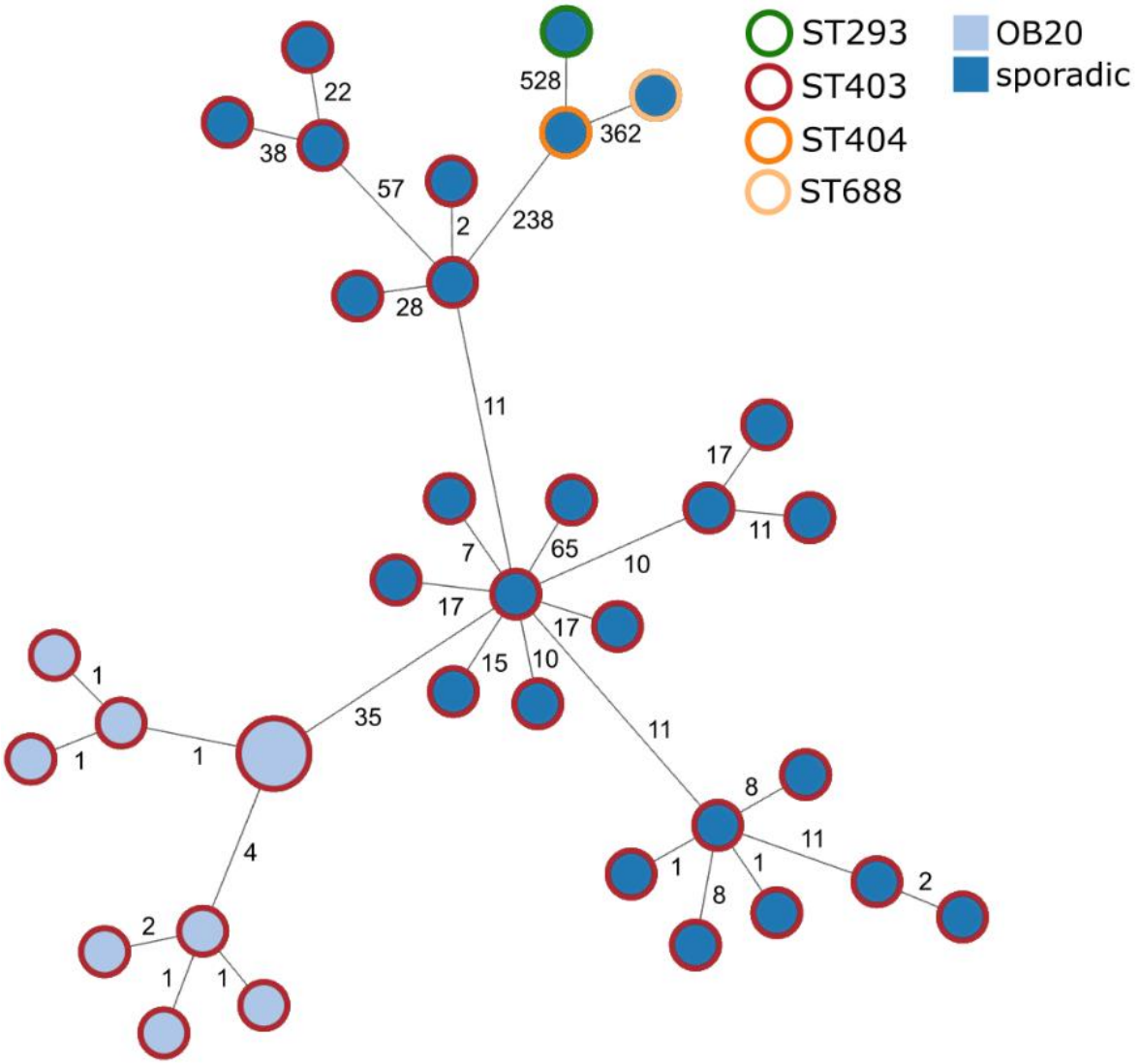


Figure 4.16: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 10 outbreak and 26 sporadic *emm*11 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to outbreak number and the outer ring according to ST. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,384 compared loci).

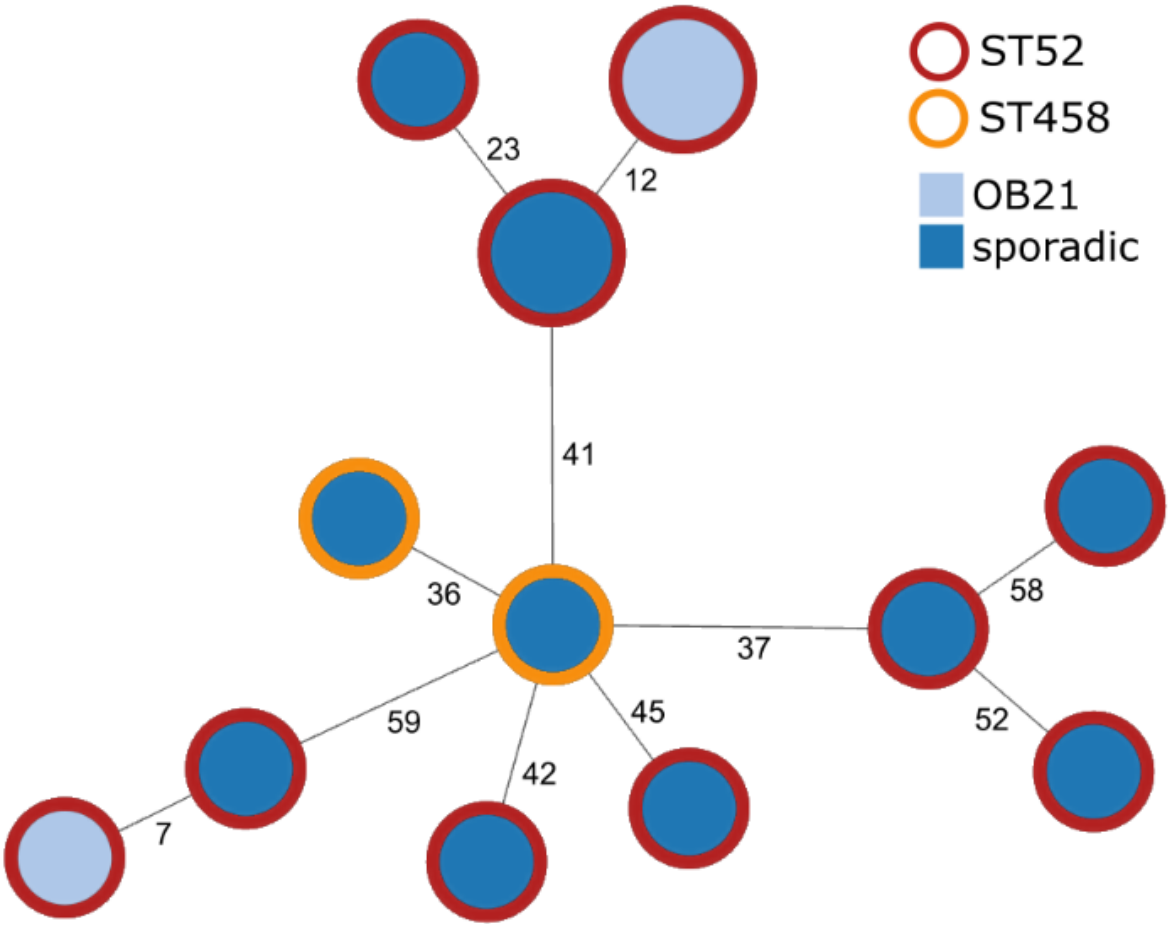


Figure 4.17: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 3 outbreak and 11 sporadic *emm*28 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to outbreak number and the outer ring according to ST. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,510 compared loci).

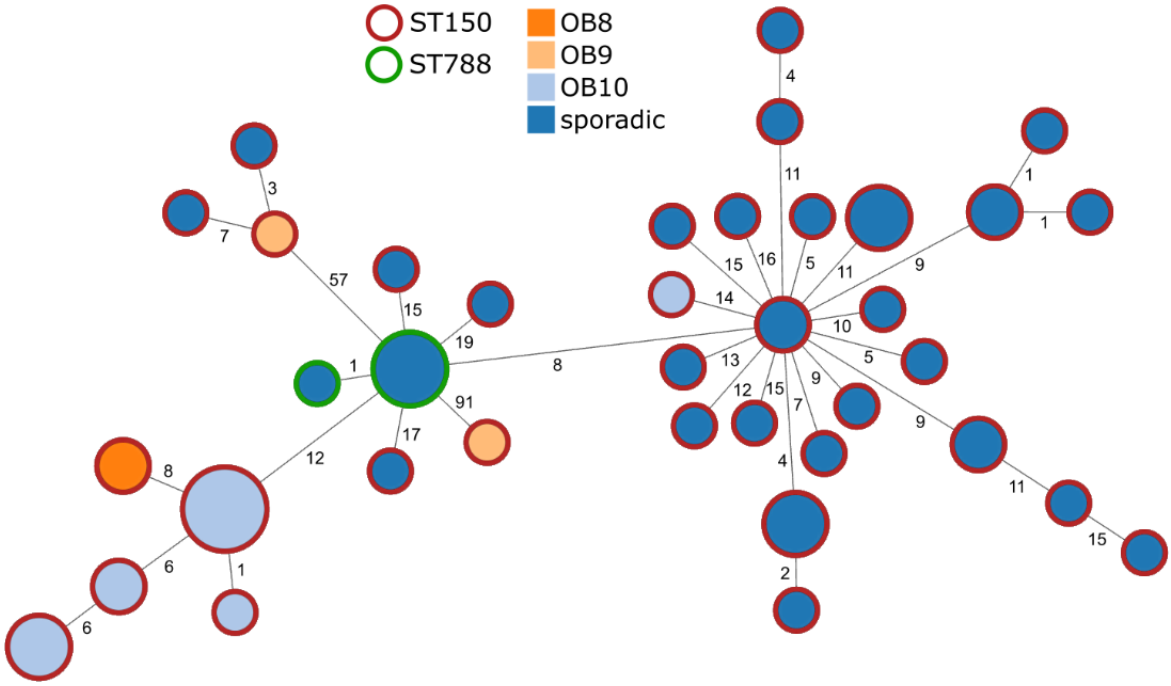


Figure 4.18: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 16 outbreak and 39 sporadic *emm*75 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to outbreak number and the outer ring according to ST. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,547 compared loci).

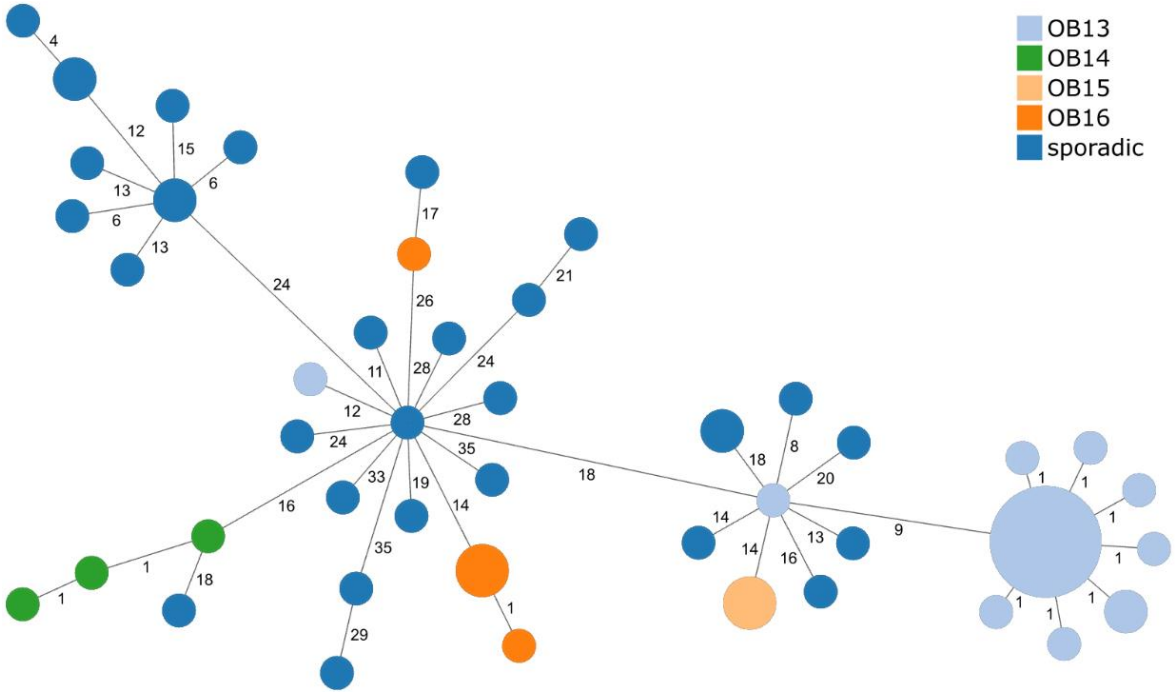


Figure 4.19: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 30 outbreak and 31 sporadic *emm*89 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Nodes are colored according to outbreak number. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,392 compared loci). All isolates belong to ST101.

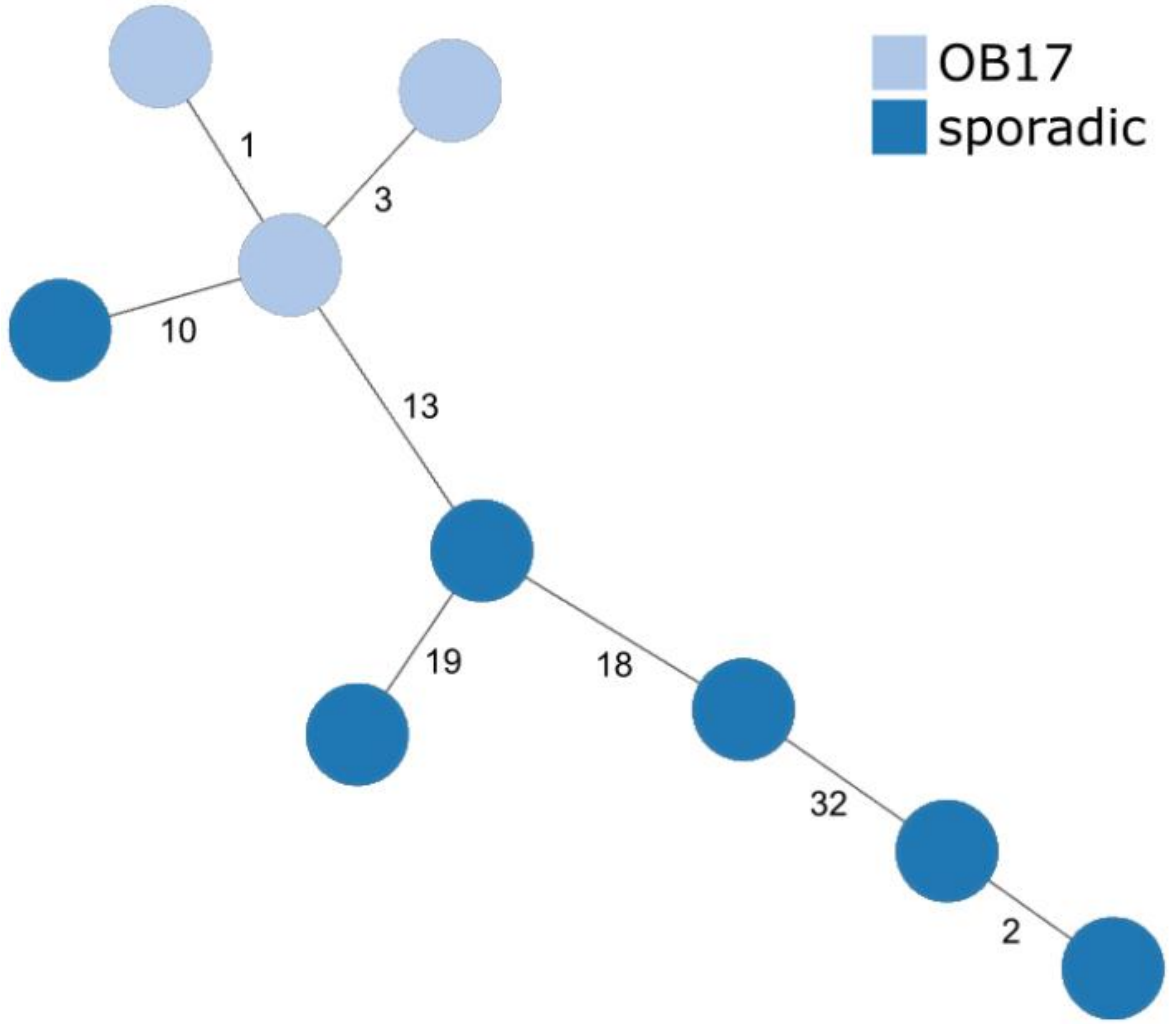


Figure 4.20: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 3 outbreak and 6 sporadic *emm*94 isolates recovered in the UK [213] [Dataset 3 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Nodes are colored according to outbreak number. Link distances are represented as the total number of allelic differences between nodes (from a total of 1,506 compared loci). All isolates belong to ST89.

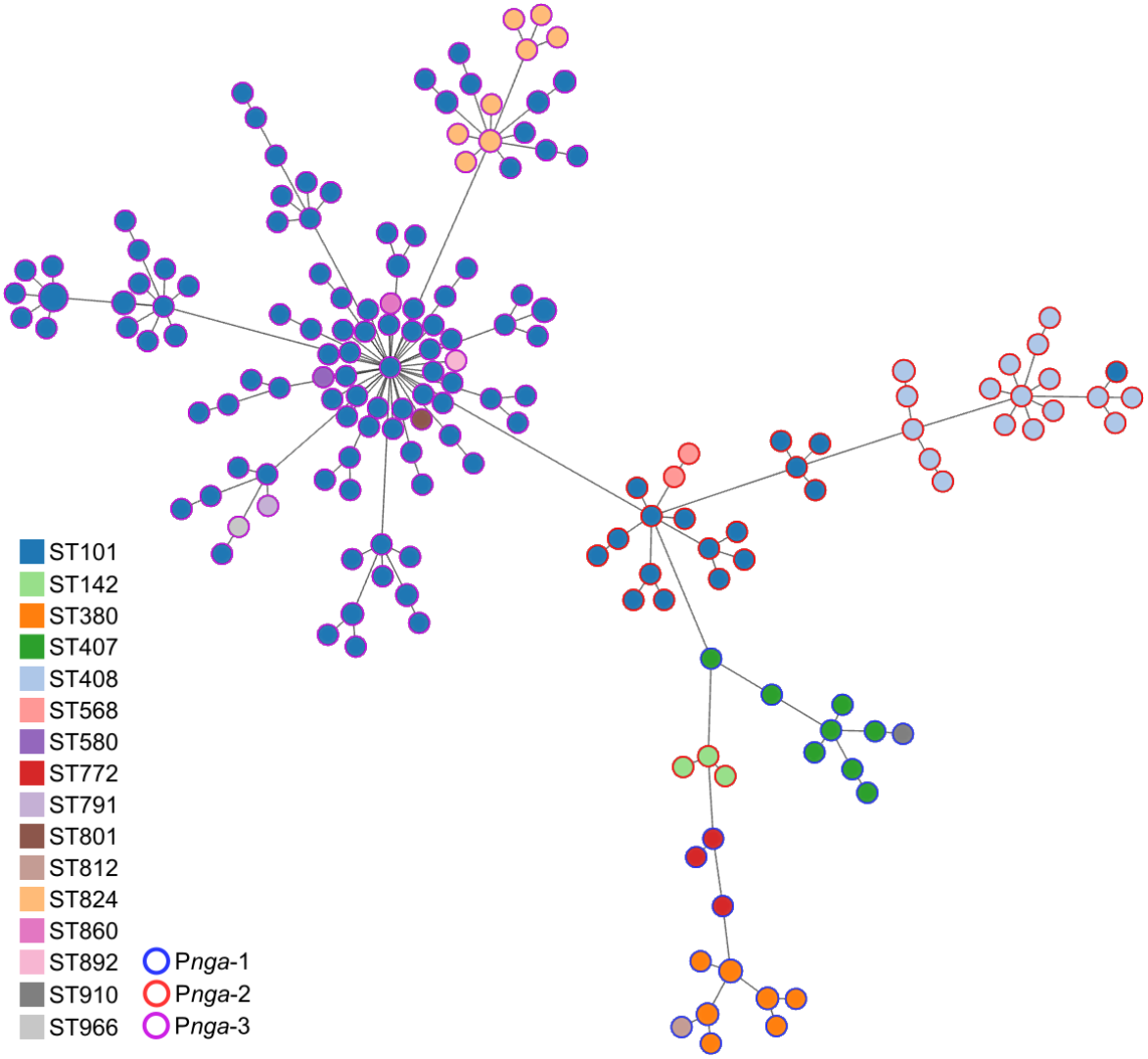


Figure 4.21: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 201 *emm*89 isolates [Dataset 5 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to ST and the outer ring according to the variant of the *nga* promoter (P*nga*). A total of 1,279 loci were compared.

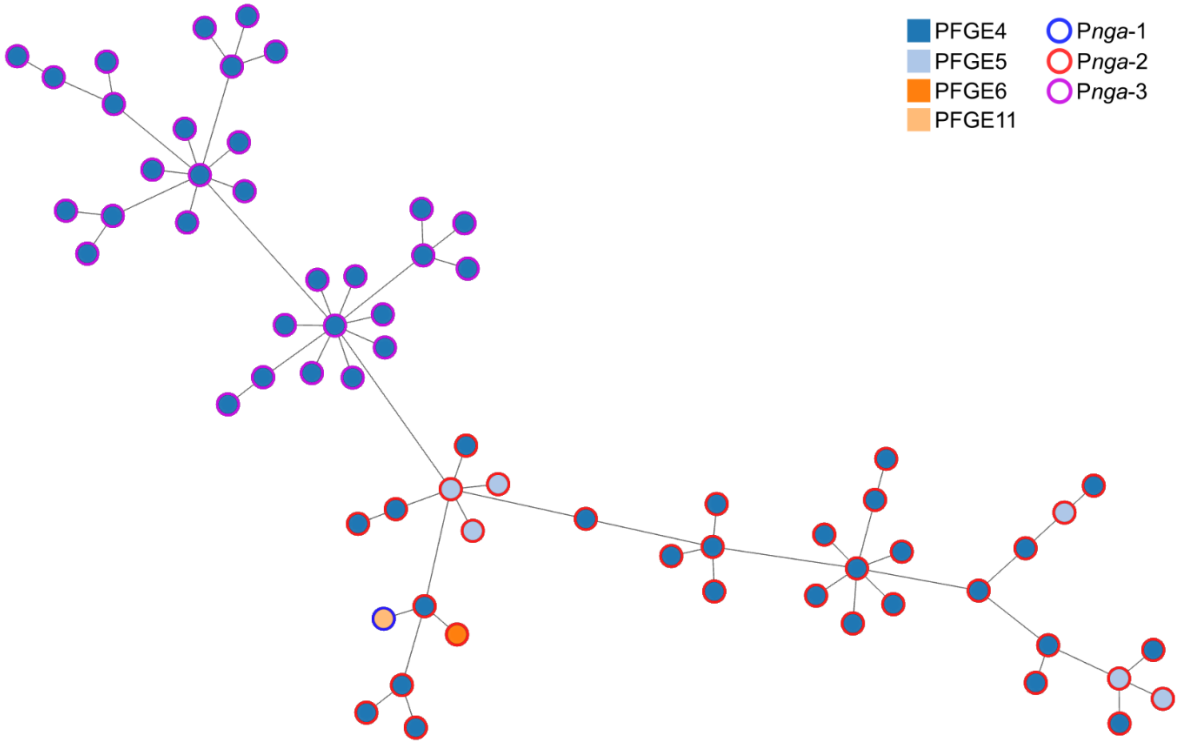


Figure 4.22: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 66 *emm*89 isolated in Portugal [Dataset 1 and Dataset 5 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to PFGE cluster and the outer ring according to the variant of the *nga* promoter (P*nga*). A total of 1,388 loci were compared.

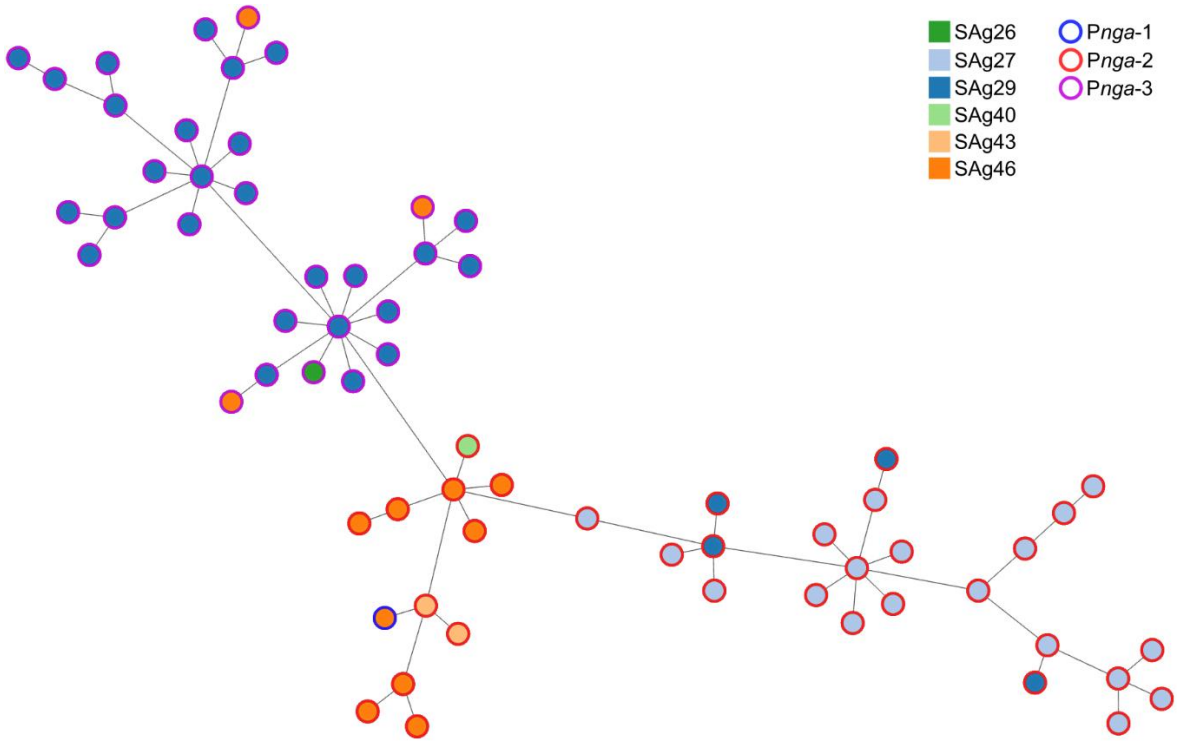


Figure 4.23: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 66 *emm*89 isolated in Portugal [Dataset 1 and Dataset 5 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to SAg gene profile and the outer ring according to the variant of the *nga* promoter (P*nga*). A total of 1,388 loci were compared.

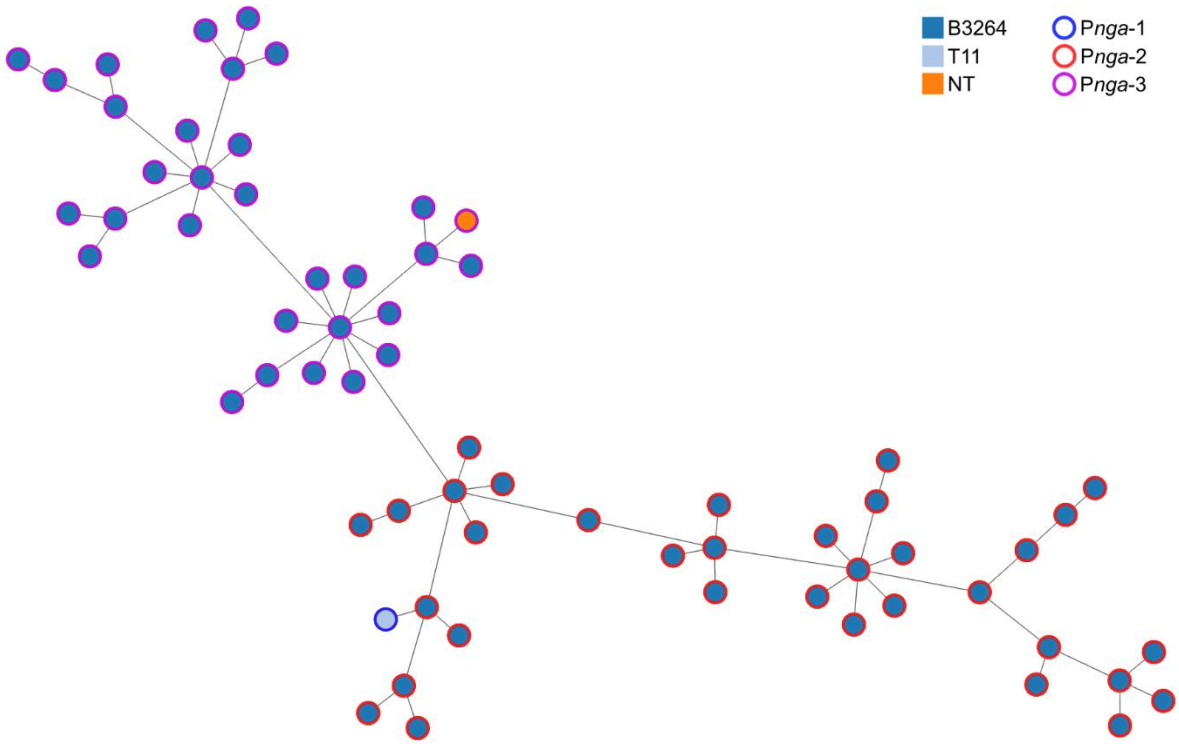


Figure 4.24: Minimum spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 66 *emm*89 isolated in Portugal [Dataset 1 and Dataset 5 [224]]. The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to T-type and the outer ring according to the variant of the *nga* promoter (P*nga*). A total of 1,388 loci were compared.

4.7.2 Supplemental Tables

Table S1 – Complete genomes used for schema creation. See separate *xlsx* file [224].

Table S2 – Changes applied to the initial schema. See separate *xlsx* file [224].

Table 4.3: Adjusted Rand (AR) values for the clustering methods used in the analysis of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]].

	SAg profile	*emm* type	ST	PFGE	T-type[a]	MST_{1000}[b]
emm type	0.725					
ST	0.65	0.755				
PFGE	0.705	0.861	0.675			
T-type[a]	0.649	0.927	0.755	0.806		
MST_{1000}[b]	0.729	0.996	0.759	0.865	0.927	
MST_{45}[b]	0.729	0.815	0.709	0.846	0.744	0.818

[a] The AR values for T-type were calculated for the subset of 248 isolates with a defined T-type (17 isolates were non-typeable).

[b] Groups of isolates linked by up to n different loci in the MST (MSTn).

Table 4.4: Adjusted Wallace (AW) values (95% confidence intervals) for the clustering methods used in the analysis of 265 *S. pyogenes* isolates recovered in Portugal [Dataset 1 [224]].

	cgMLST-100	SAg profile	T-type[a]	*emm* type	ST	PFGE	MST_{1000}[b]	MST_{45}[b]
cgMLST-100		0.904 (0.814-0.995)	0.929 (0.858-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	0.848 (0.688-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)
SAg profile	0.003 (0.000-0.008)		0.885 (0.844-0.927)	1.000 (0.999-1.000)	0.671 (0.599-0.743)	0.821 (0.791-0.851)	1.000 (0.999-1.000)	0.804 (0.777-0.831)
T-type[a]	0.002 (0.000-0.005)	0.512 (0.430-0.594)		0.948 (0.897-0.998)	0.627 (0.553-0.702)	0.721 (0.648-0.794)	0.948 (0.897-0.998)	0.632 (0.586-0.679)
emm type	0.002 (0.000-0.005)	0.569 (0.493-0.644)	0.907 (0.870-0.945)		0.606 (0.542-0.670)	0.756 (0.698-0.815)	0.992 (0.977-1.000)	0.687 (0.658-0.716)
ST	0.003 (0.000-0.008)	0.630 (0.539-0.721)	0.948 (0.915-0.980)	1.000 (1.000-1.000)		0.759 (0.703-0.816)	1.000 (1.000-1.000)	0.756 (0.725-0.787)
PFGE	0.002 (0.000-0.006)	0.617 (0.535-0.699)	0.915 (0.889-0.942)	1.000 (0.999-1.000)	0.608 (0.541-0.675)		1.000 (0.999-1.000)	0.807 (0.787-0.828)
MST_{1000}	0.002 (0.000-0.005)	0.573 (0.498-0.648)	0.907 (0.870-0.945)	1.000 (1.000-1.000)	0.611 (0.548-0.674)	0.763 (0.704-0.821)		0.693 (0.664-0.721)
MST_{45}	0.003 (0.000-0.007)	0.666 (0.575-0.757)	0.903 (0.872-0.935)	1.000 (1.000-1.000)	0.667 (0.595-0.739)	0.889 (0.833-0.945)	1.000 (1.000-1.000)	

[a] The AW values for T-type were calculated for the subset of 248 isolates with a defined T-type (17 isolates were non-typeable).

[b] Groups of isolates linked by up to *n* different loci in the MST (MST*n*).

Table S5 – Minimum, maximum and mean distances (number of allelic differences) among isolates of *emm* types comprising ≥ 10 isolates and shortest distances to other *emm* types. See separate *xlsx* file [224].

Table 4.5: Simpson's index of diversity (SID) and 95% confidence intervals ($CI_{95\%}$) for the clustering methods used in the analysis of 2,006 genetically diverse *S. pyogenes* isolates recovered worldwide [Dataset 2 [224]].

Clustering method	No. of partitions	SID ($CI_{95\%}$)
emm type	140	0.985 (0.984-0.986)
ST	443	0.993 (0.992-0.993)
PopPUNK	292	0.989 (0.989-0.990)
MST_{450}[a]	192	0.985 (0.984-0.986)
MST_{200}[a]	306	0.990 (0.989-0.991)
MST_{50}[a]	474	0.992 (0.991-0.993)
cgMLST-100	1700	1.000 (1.000-1.000)

[a] Groups of isolates linked by up to different loci in the MST (MST*n*)

Table 4.6: Adjusted Rand (AR) values for the clustering methods used in the analysis of 2,006 genetically diverse *S. pyogenes* isolates recovered worldwide [Dataset 2 [224]].

	emm type	ST	PopPUNK	MST_{450}[a]	MST_{200}[a]
ST	0.624				
PopPUNK	0.773	0.802			
MST_{450}[a]	0.694	0.65	0.823		
MST_{200}[a]	0.755	0.818	0.963	0.808	
MST_{50}[a]	0.683	0.81	0.856	0.688	0.872

[a] Groups of isolates linked by up to *n* different loci in the MST (MST*n*).

Table 4.7: Adjusted Wallace (AW) values (95% confidence intervals) for the clustering methods used in the analysis of 2,006 genetically diverse *S. pyogenes* isolates recovered worldwide [Dataset 2 [224]].

	emm type	ST	PopPUNK	MST_{450}[a]	MST_{200}[a]	MST_{50}[a]	cgMLST-100
emm type		0.463 (0.438-0.488)	0.657 (0.636-0.679)	0.695 (0.674-0.716)	0.634 (0.614-0.654)	0.521 (0.499-0.542)	0.023 (0.018-0.028)
ST	0.958 (0.939-0.977)		0.982 (0.974-0.991)	0.998 (0.996-1.000)	0.984 (0.976-0.992)	0.845 (0.821-0.869)	0.047 (0.039-0.056)
PopPUNK	0.937 (0.922-0.953)	0.677 (0.646-0.708)		1.000 (1.000-1.000)	0.948 (0.941-0.955)	0.749 (0.725-0.773)	0.033 (0.026-0.039)
MST_{450}[a]	0.694 (0.669-0.718)	0.482 (0.452-0.511)	0.700 (0.675-0.725)		0.678 (0.653-0.704)	0.524 (0.496-0.552)	0.023 (0.018-0.028)
MST_{200}[a]	0.933 (0.916-0.949)	0.7 (0.668-0.732)	0.978 (0.972-0.983)	1.000 (1.000-1.000)		0.772 (0.748-0.797)	0.034 (0.027-0.040)
MST_{50}[a]	0.991 (0.983-1.000)	0.778 (0.742-0.813)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)		0.044 (0.036-0.051)
cgMLST-100	0.996 (0.991-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	

[a] Groups of isolates linked by up to n different loci in the MST (MSTn).

Table S9 – Minimum, maximum and mean distances (number of allelic differences) among isolates of PopPUNK phylogroups comprising ≥ 10 isolates and shortest distances to other phylogroups. See separate *xlsx* file [224].

Table 4.8: Isolates with epidemiological links to outbreak isolates [213] that were excluded based on cgMLST analysis. Distances are expressed as number of allelic differences.

Excluded isolate (Accession No.)	Potential outbreak[a]	*emm* type (total no. of loci in cgMLST-100)	Minimum distance to outbreak isolates	Mean distance among true outbreak isolates (range)	Mean distance among sporadic isolates (range)
ERR1732732			9		
ERR1733247	OB5	1 (1,488)	9	NA[b]	26.9 (3-63)
ERR1735250			51		
ERR1734190	OB13	89 (1,392)	25	0.93 (0-2)	31.7 (0-50)
ERR1734292			9		
ERR1733331	OB16	89 (1,392)	34	0.5 (0-1)	31.7 (0-50)
ERR1733070	OB21	28 (1,510)	70	0	44 (12-67)
ERR1734077	OB9	75 (1,547)	114	NA[b]	19.5 (0-76)
ERR1734858			114		
ERR1732609	OB10	75 (1,547)	19	5.4 (0-11)	19.5 (0-76)

[a] Outbreak nomenclature is the same from Coelho et al. 2019 [213].

[a] OB5 and OB9 are not likely to represent true outbreaks, as concluded also in the SNP analysis [213].

4.7.3 Other Supplemental Material

The following supplemental files are available on Zenodo ([224]).

Dataset1.xlsx – relevant metadata for the isolates included in Dataset 1.

Dataset2.xlsx – relevant metadata for the isolates included and excluded from Dataset 2 [214].

Dataset3.xlsx – relevant metadata for the isolates included and excluded from Dataset 3 ([213].

Dataset4.xlsx – relevant metadata for the isolates included in Dataset 4 [168].

Dataset5.xlsx – relevant metadata for the isolates included in Dataset 5.

Dataset1_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 265 isolates included in Dataset 1.

Dataset2_selected_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 2,006 isolates from Davies *et al.* [214] that were selected to create Dataset 2.

Dataset2_excluded_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 77 isolates from Davies *et al.* [214] that were not included in Dataset 2.

Dataset3_selected_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 289 isolates from Coelho *et al.* [213] that were selected to create Dataset 3.

Dataset3_excluded_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 33 isolates from Coelho *et al.* [213] that were not included in Dataset 3.

Dataset4_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 136 isolates included in Dataset 4.

Dataset5_genome_assemblies.zip – archive file that contains the genome assemblies in FASTA format for the 201 isolates included in Dataset 5.

Streptococcus_pyogenes_wgMLST_schema.zip – archive file with the schema populated after performing allele calling with the genome assemblies from all datasets and sourced from public databases and after applying all changes listed in the file Table S2.

cgMLST95_schema_loci_ids.txt – subset of loci present in 95% of the strains included in

Dataset 2.

cgMLST99_schema_loci_ids.txt – subset of loci present in 99% of the strains included in Dataset 2.

cgMLST100_schema_loci_ids.txt – subset of loci present in 100% of the strains included in Dataset 2.

Transcriptional_regulators_schema_loci_ids.txt – subset of loci encoding transcriptional regulators.

Virulence_factors_schema_loci_ids.txt – subset of loci encoding virulence factors.

Dataset1_allelecall_results_clean.tsv – masked allele call matrix with the allelic profiles for the 265 isolates included in Dataset 1.

Dataset1_allelecall_results_raw.tsv – raw allele call matrix with the allelic profiles for the 265 isolates included in Dataset 1.

Dataset2_allelecall_results_clean.tsv – masked allele call matrix with the allelic profiles for the 2,006 isolates included in Dataset 2.

Dataset2_allelecall_results_raw.tsv – raw allele call matrix with the allelic profiles for the 2,006 isolates included in Dataset 2.

Dataset3_allelecall_results_clean.tsv – masked allele call matrix with the allelic profiles for the 289 isolates included in Dataset 3.

Dataset3_allelecall_results_raw.tsv – raw allele call matrix with the allelic profiles for the 289 isolates included in Dataset 3.

Dataset4_allelecall_results_clean.tsv – masked allele call matrix with the allelic profiles for the 136 isolates included in Dataset 4.

Dataset4_allelecall_results_raw.tsv – raw allele call matrix with the allelic profiles for the 136 isolates included in Dataset 4.

Dataset5_allelecall_results_clean.tsv – masked allele call matrix with the allelic profiles for the 201 isolates included in Dataset 5.

Dataset5_allelecall_results_raw.tsv – raw allele call matrix with the allelic profiles for the 201 isolates included in Dataset 5.

exclusive_loci_emm4_Mphenotype.txt – list with the 46 loci that were present universally and exclusively in the subset of erythromicin-resistant *emm*4 isolates from Dataset 1.

Dataset2_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances

(number of allelic differences amongst shared loci) between the 2,006 isolates included in Dataset 2.

Dataset3_emm1_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 52 *emm*1 isolates included in Dataset 3.

Dataset3_emm5_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 41 *emm*5 isolates included in Dataset 3.

Dataset3_emm11_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 36 *emm*11 isolates included in Dataset 3.

Dataset3_emm28_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 14 *emm*28 isolates included in Dataset 3.

Dataset3_emm75_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 55 *emm*75 isolates included in Dataset 3.

Dataset3_emm89_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 61 *emm*89 isolates included in Dataset 3.

Dataset3_emm94_cgMLST_pairwise_distances.tsv – distance matrix with the pairwise distances between the 9 *emm*94 isolates included in Dataset 3.

Figure2A.html – HTML file for the interactive version of Figure 2, panel A, included in the manuscript.

Figure2B.html – HTML file for the interactive version of Figure 2, panel B, included in the manuscript.

Chapter 5

General Discussion

5.1 A brief note on software development practices

I start this discussion by mentioning a subject that tends to be overlooked in bioinformatics: good practices in software development. After the initial thrill of coming up with a concept for a tool and implementing a rudimentary proof-of-concept, it is time to tidy up. Developing bioinformatics software that is efficient, maintainable, and reliable is the shortest route to contribute to its wide adoption and long-term success [181, 244, 245]. The process and effort of writing well-structured and documented code can take up as much time as the algorithmic intricacies, but it is essential to ensure maintainability. While I think it is not necessary to follow each piece of advice given by Uncle Bob[1], defining a handful of simple requirements and practices for sustained software development over time can maximize software applicability and help avoid turning something simple into a convoluted mess [246].

While working on chewBBACA 3 and Chewie-NS, presented in **Chapter 2** and **Chapter 3**, I aimed to maintain good code quality standards, as I knew that this could greatly influence long-term support. In chewBBACA 3, the compartmentalization of the code into multiple modules containing smaller functions played an important role in organizing the code to maximize reusability and maintainability. When evident, functions were grouped into modules based on more general categories, such as functions for file operations (reading and writing files), sequence manipulation and clustering (processing and clustering of DNA and protein sequences), and multiprocessing operations (functions to divide and process inputs in parallel to improve efficiency). Compartmentalization helped to keep track of the set of functions that I had already implemented, reducing the tendency to reimplement functions in distinct parts

[1] `https://en.wikipedia.org/wiki/Robert_C._Martin`

of the code, and providing a collection of reusable functions interconnecting the code and making the codebase more coherent.

When development stretches over long periods, as was the case for chewBBACA 3, often forgetful developers (!) or new contributors will have difficulty understanding the code if it is not well documented. That is why I included *docstrings* in almost every function in chewBBACA 3 and added comments to code parts whose function may seem cryptic or when it is important to explain design choices. The code documentation made it easier to jump right into development and avoid adding redundant or code-breaking changes by better understanding what the code was doing and the parts that needed to be improved or expanded. From an end-user perspective, especially for non-technical users, usage documentation is more important than code documentation. A well-structured and well-written usage documentation helps users quickly understand how to start using software and explore its functionalities [247]. In that regard, I co-created usage documentation for chewBBACA 3 and Chewie-NS, with quick start guides, step-by-step tutorials, and detailed instructions about each module's functionalities or for the API endpoints and UI in the case of Chewie-NS. For command line software such as chewBBACA 3 it is also important to provide clear usage documentation and relevant information about each step being executed and any warnings or errors that might be raised [248]. I worked on adding clear and detailed help messages for chewBBACA's modules and on printing information about normal execution or warnings and errors to the command line to help users understand each step and help resolve any issues. However, I did not add the creation of a log file, which is an excellent way to save detailed information about the execution of the process for post-analysis and debugging. Implementing a log file should be a top priority for future versions of chewBBACA, and it is something that can be easily done using the built-in *logging*[2] module.

Providing clear instructions and multiple options for installation is also of utmost importance, as difficulties during software installation often discourage users from using it [249]. In this regard, making software available through a package management system, such as *conda*[3] (for bioinformatics software, the Bioconda distribution channel [250] is often used) or *PyPI*[4] for Python packages simplifies the installation process. chewBBACA 3 is available through both *conda* and *PyPI*, which allows users to install it quickly and painlessly, avoiding common installation issues, such as having to resolve conflicts between incompatible dependencies versions. Another common practice in bioinformatics is the use of containerization, provided by platforms such as Docker, to package and isolate software and their dependencies, ensuring portability and consistency of the results across platforms [251–254]. A Docker image for chewBBACA 3 is available[5] and the deployment and management of Chewie-NS instances is only possible through the orchestration of multiple Docker containers with Docker Compose[6].

[2] `https://docs.python.org/3/library/logging.html`
[3] `https://anaconda.org/anaconda/conda`
[4] `https://pypi.org/`
[5] `https://hub.docker.com/repository/docker/ummidock/chewbbaca/general`
[6] `https://docs.docker.com/compose/`

Workflow managers, such as Nextflow[7] and Snakemake[8], further simplify software usage by helping users set up and automate tasks [255, 256]. Using a workflow manager to streamline analyses with chewBBACA 3 could further simplify and promote its usage, and while it is something I did not have the opportunity to work on, it should be a future priority.

Another practice that promotes reproducibility and sustainability is software testing, as regularly testing the functionality of software, especially if automated, can help assess software performance and detect unexpected issues [257]. Testing is a critical, but underused, aspect of scientific software development that I and others have advocated in a publication [258]. I implemented automated standard functionality testing for chewBBACA 3 to ensure that it produces the expected results and detect issues arising due to minor changes or the introduction of new functionalities.

Ultimately, I think that learning and upholding good practices for sustained software development is crucial to developing quality software and maximizing its applicability in the long-term, as well as being a valuable skill for any developer. For Pythonistas like me, it is worth it to once in a while open a Python console and type *import this*[9].

5.2 Limitations and further improvements to CDS prediction in wg/cgMLST

Accurate CDS prediction is of the utmost importance in wg/cgMLST. This is generally done in one of two ways: by aligning reference alleles to genome assemblies to identify similar alleles based on parameters such as sequence identity and coverage [130]; or through *ab initio* or model-based CDS prediction tools that scan genome assemblies to identify ORFs that are subsequently filtered to identify putative CDSs [259]. Prodigal is a well-known *ab initio* prokaryotic CDS predictor that chewBBACA used up to version 3.3.0 [143, 159]. However, Prodigal has not been actively maintained, raising concerns about its long-term support. As part of the work developed in **Chapter 2** to improve chewBBACA, I substituted Prodigal by Pyrodigal [158]. Pyrodigal provides Python bindings to Prodigal, improving the performance and offering greater control over the CDS prediction process, while also ensuring results consistent with Prodigal's latest stable version[10]. **Chapter 2** evaluated the performance of chewBBACA only at the module level. Benchmarking the key steps in each of chewBBACA's modules, especially if varying the number of Central Processing Unit (CPU) cores used to better assess scalability, would have provided valuable information to understand which steps were most optimized and which still have room for improvement. Although the performance of the CDS prediction step was not directly measured, it is safe to assume that

[7] `https://www.nextflow.io/`

[8] `https://snakemake.readthedocs.io/en/stable/`

[9] `https://peps.python.org/pep-0020/`

[10] `https://github.com/hyattpd/Prodigal/releases/tag/v2.6.3`

Pyrodigal, which was optimized to reduce memory usage and runtime compared to Prodigal [158], contributed to the efficiency of chewBBACA 3 compared to its previous versions.

Although efficiency gains are definitely important, the true value of switching to Pyrodigal lies in having greater control over the CDS prediction process. With Prodigal, CDS prediction was only possible by calling Prodigal directly. In contrast, Pyrodigal provides Python bindings that enable finer control of the CDS prediction parameters. This allowed to seamlessly integrate CDS prediction into chewBBACA 3 and maintain results consistent with previous chewBBACA versions. In addition, this can be further explored to implement changes that can significantly improve the CDS prediction process in future versions of chewBBACA 3. Possible improvements include selecting from multiple file formats for the output file containing the CDS prediction results, such as GenBank and translated CDSs, which currently can only be written in FASTA format. Furthermore, the scores attributed to each CDS can be accessed before writing the results, which, through a thorough validation process, may allow the definition of a score threshold that can be used to exclude spurious CDSs predicted by Pyrodigal, leading to improvements in allele calling and a reduction in the number of spurious loci included in the schema seeds created by the *CreateSchema* module.

The prediction of spurious CDSs is the main disadvantage of using a *ab initio* or model-based CDS predictor. The predicted CDSs may not match any known gene, making validation difficult, or there may be a consistent deviation from reference alleles that have been experimentally validated, such as when the predicted CDSs are in-frame but shorter than the reference alleles or are out of frame compared to reference alleles [259]. This may contrast with alignment-based CDS prediction, used in wg/cgMLST platforms such as BIGSdb, which may predict less spurious CDSs due to the use of more conservative parameters, but will have a limited capacity to identify more divergent alleles or CDSs corresponding to genes not in the reference database. The discrepancies between both CDS prediction strategies hinder reproducibility. wg/cgMLST schemas and results generated with chewBBACA 3 will not be directly comparable and may lead to different conclusions than the results generated on a platform that uses an alignment-based CDS predictor. Adding alignment-based CDS prediction to chewBBACA 3 can help complement Pyrodigal results, while also allowing users to define parameters that approximate those used by other wg/cgMLST platforms, promoting interoperability and further adoption of chewBBACA 3. In fact, a combinatory approach is already used in some genome annotation pipelines, such as Prokka, Bakta, and PGAP, to improve CDS prediction and annotation [260–262]. Adopting a similar strategy in chewBBACA 3, especially if also defining a threshold based on the scores computed by Pyrodigal to filter out potential spurious CDSs, would definitely improve CDS prediction and, consequently, the accuracy of the results. Validating the results produced by this strategy would require a thorough evaluation with reference datasets for both experimentally validated and spurious CDSs, which could prove to be a laborious process, but would certainly help establish a framework for more accurate CDS prediction, especially in wg/cgMLST.

5.3 The assumption of dataset redundancy for large-scale wg/cgMLST and its potential limitations

The average pairwise ANI values of the species datasets used to evaluate chewBBACA 3 in **Chapter 2** reflect the composition of the public databases from which the genomes were sourced [175, 229]. A more uniform database composition does not necessarily correlate with global species diversity, as biases introduced by experimental protocols and uneven sampling, which favors specific bacterial strains, such as clinically relevant strains, may narrow our view on species diversity [263–266]. Nonetheless, the species definitions in use correlate well with an intra-species ANI$\geq$ 95% [267–272], indicating that strains of the same species have very similar genomes and, as reported in **Chapter 2** when identifying the distinct set of CDSs for each dataset (see Table 2.7), may share a high proportion of identical CDSs. As demonstrated in **Chapter 2**, identifying the distinct set of CDSs shared by a group of strains by equality comparisons can significantly simplify allele identification and strain classification in wg/cgMLST compared to approaches that rely more heavily on alignment.

The first published version of chewBBACA [143] already compared the CDSs predicted from the input genomes with the schema without using alignment, but would do this for every single genome, even if the same CDS was shared by all genomes. To further optimize the exact matching process, I decided to implement a deduplication step after CDS prediction. The deduplication identifies the set of distinct CDSs, returning a hash table with CDS SHA-256 hashes as keys and the list of genomes containing each CDS as values. The SHA-256 hashes are used as unique identifiers for exact matching between the predicted CDSs and the schema allele hashes, and are shorter than the actual CDS sequences, which reduces memory usage compared to using nucleotide or protein sequences as keys. By associating the list of genomes with a particular CDS to each distinct CDS SHA-256 hash it is possible to identify and classify all genomes with a particular CDS, avoiding redundant comparisons. An issue that I identified while testing CDS deduplication with large datasets was that the lists of genome string identifiers associated to each key could contain many values and occupy a lot of memory, which meant that storing the hash table in memory could become impractical for systems with less than 16GB of Random Access Memory (RAM). Substituting genome string identifiers with integer identifiers and using lists of integer identifiers as values reduced memory usage, but it was not enough to drop memory usage to a level at which most users could perform large-scale analyses. To further reduce memory usage, I implemented modified polyline encoding (inspired by the numcompress[11] package) to compress the lists of integers, producing strings that occupy less memory than the original lists, and allowing large-scale analyses with the memory available in most modern laptops, as reported in **Chapter 2**. Although implementing a solution for exact matching that integrates sequence deduplication, sequence hashing, and modified polyline encoding is more complex than simply comparing the CDSs sequences identified in each genome against a wg/cgMLST schema, the gains in

[11] `https://github.com/amit1rrr/numcompress`

efficiency far outweigh the increase in complexity since it allows chewBBACA to be used at a scale that was not previously possible in a reasonable time frame.

A possible limitation of the strategy used for exact matching in chewBBACA 3 is that it assumes that the genomes in a dataset, especially as the size of the dataset increases, share a high proportion of identical CDSs. This was observed for the datasets used in **Chapter 2** and **Chapter 4**, and may be broadly generalizable when working with single-species datasets. However, at higher taxonomic ranks, such as at the genus level, sequence diversity may be considerably higher, increasing the size of the hash table and, consequently, memory usage. chewBBACA's performance at the genus level has not been tested in the results presented in this thesis, but there have been applications at that level, as evidenced by the *Brucella*[12] [273] and *Shewanella*[13] schemas available on Chewie-NS. Other wg/cgMLST platforms, such as BIGSdb[14], BIGSdb-Pasteur[15] and EnteroBase[16] also store and manage schemas for genus-level typing, highlighting the interest and relevance of creating schemas for high-resolution typing at the genus level, especially when multiple species within a genus display pathogenic potential [273–276]. Genus-level wg/cgMLST schemas may be more broadly applicable than single-species schemas, providing high-resolution typing and a framework to study the inter-species loci diversity and gene flow, helping to elucidate key aspects in bacterial evolution and processes such as pathogenesis and virulence [277, 278]. However, defining a set of target loci for genus-level applications may only be possible for species with sufficient genomic overlap. For genus that encompass very diverse species, it may only be possible to define a useful set of target loci for some of the species within the genus, if at all. Applications above the genus level are not envisioned, mainly because the degree of sequence divergence between taxa at higher taxonomic ranks may compromise the operational definition of loci used in wg/cgMLST.

Another potential limitation is related to taxonomic underrepresentation in sequence databases, which is caused by both technical and non-technical factors. From a technical standpoint, sequencing and assembling the genome of certain bacterial species is more challenging due to limitations of sequencing technologies, difficulty or inability to culture, higher frequency of repetitive elements, and limitations of genome assembly software, among others. From a non-technical perspective, sequencing capacity can vary enormously between countries and tends to be used in studies or interventions deemed more relevant, such as sequencing bacterial species with a significant clinical or economic impact, leading to an incomplete representation of the diversity of bacterial species [279]. The application of methodologies that allow for less selective processing of biological samples, such as metagenomics, can help capture species diversity more accurately, promising a future in which reference databases are less biased. However, the impact of such approaches will not be immediate, nor are these ap-

[12] `https://chewbbaca.online/species/11/schemas/1`
[13] `https://chewbbaca.online/species/17/schemas/1`
[14] `https://pubmlst.org/`
[15] `https://bigsdb.pasteur.fr/`
[16] `https://enterobase.warwick.ac.uk/`

proaches without challenges. For example, metagenomics may offer limited capacity to detect low-abundance bacterial species in samples with an overabundance of host DNA [280–284]. However, technical improvements are promoting the broader adoption of metagenomics in multiple areas, such as diagnostics, the detection of novel pathogens, and microbiome studies [285]. This will inevitably reveal more of the diversity of bacterial species, leading to an increase in allele diversity and, consequently, an increase in memory usage for comprehensive wg/cgMLST. The implementation of chewBBACA 3, presented in **Chapter 2**, demonstrates that it is possible to perform large-scale wg/cgMLST with less than 8GB of RAM in a reasonable time frame. Since most modern laptops include 8GB or more RAM, using a laptop for large-scale wg/cgMLST for any bacterial species is already a possibility, either by processing datasets representing the full diversity of the species or by dividing huge datasets into smaller datasets for species that are disproportionally represented in databases, such as *Salmonella enterica*[17]. In addition, any concerns about increased memory usage may be minimized in the near future as the computational resources available to the general public become more and more powerful.

For large-scale wg/cgMLST in systems with more modest computational resources, such as a laptop, storage capacity can also be a bottleneck. The input files in FASTA format and the intermediate data generated by chewBBACA 3 while processing large datasets (i.e., tens of thousands of genomes) may require a lot of storage space. The storage space required to process the genome collections of the most represented species in public databases such as the NCBI may far exceed the storage capacity of most laptops. The storage and management of the huge amount of biological data that is generated has become a major challenge, representing significant costs and promoting research into more efficient data compression methods to reduce these costs [286–288]. Sequencing reads in FASTQ format and genome assemblies in FASTA format are usually compressed with *gzip*[18] to reduce the storage space used. Contrary to many bioinformatics software which accept compressed input files, chewBBACA 3 does not. Adding support for compressed input files and reducing the number and size of intermediate files created by chewBBACA 3 would allow more users to store and process larger datasets. General purpose compression algorithms are a convenient solution for the compression of biological data that is already included in many Operating System (OS) or is easy to install. However, as more and more sequencing data is generated, the compression ratio provided by compression tools such as *gzip* is insufficient, and it is necessary to use more efficient data compression methods specifically designed to compress biological data, such as AGC and MiniPhy [178, 289]. To greatly reduce costs associated with data storage and enable efficient and granular searches over terabyte- or petabyte-scale datasets, institutions managing reference databases may have no choice but to switch and promote the adoption of much more efficient compression algorithms specifically designed for biological data.

[17] `https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=28901`
[18] `https://www.gzip.org/`

5.4 The importance and potential impact of combinatory approaches for the accuracy of wg/cgMLST

The choice of methods for sequence comparison can have a great impact on the accuracy of bioinformatics analysis. Although alignment-based approaches remain the reference for sequence comparison, they can have some limitations [290]. For example, they tend to scale poorly for large datasets and provide less accurate results when comparing more divergent sequences with lower sequence identity due to the accumulation of point mutations and small indels or sequence rearrangements caused by recombination or Mobile Genetic Elements (MGEs) [291]. In the specific case of BLAST [155, 190] applied to CDS prediction and allele identification, local alignment extension can stop due to significant sequence divergence, resulting in a single or multiple local alignments that, while having a high score, do not meet the identity and coverage criteria enforced by wg/cgMLST allele callers. This may result in the inability to identify loci that are, in fact, present in genomes.

In the case of chewBBACA 3, which only considers the highest scoring alignment reported by BLASTp, the BSR calculated from the best alignment may not reach the defined BSR threshold simply because the reported alignment does not cover the majority of sequences that, apart from a single or a few dissimilar regions that interfere with alignment extension, are highly similar. In theory, combining multiple alignments may yield a BSR that reaches the threshold, but there is no guarantee that BLAST will report alignments between all similar regions. Moreover, BLAST computes local alignments, and it seems reasonable that it does not provide any option to combine multiple local alignments to obtain a combined score. Initially, and perhaps naively, I thought it may be possible to implement a simple script to compute the combined score. But lack of knowledge about how the alignment raw score is computed, especially with compositional adjustment [292, 293], made it seem like I was trying to probe a black box. Thus, it appeared reasonable to consider other options.

At first glance, one of the options could be to identify pairs of sequences for which BLASTp reports multiple alignments that cover a significant fraction of the sequences and to compute global alignments to get a better estimate of global sequence similarity. Although a global alignment would cover the full length of the sequences being compared, two factors contributed to this option being ruled out nearly immediately. Firstly, global alignment yields different results than local alignment and I would have to use a different aligner than BLAST. This would lead to completely different results, invalidating the BSR-based approach used in chewBBACA, and could potentially diverge greatly from the results determined with BLAST, which would warrant a complete overhaul of chewBBACA's algorithm, both conceptually and algorithmically. Secondly, global alignment is computationally expensive and is not adequate to align large collections of potentially divergent sequences. Since I intended to implement a new version of chewBBACA tailored for scalable and efficient analysis of large genome collections, I decided instead to complement the alignment-based strategy used by

chewBBACA with alignment-free methods, more specifically *k*-mer-based methods [290, 294]. The concept of *k*-mers, although seemingly simple, offers extreme versatility by simply varying the value of *k*, which defines the size of the *k*-mers, and choosing a sampling method to select subsets of *k*-mers from a sequence [137]. Research into the optimization of *k*-mer-based methods has produced concepts such as spaced seeds [295, 296], minimizers [132, 134, 149, 162–164], syncmers [297, 298], and more recently strobemers [133, 135, 299]. These concepts are applicable to virtually any step of the bioinformatics workflow, contributing to the efficiency and accuracy of methods for read mapping [300–303], sequence alignment [155, 304, 305], taxonomic classification [140, 306–310], and metagenomic binning [311–313], among others [299]. Unsurprisingly, there are also examples of the application of alignment-free methods in wg/cgMLST, such as using KMA [305] to perform cgMLST directly from sequencing reads[19] and the EToKi[20] suite of methods used by Enterobase [151, 314]. Alignment-free methods have become undeniably ubiquitous in bioinformatics, either because they allow for more efficient and scalable analysis or because they overcome some of the limitations of alignment-based methods [290, 315].

I chose minimizer-based clustering to complement the alignment-based strategy used by chewBBACA because minimizers are simple to implement, fast to determine, and enabled me to implement a clustering method that was accurate enough for the application I had in mind. Performing intra-cluster alignment to classify the CDSs is much faster than all-vs-all or pairwise alignment between the CDSs identified in all inputs. Minimizer-based clustering also enables the identification of highly similar sequences with differences concentrated in a single or a few sequence regions because it counts the number of shared minimizers along the complete sequences, whereas BLAST may only report smaller local alignments, making it difficult to determine global sequence identity. The strength of the strategy lies then in combining alignment-based and alignment-free methods to improve speed and accuracy, an approach that has been increasingly used in bioinformatics. Although minimizer-based clustering enabled considerably faster and slightly more accurate wg/cgMLST with chewBBACA 3, as reported in **Chapter 2**, it is not sufficiently sensitive to identify some of the most divergent alleles within the $0.6 \geq BSR > 0.7$ interval. Due to this limitation, the identification of the most divergent alleles, performed by the default execution mode, mode 4, uses only BLASTp, which can increase the runtime with larger and more diverse datasets. An alternative approach would be to accept a slight reduction in sensitivity and use minimizer-based clustering anyway, since the number of more divergent alleles not identified would probably be small, as evidenced by the small differences between chewBBACA execution modes 3 and 4 (see Figures 2.19 and 2.20). However, since I did not test this hypothesis with datasets of other species, I decided to opt for the more conservative approach. The limitation of minimizer-based clustering in the identification of divergent alleles may be surpassed by the implementation of spaced-minimizer- or strobemer-based clustering. Spaced-minimizers and strobemers allow for sequence variation between matching regions, which can better

[19] `https://github.com/BCCDC-PHL/kma-cgmlst`

[20] `https://github.com/zheminzhou/EToKi`

accommodate the differences identified in more divergent alleles [133, 135]. If successful, this approach could further reduce chewBBACA's runtime and improve accuracy.

To further increase the accuracy of schema creation and allele calling, creating graphs representing intra- and inter-cluster connections between similar CDSs could enable the identification of connected components matching groups of similar CDSs. These connected components would facilitate the identification of alleles belonging to the same locus, including divergent ones not directly linked but sharing a BSR ≥ 0.6 with other alleles in the same connected component, and improve the detection of paralogous loci and pseudogenes based on sequence similarity and genomic context analysis. Furthermore, including the frequency of each CDS in the input genomes as a node attribute could help identify and select the most frequent CDSs as loci representatives during schema creation to better capture loci diversity. This contrasts with the current strategy used in chewBBACA 3, which selects a single representative allele, limiting the identification of new alleles based on a maximum sequence divergence from the single representative allele defined by the BSR and the size threshold parameters. The resolution of the wg/cgMLST analysis performed by chewBBACA 3 can also be expanded by performing Synteny Network Analysis (SNA) based on the CDS coordinate data determined by chewBBACA 3, a resource that has remained largely unused. Loci adjacency can be represented as a graph, allowing to identify paths corresponding to groups of loci that are frequently conserved in order and orientation, called Synteny Blocks (SBs). Variations in the composition of similar SBs can help identify structural changes caused by events such as Homologous Recombination (HR) and HGT [316, 317]. Comparison of strains based on the composition and presence-absence analysis of SBs can potentially improve the estimation of relatedness in wg/cgMLST and help identify groups of loci correlated with relevant phenotypes associated with increased virulence and AMR. The results of the SNA could be included in the report generated by the *AlleleCallEvaluator* module, enabling users to explore the composition of identified SBs and compare bacterial strains based on synteny.

5.5 Providing functionalities for comprehensive and user-centered wg/cgMLST data analysis is essential to fully leverage the potential of wg/cgMLST

Improving the accuracy of wg/cgMLST schema creation and allele calling is crucial to provide reliable results. However, assuming that minimum requirements for accuracy have been met, end-users are usually more interested in downstream analyses, such as the computation of distance matrices and MSTs to identify closely related strains, and the attribution of STs based on pre-defined nomenclatures. Those analyses can provide valuable information in surveillance and outbreak detection settings, and there are several tools and platforms that enable them [235, 236, 318–320]. However, wg/cgMLST schemas and results

hold the potential for much more detailed analyses that go beyond traditional applications in surveillance and outbreak detection. For example, wg/cgMLST was used to evaluate the genomic diversity of an emerging human pathogen and characterize candidate antigens for vaccine development [183, 321, 322].

I developed chewBBACA's *SchemaEvaluator* and *AlleleCallEvaluator* modules to allow users to explore wg/cgMLST schemas and results intuitively and comprehensively. The information and functionalities provided in the reports generated by both modules allow users to evaluate loci diversity and assess strain similarity in great detail at the dataset level, and up to the species level if analyzing wg/cgMLST schemas and datasets that capture the full diversity of the species. In addition, reports are generated locally to enable more scalable and shareable analyzes compared to centralized platforms with limited resources and to avoid issues related to data privacy concerns. Although the reports already include functionalities for a detailed analysis of wg/cgMLST results, I think they can be further expanded to improve their applicability.

While the *SchemaEvaluator* module has been used to evaluate schemas not generated with chewBBACA, such as schemas available in BIGSdb and Enterobase [129, 151], this option was not extensively tested and it remains unclear if the module can be used to evaluate any schema regardless of provenance. The *AlleleCallEvaluator* module only accepts results generated with chewBBACA, limiting its applicability. Changing both modules to make them compatible with the most widely used schema and allele call results data formats would definitely benefit most users. Additionally, the allele call results can be utilized more effectively by developing new functionalities that enable the computation of loci and sample sets based on loci presence thresholds directly in the report and identify lists of potentially spurious loci and low-quality genomes based on the percentage of missing data.

The special classifications assigned by chewBBACA 3 (see Figures 2.5-2.8) are usually ignored or discarded while analyzing allele calling results. This is a reasonable approach because the ambiguous nature of these classifications does not allow to infer loci presence confidently. Inferring loci presence based on these classifications can reduce the accuracy of common downstream analysis, such as distance estimation and core genome computation. However, simply treating these classifications as missing data also undervalues their potential. CDSs assigned special classifications may correspond to valid alleles of biological relevance, such as more divergent alleles, alleles for new loci that are not in the schema and alleles of paralogous genes. Special classifications are also assigned to spurious CDSs more frequently, with higher frequencies of special classifications being a sign of lower-quality data. Including a more detailed analysis of the CDSs assigned special classifications in chewBBACA's reports, including their genomic context, could provide valuable information to identify new alleles, paralogous genes, pseudogenes, and spurious CDSs resulting from misassembly and frameshift mutations.

There are also features that, while perhaps more challenging to implement, would extend

the report's functionalities to the point where most users would not need to resort to other software for further insight. For example, determining the variable positions or SNPs per locus would tap into the resolution level of SNP-based approaches. New features for sample clustering and ST assignment, which are commonly performed by users using other tools, based on a *ad hoc* or pre-defined nomenclature have been requested by chewBBACA users, and could probably be incorporated into chewBBACA by integrating the results generated by ReporTree [319]. The report generated by the *SchemaEvaluator* module provides detailed information about the number of alleles and allele size per locus, as well as a MSA component to compare translated alleles and identify sequence differences, but it does not provide a more general measure of similarity at the intra- and inter-locus level. To that end, representations of intra- and inter-locus similarity based on the BSR could help users quickly identify loci with greater variability and groups of similar loci. A dedicated page for each sample analyzed would be a valuable addition to the report generated by the *AlleleCallEvaluator* module. Each sample page would include summary classification statistics and a circular genome viewer, such as CGView.js[21] [323], which supports the representation of annotated loci features, providing an easy way to explore genome structure, identify loci of interest, and explore their genomic context.

chewBBACA 3, as many other bioinformatics tools, is a Command Line Interface (CLI) tool compatible only with Unix-like operating systems such as Linux. While running chewBBACA's modules is not very complex, it still requires users to be familiar with the command line. The lack of a General User Interface (GUI) is perhaps the greatest hurdle to the wide adoption of bioinformatics tools by users without bioinformatics training. Implementing a GUI and making chewBBACA compatible with Windows would definitely encourage more researchers to perform wg/cgMLST and explore the functionalities provided by chewBBACA.

Chewie-NS, presented in **Chapter 3**, was created to be less restrictive in terms of data sharing policies compared to centralized platforms such as BIGSdb. In that regard, I think Chewie-NS succeeds in providing uncomplicated access to wg/cgMLST schemas that can be used for local and private analysis based on a common allelic nomenclature. However, Chewie-NS lacks functionalities for a comprehensive analysis of schema structure and loci diversity compared to other well-known wg/cgMLST platforms. Although the schema and loci pages in Chewie-NS served as templates for the initial versions of the reports created by chewBBACA's *SchemaEvaluator* and *AlleleCallEvaluator* modules, the pages in Chewie-NS have become outdated compared to chewBBACA's reports. Updating the schema and loci pages in Chewie-NS to at least match the functionalities provided by the chewBBACA reports would benefit the users browsing the schemas deposited in Chewie-NS, essentially providing a catalog of the loci identified in bacterial species and their diversity. The decentralization of wg/cgMLST analysis is the main strength of Chewie-NS. However, it can also be a major disadvantage if users do not have the expertise to perform local wg/cgMLST and request or send data through Chewie-NS' API. While it is not the objective of Chewie-NS to centralize

[21] `https://js.cgview.ca/index.html`

wg/cgMLST analysis, I think implementing pages for the submission of data and the analysis of results would allow more users, especially those without bioinformatics training, to use the schemas deposited in Chewie-NS. These operations would be optional, with a fully decentralized workflow still possible. Furthermore, to promote interoperability and the applicability of the allelic nomenclatures managed by Chewie-NS, it would be beneficial to store allelic profiles and allow users to compare their results against stored profiles based on Chewie-NS' nomenclature and the nomenclatures implemented by other wg/cgMLST platforms. Since data submission would be optional, the database of allelic profiles would be populated by periodically downloading high-quality genomes from public databases such as the NCBI and performing allele calling in Chewie-NS, providing a strong basis for users to contextualize and compare their strains.

5.6 The unrealized potential of wgMLST

I have frequently referred to wg/cgMLST, but in reality the majority of analyses is performed at the cgMLST level and wg/cgMLST platforms store mostly cgMLST schemas[22] [130, 324]. Working at the cgMLST level provides robust results for most applications, such as surveillance, and avoids having to deal with the issue of missing data when considering a greater number of loci, such as when working at the wgMLST level [127, 146]. When greater resolution is needed, such as in outbreak investigation, cgMLST results can be complemented with other methods, such as SNP-based approaches using a closely related reference genome or wgMLST [119, 127, 128]. The process of creating a cgMLST schema is relatively simple, especially compared to the creation of a robust wgMLST schema, such as the one created for *S. pyogenes* in **Chapter 4**. A simple workflow for cgMLST schema creation that takes advantage of chewBBACA's functionalities consists in selecting a dataset of high-quality closed genomes for a bacterial species to create a schema with the *CreateSchema* module, followed by allele calling with a more diverse dataset and core-genome determination with the *ExtractCgMLST* module. This process has been applied many times to create cgMLST schemas for high-resolution typing [325–333]. Identifying and removing spurious and paralogous loci may further increase the quality of cgMLST schemas, but these issues tend to be few if the genomes used for schema creation are carefully selected. Scaling to wgMLST is more complex, as it incorporates the accessory genome, which can be highly variable, in part due to the effect of recombination and MGEs [334–336].

The variability of the accessory genome represents a challenge in the creation of wgMLST schemas. Incorporating the accessory genome leads to an increase in the number of spurious loci in wgMLST schemas due to a higher frequency of pseudogenes, gene fusions, paralogous genes, repetitive elements, and genes whose allele diversity can only be fully captured by

[22]The cgMLST schemas used with the commercial software Ridom SeqSphere+ are available at `https://www.cgmlst.org/ncs` and an instance of the BIGSdb platform managed by the Institut Pasteur that stores cgMLST schemas for multiple *taxa* is available at `https://bigsdb.pasteur.fr/`

defining custom parameters, rather than using schema-wide parameters. This diversity of features blurs the limits of loci and allele definitions. To make matters worse, features such as repetitive elements complicate the genome assembly process, potentially leading to a higher number of sequence assembly errors, which introduces uncertainty about the validity of the identified features [337, 338]. These issues were encountered when creating the wgMLST schema for *S. pyogenes*, as described in **Chapter 4**. The initial wgMLST schema, containing 3,318 distinct loci, identified from a dataset of high-quality 208 complete genomes, was curated to identify and correct issues, resulting in a final schema with 3,044 loci. The reduction in the number of loci was not due to simply excluding loci from the schema. It was the result of a laborious curation process guided by an expert in *S. pyogenes* to identify spurious schema loci corresponding to pseudogenes, gene fusions, and paralogous loci based on the functional annotation of all loci and on the inspection of the sequences and genomic context of hundreds of schema loci. The final curation process consisted of the substitution of schema loci that matched pseudogenes and gene fusions by valid alleles, merging schema loci corresponding to the same gene, and removing schema loci only when it was not possible to identify valid alternative alleles. Removing groups of paralogous loci identified by chewBBACA and loci mostly assigned special classifications is a simpler alternative to the laborious curation process used to create the *S. pyogenes* wgMLST schema, but it potentially removes a lot of loci that can be corrected through a more careful curation process. A detailed curation process maximizes the quality and loci diversity accurately captured by a wgMLST schema, but the requirements of such a process are a great limitation to developing comprehensive and reliable wgMLST schemas. Moreover, some studies suggest that wgMLST does not constitute a very significant improvement in terms of discriminatory power over cgMLST [126, 146, 150, 339–343]. So why even bother creating wgMLST schemas?

Regarding the requirements for creating wgMLST schemas, I consider that we are simply missing the tools to create high-quality wgMLST schemas. In fact, I started a project called Schema Refinery[23] with the objective of creating a set of tools to help users perform common steps of wg/cgMLST schema creation. Initially, I included Python scripts to download and select genome assemblies for schema creation and allele calling, annotate schema loci based on multiple sources, and perform basic operations to refine schemas such as merging, splitting, and excluding loci. Schema Refinery has been under active development, although the bulk of the development has been passed to other members of the lab, while I participate in the discussions about implementation design, test the functionalities, and occasionally contribute with code changes. Since the old adage of Garbage In, Garbage Out (GIGO) is also valid for wg/cgMLST, Schema Refinery enables the download and selection of high-quality genomes for schema creation and allele calling, which will greatly reduce the number of issues related to spurious loci identified in low-quality genome assemblies. Furthermore, Schema Refinery includes a module to refine and expand wg/cgMLST schemas. This module allows users to identify and resolve issues such as those identified during the creation of

[23] `https://github.com/B-UMMI/Schema_Refinery`

the *S. pyogenes* wgMLST schema automatically, and will provide the option to identify new loci from chewBBACA's allele calling results to add to the schemas. The possibility of automatically refining schemas will greatly simplify the creation of high-quality wgMLST schemas, and the option of adding new loci has the potential to be a future-proof solution to update schemas as we reveal more of the diversity of bacterial species.

Concerning the observation that wgMLST may not constitute a considerable improvement over cgMLST, I think such a conclusion can only be drawn from experiments whose objectives do not fully explore the potential of wgMLST. Firstly, wgMLST schemas can contain a lot more loci than cgMLST schemas, often more than double the loci, capturing much more of the diversity of bacterial species. wgMLST and cgMLST results can be strongly correlated, leading to similar observations and conclusions when comparing, for example, tree topologies. However, it is important to compare the resolution provided by both approaches in terms of the number of loci being compared and the differences identified between bacterial strains. A cgMLST schema that contains half or less than the average number of loci in the genomes of a bacterial species cannot possibly provide the same resolution as a wgMLST schema that includes nearly every locus ever identified in the genomes of the same species. Thus, an apparent equivalence of both approaches is only true when the study objectives limit the analysis to a lower resolution level than what wgMLST can provide. It is important to establish a clear distinction between discriminatory power and resolution. For a specific dataset, both approaches can identify the same clusters of closely related strains, providing equivalent discriminatory power for clustering. However, this is highly dependent on the dataset being analyzed and on the parameters used for cluster definition. The higher resolution provided by wgMLST allows to compare and explore the diversity of more loci than cgMLST and offers greater discriminatory power when differences are more concentrated on the accessory genome. Multiple studies have compared cgMLST and wgMLST, demonstrating that wgMLST provides further resolution when compared with the cgMLST approaches routinely used for surveillance and outbreak investigation [119, 126, 128, 343]. In addition, a detailed analysis of the loci and population diversity of a species, such as what is performed in pangenome analysis, is only possible at the wgMLST level, as restricting the analysis to the cgMLST level would under-evaluate the genetic diversity of the species [335]. In fact, a wgMLST schema that captures a species diversity could also be called a pan-genome MLST (pgMLST) schema, and could, in theory, allow for pangenome analysis with chewBBACA 3 by analyzing a dataset representative of the diversity of the species with the *AlleleCall* module and generating reports with the the *SchemaEvaluator* and *AlleleCallEvaluator* modules. It is also important to note that genes determinant for relevant phenotypic characteristics, such as virulence and antibiotic resistance, are often part of the accessory genome and may not be identified by cgMLST [344–346]. The increased resolution of wgMLST can be especially valuable in the context of surveillance and outbreak investigation, as demonstrated by the identification of the *S. pyogenes* M1 lineages in **Chapter 4**. In that case, the analysis was performed by focusing on the core genes shared by the lineages' strains, but since the schema used was a wgMLST

schema it allowed to tune up the core-genome based on the dataset, which would not be possible with a more strict cgMLST schema. The strategy of using a wgMLST schema to scale the size of the core-genome based on the dataset under analysis can enrich surveillance and outbreak detection analyzes, but it has not been adopted as routine. Furthermore, the increased resolution resulting from the transition to the wgMLST schemas would invalidate the distance thresholds for outbreak definition currently used. Thus, transitioning to routine wgMLST would require a comprehensive analysis to identify congruence points and establish equivalent distance values, ideally by defining flexible thresholds that better accommodate the differences between the surveillance systems implemented by public health institutions[119].

Although using wgMLST schemas to dynamically define the core-genome for each dataset represents a step forward from using stricter cgMLST schemas, it still does not use all loci in wgMLST schemas, which can identify additional loci exclusive to strain subsets in the dataset under analysis. Fully transitioning to wgMLST is complex, in part due to uncertainty surrounding missing data. Missing data corresponds to loci that are not identified in the genomes, classified as absent (e.g. the *LNF* classification assigned by chewBBACA), or to loci for which the validity of the match found is somewhat ambiguous (e.g., the match is considerably shorter or longer than the matched locus, corresponding to the *ASM* and *ALM* classifications assigned by chewBBACA, respectively). Most of WGS data are available as sequencing reads in FASTQ format or, if assembled, as draft genome assemblies in FASTA format. The impossibility of having complete genomes introduces uncertainty about loci presence, as we cannot always distinguish between loci that are truly absent or appear absent due to issues introduced by sequencing and assembly errors. The loci presence threshold used to determine the core genome from wg/cgMLST data is usually set to 99% or 95%, instead of 100%, precisely to accommodate issues such as these that cause a drop in the frequency of highly frequent loci. The transition to wgMLST is also hampered by the methods and parameters used for loci identification. Defining simple parameters for loci identification, such as 80% sequence identify and coverage, or a BSR ≥ 0.6 as used by chewBBACA 3, represents an oversimplification of locus diversity. A single or a few parameter values do not allow for perfect loci identification. Perfect loci identification, if possible at all, may require the definition of finely tuned parameter values for each locus. Some wg/cgMLST platforms, such as BIGSdb allow to define locus-specific parameter values. However, we cannot determine the optimal set of parameters for every locus. Thus, I think that, in the meantime, we will have to continue to adjust our generalizations. Uncertainty and doubt remain, both important in science.

5.7 Perspectives on the future of wg/cgMLST

Bioinformatics, as an interdisciplinary field, has greatly benefited from the application of concepts and methods derived from other scientific fields. The achievement of technological milestones and growing interest are propelling bioinformatics into a flourishing age and to the

center stage of scientific research. The performance of computing hardware has been steadily improving and has reached a point where it is possible to generate, store, and analyze huge amounts of biological data. More widespread access to powerful computational resources allows more researchers to apply existing or newly developed methods to tackle challenging problems that were once off limits due to technological and technical constraints. Several public health emergencies, such as the COVID-19 pandemic, caused by the SARS-CoV-2 virus, and mortality caused by major bacterial pathogens, especially in low- and middle-income countries, have also raised the public and governmental organizations' awareness of the importance of research and preventive measures [3].

The boom in Graphics Processing Unit (GPU) computing, in conjunction with the growing interest and advances in Artificial Intelligence (AI), will definitely continue to accelerate biomedical research and expand the realm of research possibilities. A good example of a recent and monumental breakthrough due to the application of GPU computing and AI is AlphaFold [347, 348], which has greatly improved our ability to accurately predict protein structures. The true impact of AlphaFold will probably only be fully realized in the coming years. AI has already been applied to prokaryotic gene prediction [161], a crucial aspect for accurate wg/cgMLST. A more accurate prediction of protein structure and function can also help validate gene prediction data, contributing to improvements in gene prediction and wg/cgMLST schema refinement.

When talking about the future of wg/cgMLST, one cannot simply forget to mention the relatively recent improvements to DNA sequencing technologies that have enabled WGS and consequently wg/cgMLST. Further improvements to these technologies are expected. Any improvements, especially in the accuracy of long-read sequencing, will increase the contiguity of assembled genomes and perhaps even make accurate wg/cgMLST from sequencing reads a reality. More contiguous genome assemblies should also lead to more accurate gene prediction. The quality of a genome assembly is influenced by multiple factors, including the methods used for sample preparation, the sequencing technology used, and the tools used to process the sequencing data and generate a genome assembly. Each step of the process and the combination of methods used throughout the process can introduce biases that negatively influence the end result [337]. These issues contribute differently to the degree of fragmentation of genome assemblies, often leading to a cumulative effect that affects a greater number of loci as more genomes are analyzed. In my view, improvements to DNA sequencing technologies and the standardization of other steps can lead to much more than a slight increase in the accuracy of wg/cgMLST. It can greatly reduce the number of absent or fragmented genes, facilitating the transition to wgMLST, and possibly revealing that we need to rethink the definitions of the core and accessory genome for many bacterial species.

I also consider that we are nearing a crossroads where we will have to rethink, or at least adapt, the methods used for wg/cgMLST. The Web platforms and other software for wg/cgMLST are responsible and are also a consequence of the success of wg/cgMLST. wg/cgMLST allows for a comprehensive analysis of loci diversity and high-resolution typing

in surveillance and outbreak scenarios, but even at its highest potential, it may lack some of the features of SNP-based and *k*-mer-based methods. In fact, wg/cgMLST results are often complemented by SNP analysis for finer resolution, such as in outbreak detection. The integration of *k*-mer-based methods can also increase the accuracy and speed of the analyzes, as achieved with chewBBACA 3 in **Chapter 2**. I think these different approaches will be increasingly combined to take advantage of the strengths of each one, with discussions such as allele-based vs. SNP-based becoming increasingly less relevant and priorities shifting towards the implementation of software that performs comprehensive and multifaceted analyzes. In addition, I think graph-based methods can emerge as an alternative that surpasses the accuracy of other approaches. Improvements in computing hardware and graph algorithms, as well as the availability of a large number of sequenced genomes, have made it possible to index and examine the variation of larger genome collections [349–351]. Pangenome graphs have been used to index and examine the variability of bacterial genomes, allowing the identification of population-level nucleotide and structural polymorphisms. Using graph data structures to index the variability of coding and non-coding regions of large collections of bacterial genomes may provide a framework that achieves results more accurate than the combination of wg/cgMLST and SNP-based approaches by allowing the analysis of non-coding regions, which cannot be done with wg/cgMLST, and surpassing the reference bias issue of SNP-based methods. All while also providing information about genome structure, which would allow to obtain information about genomic context through synteny analysis. Notwithstanding the potential of graph-based methods, the computational requirements for the construction of graphs that encompass the diversity of thousands of bacterial strains may be prohibitive for most users, preventing the applicability of graph-based methods compared to wg/cgMLST.

The comparability of the wg/cgMLST results at the national and international levels is essential for effective surveillance and outbreak detection. During the COVID-19 pandemic, the scientific community quickly adapted and established data analysis and sharing standards to promote easy sharing and comparability of results to track SARS-CoV-2 variants [352, 353]. In contrast, the stage of implementation of WGS-based surveillance systems for bacterial pathogens, such as for FWD pathogens, can differ significantly between countries [119]. The disparities between the surveillance systems implemented by different countries hinder the comparability of the results at the intersectoral and international levels, which affects outbreak detection and investigation, especially for multi-country outbreaks. A detailed congruence analysis of the results generated by eleven European institutes revealed that, while there may be general concordance between the results generated by similar surveillance systems, the results are not directly comparable [119]. Different surveillance systems provide different levels of resolution, and a congruence analysis is necessary to establish threshold equivalence between systems. Since the implementation of a surveillance system represents a significant investment, it is highly unlikely that any country will switch to a different approach in the short term, even if that would promote interoperability and provide more accurate results. A more immediate solution would be to determine congruence points between all the different systems and using flexible thresholds to accommodate for incompatibilities arising

due to the use of single value thresholds. However, this approach would require a significant collaboration effort between institutions of many different countries and the validity of the congruence system would have to be continuously assessed. In the long term, the systems used by different countries may converge toward methods that are provably better, promoting interoperability and approximating a global One Health surveillance system. In the meantime, countries will likely continue sharing data on an *ad hoc* basis or through centralized systems such as EFSA's One Health WGS system when a more concerted action is necessary to track and resolve multi-country outbreaks.

It is clear that the potential of the wg/cgMLST approach has not yet been fully realized, with room for improvement both conceptually and technically. The ongoing technological revolution also expands the realm of possibilities, encouraging researchers to apply methodologies that were once unfeasible from a technological standpoint or come up with revolutionary strategies arising from a shift in perspective. The adoption and evolution of wg/cgMLST is influenced by multiple factors that go beyond scientific objectiveness, making it difficult to predict the magnitude of its future role. However, future bacterial surveillance systems and population diversity studies, even if methodologically distinct from wg/cgMLST, will surely have to incorporate the advantageous properties of wg/cgMLST that distinguish it from other approaches currently used. The current scientific revolution will surely continue to provide a multitude of possibilities and opportunities that will approximate and blur the boundaries between scientific fields. The uncertainty of how it will all unfold is undoubtedly exciting.

Bibliography

[1] C R Woese, O Kandler, and M L Wheelis. “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.” In: *Proceedings of the National Academy of Sciences* 87.12 (1990). Publisher: Proceedings of the National Academy of Sciences, pp. 4576–4579. DOI: 10.1073/pnas.87.12.4576.

[2] Mohsen Naghavi et al. “Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050”. In: *The Lancet* 404.10459 (Sept. 28, 2024). Publisher: Elsevier, pp. 1199–1226. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(24)01867-1.

[3] Kevin S. Ikuta et al. “Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019”. In: *The Lancet* 400.10369 (Dec. 17, 2022). Publisher: Elsevier, pp. 2221–2248. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(22)02185-7.

[4] *WHO Bacterial Priority Pathogens List 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control Antimicrobial Resistance*. 1st ed. Geneva: World Health Organization, 2024. 1 p. ISBN: 978-92-4-009346-1.

[5] Wenjun Li, Didier Raoult, and Pierre-Edouard Fournier. “Bacterial strain typing in the genomic era”. In: *FEMS Microbiology Reviews* 33.5 (Sept. 1, 2009), pp. 892–916. ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2009.00182.x.

[6] Mike S. Son and Ronald K. Taylor. “Growth and Maintenance of Escherichia coli Laboratory Strains”. In: *Current protocols* 1.1 (Jan. 2021), e20. ISSN: 2691-1299. DOI: 10.1002/cpz1.20.

[7] Stephen V. Gordon and Tanya Parish. “Microbe Profile: Mycobacterium tuberculosis: Humanity’s deadly microbial foe: This article is part of the Microbe Profiles collection.” In: *Microbiology* 164.4 (Apr. 1, 2018), pp. 437–439. ISSN: 1350-0872, 1465-2080. DOI: 10.1099/mic.0.000601.

[8] Xavier Didelot et al. “Transforming clinical microbiology with bacterial genome sequencing”. In: *Nature Reviews Genetics* 13.9 (Sept. 2012), pp. 601–612. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3226.

[9] Antony Croxatto, Guy Prod'hom, and Gilbert Greub. "Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology". In: *FEMS Microbiology Reviews* 36.2 (Mar. 1, 2012), pp. 380–407. ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2011.00298.x.

[10] Peter Lasch et al. "A MALDI-ToF mass spectrometry database for identification and classification of highly pathogenic bacteria". In: *Scientific Data* 12.1 (Jan. 31, 2025). Publisher: Nature Publishing Group, p. 187. ISSN: 2052-4463. DOI: 10.1038/s41597-025-04504-z.

[11] Maryam Alizadeh et al. "MALDI-TOF Mass Spectroscopy Applications in Clinical Microbiology". In: *Advances in Pharmacological and Pharmaceutical Sciences* 2021.1 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/9928238, p. 9928238. ISSN: 2633-4690. DOI: 10.1155/2021/9928238.

[12] Piseth Seng et al. "Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry". In: *Clinical Infectious Diseases* 49.4 (Aug. 15, 2009), pp. 543–551. ISSN: 1058-4838. DOI: 10.1086/600885.

[13] E. A. Idelevich et al. "Rapid detection of antibiotic resistance by MALDI-TOF mass spectrometry using a novel direct-on-target microdroplet growth assay". In: *Clinical Microbiology and Infection* 24.7 (July 1, 2018), pp. 738–743. ISSN: 1198-743X. DOI: 10.1016/j.cmi.2017.10.016.

[14] Jack Hassall et al. "Limitations of current techniques in clinical antimicrobial resistance diagnosis: examples and future prospects". In: *npj Antimicrobials and Resistance* 2.1 (June 17, 2024). Publisher: Nature Publishing Group, p. 16. ISSN: 2731-8745. DOI: 10.1038/s44259-024-00033-8.

[15] Jessica E. Buddle and Robert P. Fagan. "Pathogenicity and virulence of Clostridioides difficile". In: *Virulence* 14.1 (Dec. 31, 2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21505594.2022.2150452, p. 2150452. ISSN: 2150-5594. DOI: 10.1080/21505594.2022.2150452.

[16] Gerard Lina et al. "Involvement of Panton-Valentine Leukocidin—Producing Staphylococcus aureus in Primary Skin Infections and Pneumonia". In: *Clinical Infectious Diseases* 29.5 (Nov. 1, 1999), pp. 1128–1132. ISSN: 1058-4838. DOI: 10.1086/313461.

[17] François Vandenesch et al. "Community-Acquired Methicillin-Resistant Staphylococcus aureus Carrying Panton-Valentine Leukocidin Genes: Worldwide Emergence - Volume 9, Number 8—August 2003 - Emerging Infectious Diseases journal - CDC". In: (). DOI: 10.3201/eid0908.030089.

[18] Anne Tristan et al. “Global Distribution of Panton-Valentine Leukocidin–positive Methicillin-resistant Staphylococcus aureus, 2006 - Volume 13, Number 4—April 2007 - Emerging Infectious Diseases journal - CDC”. In: (). DOI: 10.3201/eid1304.061316.

[19] Binh An Diep et al. “Contribution of Panton-Valentine Leukocidin in Community-Associated Methicillin-Resistant Staphylococcus aureus Pathogenesis”. In: *PLOS ONE* 3.9 (Sept. 12, 2008). Publisher: Public Library of Science, e3198. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0003198.

[20] Andie S. Lee et al. “Methicillin-resistant Staphylococcus aureus”. In: *Nature Reviews Disease Primers* 4.1 (May 31, 2018). Publisher: Nature Publishing Group, p. 18033. ISSN: 2056-676X. DOI: 10.1038/nrdp.2018.33.

[21] J Henrichsen. “Six newly recognized types of Streptococcus pneumoniae”. In: *Journal of Clinical Microbiology* 33.10 (Oct. 1995). Publisher: American Society for Microbiology, pp. 2759–2762. DOI: 10.1128/jcm.33.10.2759-2762.1995.

[22] D. Tarragó et al. “Identification of pneumococcal serotypes from culture-negative clinical specimens by novel real-time PCR”. In: *Clinical Microbiology and Infection* 14.9 (Sept. 1, 2008). Publisher: Elsevier, pp. 828–834. ISSN: 1198-743X. DOI: 10.1111/j.1469-0691.2008.02028.x.

[23] Catarina Silva-Costa et al. “Adult non-invasive pneumococcal pneumonia in Portugal is dominated by serotype 3 and non-PCV13 serotypes 3-years after near universal PCV13 use in children”. In: *Frontiers in Public Health* 11 (Dec. 20, 2023), p. 1279656. ISSN: 2296-2565. DOI: 10.3389/fpubh.2023.1279656.

[24] Daniel M. Musher, Ronald Anderson, and Charles Feldman. “The remarkable history of pneumococcal vaccination: an ongoing challenge”. In: *Pneumonia* 14.1 (Sept. 25, 2022), p. 5. ISSN: 2200-6133. DOI: 10.1186/s41479-022-00097-y.

[25] Lennard Epping et al. “SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data”. In: *Microbial Genomics* 4.7 (2018). Publisher: Microbiology Society, e000186. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000186.

[26] Oliver Lorenz et al. *SeroBA(v2.0) and SeroBAnk: a robust genome-based serotyping scheme and comprehensive atlas of capsular diversity in Streptococcus pneumoniae*. Pages: 2025.04.16.648953 Section: New Results. Apr. 21, 2025. DOI: 10.1101/2025.04.16.648953.

[27] Donald R. Walkinshaw et al. “The Streptococcus pyogenes vaccine landscape”. In: *npj Vaccines* 8.1 (Feb. 14, 2023). Publisher: Nature Publishing Group, p. 16. ISSN: 2059-0105. DOI: 10.1038/s41541-023-00609-x.

[28] Catarina Inês Mendes. “Towards accreditation in metagenomics for clinical microbiology”. Accepted: 2024-03-22T18:02:37Z. doctoralThesis. Apr. 2023.

[29] D. C. Schwartz and C. R. Cantor. “Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis”. In: *Cell* 37.1 (May 1984), pp. 67–75. ISSN: 0092-8674. DOI: 10.1016/0092-8674(84)90301-5.

[30] Jill Herschleb, Gene Ananiev, and David C. Schwartz. “Pulsed-field gel electrophoresis”. In: *Nature Protocols* 2.3 (2007), pp. 677–684. ISSN: 1750-2799. DOI: 10.1038/nprot.2007.94.

[31] Lilia Lopez-Canovas et al. “Pulsed Field Gel Electrophoresis: Past, present, and future”. In: *Analytical Biochemistry* 573 (May 15, 2019), pp. 17–29. ISSN: 0003-2697. DOI: 10.1016/j.ab.2019.02.020.

[32] Hui-min Neoh et al. “Pulsed-field gel electrophoresis (PFGE): A review of the “gold standard” for bacteria typing and current alternatives”. In: *Infection, Genetics and Evolution* 74 (Oct. 1, 2019), p. 103935. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2019.103935.

[33] S. N. Thibodeau. “Use of restriction fragment length polymorphism analysis for detecting carriers of "fragile X" syndrome”. In: *Clinical Chemistry* 33.10 (Oct. 1987), pp. 1726–1730. ISSN: 0009-9147.

[34] R. Todd et al. “From the chromosome to DNA: Restriction fragment length polymorphism analysis and its clinical application”. In: *Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons* 59.6 (June 2001), pp. 660–667. ISSN: 0278-2391. DOI: 10.1053/joms.2001.22707.

[35] E. M. Southern. “Detection of specific sequences among DNA fragments separated by gel electrophoresis”. In: *Journal of Molecular Biology* 98.3 (Nov. 5, 1975), pp. 503–517. ISSN: 0022-2836. DOI: 10.1016/s0022-2836(75)80083-0.

[36] E. H. Bingen, E. Denamur, and J. Elion. “Use of ribotyping in epidemiological surveillance of nosocomial outbreaks”. In: *Clinical Microbiology Reviews* 7.3 (July 1994), pp. 311–327. ISSN: 0893-8512. DOI: 10.1128/CMR.7.3.311.

[37] Matthew Phillip Moore et al. “K-mer based prediction of Clostridioides difficile relatedness and ribotypes”. In: *Microbial Genomics* 8.4 (2022). Publisher: Microbiology Society, p. 000804. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000804.

[38] Luiz Irber et al. “sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets”. In: *Journal of Open Source Software* 9.98 (June 28, 2024), p. 6830. ISSN: 2475-9066. DOI: 10.21105/joss.06830.

[39] Peng Qi et al. “P-224. Accurate Clostridioides difficile Ribotype Prediction from Whole Genome Sequencing Data Using Machine Learning”. In: *Open Forum Infectious Diseases* 12 (Supplement_1 Feb. 1, 2025), ofae631.428. ISSN: 2328-8957. DOI: 10.1093/ofid/ofae631.428.

[40] K. Mullis et al. “Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction”. In: *Cold Spring Harbor Symposia on Quantitative Biology* 51 Pt 1 (1986), pp. 263–273. ISSN: 0091-7451. DOI: 10.1101/sqb.1986.051.01.032.

[41] J S Chamberlain et al. “Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification.” In: *Nucleic Acids Research* 16.23 (Dec. 9, 1988), pp. 11141–11156. ISSN: 0305-1048.

[42] R. Higuchi et al. “Simultaneous amplification and detection of specific DNA sequences”. In: *Bio/Technology (Nature Publishing Company)* 10.4 (Apr. 1992), pp. 413–417. ISSN: 0733-222X. DOI: 10.1038/nbt0492-413.

[43] Mikael Kubista et al. “The real-time polymerase chain reaction”. In: *Molecular Aspects of Medicine* 27.2 (2006), pp. 95–125. ISSN: 0098-2997. DOI: 10.1016/j.mam.2005.12.007.

[44] J. Welsh and M. McClelland. “Fingerprinting genomes using PCR with arbitrary primers”. In: *Nucleic Acids Research* 18.24 (Dec. 25, 1990), pp. 7213–7218. ISSN: 0305-1048. DOI: 10.1093/nar/18.24.7213.

[45] J. G. Williams et al. “DNA polymorphisms amplified by arbitrary primers are useful as genetic markers”. In: *Nucleic Acids Research* 18.22 (Nov. 25, 1990), pp. 6531–6535. ISSN: 0305-1048. DOI: 10.1093/nar/18.22.6531.

[46] J. Versalovic, T. Koeuth, and J. R. Lupski. “Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes”. In: *Nucleic Acids Research* 19.24 (Dec. 25, 1991), pp. 6823–6831. ISSN: 0305-1048. DOI: 10.1093/nar/19.24.6823.

[47] F. J. de Bruijn. “Use of repetitive (repetitive extragenic palindromic and enterobacterial repetitive intergeneric consensus) sequences and the polymerase chain reaction to fingerprint the genomes of Rhizobium meliloti isolates and other soil bacteria”. In: *Applied and Environmental Microbiology* 58.7 (July 1992), pp. 2180–2187. ISSN: 0099-2240. DOI: 10.1128/aem.58.7.2180-2187.1992.

[48] Bjørn-Arne Lindstedt. “Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria”. In: *Electrophoresis* 26.13 (June 2005), pp. 2567–2582. ISSN: 0173-0835. DOI: 10.1002/elps.200500096.

[49] T. A. Wichelhaus et al. “Rapid molecular typing of methicillin-resistant Staphylococcus aureus by PCR-RFLP”. In: *Infection Control and Hospital Epidemiology* 22.5 (May 2001), pp. 294–298. ISSN: 0899-823X. DOI: 10.1086/501903.

[50] Jian Ye et al. “Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction”. In: *BMC Bioinformatics* 13.1 (June 18, 2012), p. 134. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-134.

[51] Ruslan Kalendar et al. “In silico PCR analysis: a comprehensive bioinformatics tool for enhancing nucleic acid amplification assays”. In: *Frontiers in Bioinformatics* 4 (Oct. 7, 2024), p. 1464197. ISSN: 2673-7647. DOI: 10.3389/fbinf.2024.1464197.

[52] W. M. Freeman, D. J. Robertson, and K. E. Vrana. “Fundamentals of DNA hybridization arrays for gene expression analysis”. In: *BioTechniques* 29.5 (2000), pp. 1042–1055. ISSN: 0736-6205. DOI: 10.2144/00295rv01.

[53] T. M. Gress et al. “Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues”. In: *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 3.11 (1992), pp. 609–619. ISSN: 0938-8990. DOI: 10.1007/BF00352477.

[54] G. G. Lennon and H. Lehrach. “Hybridization analyses of arrayed cDNA libraries”. In: *Trends in genetics: TIG* 7.10 (Oct. 1991), pp. 314–317. ISSN: 0168-9525. DOI: 10.1016/0168-9525(91)90420-u.

[55] J. DeRisi et al. “Use of a cDNA microarray to analyse gene expression patterns in human cancer”. In: *Nature Genetics* 14.4 (Dec. 1996), pp. 457–460. ISSN: 1061-4036. DOI: 10.1038/ng1296-457.

[56] M. Schena et al. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. In: *Science (New York, N.Y.)* 270.5235 (Oct. 20, 1995), pp. 467–470. ISSN: 0036-8075. DOI: 10.1126/science.270.5235.467.

[57] D. Shalon, S. J. Smith, and P. O. Brown. “A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization”. In: *Genome Research* 6.7 (July 1996), pp. 639–645. ISSN: 1088-9051. DOI: 10.1101/gr.6.7.639.

[58] Roger Bumgarner. “DNA microarrays: Types, Applications and their future”. In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* 0 22 (Jan. 2013), Unit–22.1. ISSN: 1934-3639. DOI: 10.1002/0471142727.mb2201s101.

[59] Zachery W. Dickson et al. “Probe design for simultaneous, targeted capture of diverse metagenomic targets”. In: *Cell Reports Methods* 1.6 (Sept. 15, 2021), p. 100069. ISSN: 2667-2375. DOI: 10.1016/j.crmeth.2021.100069.

[60] Tyler K Chafin, Marlis R Douglas, and Michael E Douglas. “MrBait: universal identification and design of targeted-enrichment capture probes”. In: *Bioinformatics* 34.24 (Dec. 15, 2018), pp. 4293–4296. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty548.

[61] Rita Macedo et al. “Molecular Capture of Mycobacterium tuberculosis Genomes Directly from Clinical Samples: A Potential Backup Approach for Epidemiological and Drug Susceptibility Inferences”. In: *International Journal of Molecular Sciences* 24.3 (Feb. 2, 2023), p. 2912. ISSN: 1422-0067. DOI: 10.3390/ijms24032912.

[62] Miguel Pinto et al. “Genome-scale analysis of the non-cultivable Treponema pallidum reveals extensive within-patient genetic variation”. In: *Nature Microbiology* 2.1 (Oct. 17, 2016). Publisher: Nature Publishing Group, p. 16190. ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2016.190.

[63] Tristan P. W. Dennis et al. “Target-enrichment sequencing yields valuable genomic data for challenging-to-culture bacteria of public health importance”. In: *Microbial Genomics* 8.5 (2022). Publisher: Microbiology Society, p. 000836. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000836.

[64] O. T. Avery, C. M. Macleod, and M. McCarty. “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III”. In: *The Journal of Experimental Medicine* 79.2 (Feb. 1, 1944), pp. 137–158. ISSN: 0022-1007. DOI: 10.1084/jem.79.2.137.

[65] J. D. Watson and F. H. C. Crick. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171.4356 (Apr. 1953). Publisher: Nature Publishing Group, pp. 737–738. ISSN: 1476-4687. DOI: 10.1038/171737a0.

[66] Doris T. Zallen. “Despite Franklin’s work, Wilkins earned his Nobel”. In: *Nature* 425.6953 (Sept. 2003). Publisher: Nature Publishing Group, pp. 15–15. ISSN: 1476-4687. DOI: 10.1038/425015b.

[67] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977). Publisher: Proceedings of the National Academy of Sciences, pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

[68] James M. Heather and Benjamin Chain. “The sequence of sequencers: The history of sequencing DNA”. In: *Genomics* 107.1 (Jan. 2016), pp. 1–8. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003.

[69] Raphaël Rodriguez and Yamuna Krishnan. “Genesis of next-generation sequencing”. In: *Nature biotechnology* 41.12 (Dec. 2023), pp. 1709–1715. ISSN: 1087-0156. DOI: 10.1038/s41587-023-01986-3.

[70] R. D. Fleischmann et al. “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd”. In: *Science (New York, N.Y.)* 269.5223 (July 28, 1995), pp. 496–512. ISSN: 0036-8075. DOI: 10.1126/science.7542800.

[71] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. In: *Nature* 431.7011 (Oct. 2004). Publisher: Nature Publishing Group, pp. 931–945. ISSN: 1476-4687. DOI: 10.1038/nature03001.

[72] P. Nyren, B. Pettersson, and M. Uhlen. “Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay”. In: *Analytical Biochemistry* 208.1 (Jan. 1, 1993), pp. 171–175. ISSN: 0003-2697. DOI: 10.1006/abio.1993.1024.

[73] Mostafa Ronaghi et al. “Real-Time DNA Sequencing Using Detection of Pyrophosphate Release”. In: *Analytical Biochemistry* 242.1 (Nov. 1, 1996), pp. 84–89. ISSN: 0003-2697. DOI: 10.1006/abio.1996.0432.

[74] Marcel Margulies et al. “Genome sequencing in microfabricated high-density picolitre reactors”. In: *Nature* 437.7057 (Sept. 2005). Publisher: Nature Publishing Group, pp. 376–380. ISSN: 1476-4687. DOI: 10.1038/nature03959.

[75] Nicholas J. Loman and Mark J. Pallen. “Twenty years of bacterial genome sequencing”. In: *Nature Reviews Microbiology* 13.12 (Dec. 2015). Publisher: Nature Publishing Group, pp. 787–794. ISSN: 1740-1534. DOI: 10.1038/nrmicro3565.

[76] Sara Goodwin, John D. McPherson, and W. Richard McCombie. “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17.6 (June 2016). Publisher: Nature Publishing Group, pp. 333–351. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.49.

[77] Pål Nyrén. “The History of Pyrosequencing®”. In: *Pyrosequencing: Methods and Protocols*. Ed. by Ulrich Lehmann and Jörg Tost. New York, NY: Springer, 2015, pp. 3–15. ISBN: 978-1-4939-2715-9. DOI: 10.1007/978-1-4939-2715-9_1.

[78] M. Ronaghi, M. Uhlén, and P. Nyrén. “A sequencing method based on real-time pyrophosphate”. In: *Science (New York, N.Y.)* 281.5375 (July 17, 1998), pp. 363, 365. ISSN: 0036-8075. DOI: 10.1126/science.281.5375.363.

[79] P. Nyrén. “Enzymatic method for continuous monitoring of DNA polymerase activity”. In: *Analytical Biochemistry* 167.2 (Dec. 1987), pp. 235–238. ISSN: 0003-2697. DOI: 10.1016/0003-2697(87)90158-8.

[80] Koen Andries et al. “A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis”. In: *Science (New York, N.Y.)* 307.5707 (Jan. 14, 2005), pp. 223–227. ISSN: 1095-9203. DOI: 10.1126/science.1106753.

[81] Jonathan M. Rothberg and John H. Leamon. “The development and impact of 454 sequencing”. In: *Nature Biotechnology* 26.10 (Oct. 2008). Publisher: Nature Publishing Group, pp. 1117–1124. ISSN: 1546-1696. DOI: 10.1038/nbt1485.

[82] David A. Wheeler et al. “The complete genome of an individual by massively parallel DNA sequencing”. In: *Nature* 452.7189 (Apr. 2008). Publisher: Nature Publishing Group, pp. 872–876. ISSN: 1476-4687. DOI: 10.1038/nature06884.

[83] Gerardo Turcatti et al. “A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis”. In: *Nucleic Acids Research* 36.4 (Mar. 2008), e25. ISSN: 1362-4962. DOI: 10.1093/nar/gkn021.

[84] Mathias Uhlen and Stephen R. Quake. “Sequential sequencing by synthesis and the next-generation sequencing revolution”. In: *Trends in Biotechnology* 41.12 (Dec. 2023), pp. 1565–1572. ISSN: 01677799. DOI: 10.1016/j.tibtech.2023.06.007.

[85] David R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456.7218 (Nov. 2008). Publisher: Nature Publishing Group, pp. 53–59. ISSN: 1476-4687. DOI: 10.1038/nature07517.

[86] Milan Fedurco et al. “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies”. In: *Nucleic Acids Research* 34.3 (Feb. 9, 2006), e22. ISSN: 1362-4962. DOI: 10.1093/nar/gnj023.

[87] Junjun Zhang et al. “The International Cancer Genome Consortium Data Portal”. In: *Nature Biotechnology* 37.4 (Apr. 2019). Publisher: Nature Publishing Group, pp. 367–369. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0055-9.

[88] 000 Genomes Project Pilot Investigators The 100. “100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report”. In: *New England Journal of Medicine* 385.20 (Nov. 10, 2021). Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2035790, pp. 1868–1880. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2035790.

[89] Aaron M. Wenger et al. “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome”. In: *Nature Biotechnology* 37.10 (Oct. 2019). Publisher: Nature Publishing Group, pp. 1155–1162. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0217-9.

[90] John Eid et al. “Real-time DNA sequencing from single polymerase molecules”. In: *Science (New York, N.Y.)* 323.5910 (Jan. 2, 2009), pp. 133–138. ISSN: 1095-9203. DOI: 10.1126/science.1162986.

[91] Alexander S. Mikheyev and Mandy M. Y. Tin. “A first look at the Oxford Nanopore MinION sequencer”. In: *Molecular Ecology Resources* 14.6 (Nov. 2014), pp. 1097–1102. ISSN: 1755-0998. DOI: 10.1111/1755-0998.12324.

[92] David Stoddart et al. “Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.19 (May 12, 2009), pp. 7702–7707. ISSN: 1091-6490. DOI: 10.1073/pnas.0901054106.

[93] Yunhao Wang et al. “Nanopore sequencing technology, bioinformatics and applications”. In: *Nature Biotechnology* 39.11 (Nov. 2021). Publisher: Nature Publishing Group, pp. 1348–1365. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01108-x.

[94] Ryan R. Wick, Benjamin P. Howden, and Timothy P. Stinear. *Autocycler: long-read consensus assembly for bacterial genomes*. Pages: 2025.05.12.653612 Section: New Results. May 15, 2025. DOI: 10.1101/2025.05.12.653612.

[95] Ebenezer Foster-Nyarko et al. “Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, Klebsiella pneumoniae”. In: *Microbial Genomics* 9.2 (2023). Publisher: Microbiology Society, p. 000936. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000936.

[96] Ryan R. Wick et al. “Trycycler: consensus long-read assemblies for bacterial genomes”. In: *Genome Biology* 22.1 (Sept. 14, 2021), p. 266. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02483-z.

[97] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. "Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing". In: *PLOS Computational Biology* 19.3 (Mar. 2, 2023). Publisher: Public Library of Science, e1010905. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010905.

[98] George Bouras et al. "Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies". In: *Microbial Genomics* 10.5 (2024). Publisher: Microbiology Society, p. 001244. ISSN: 2057-5858. DOI: 10.1099/mgen.0.001244.

[99] Bo Maxwell Stevens et al. "Comparison of Oxford Nanopore Technologies and Illumina MiSeq sequencing with mock communities and agricultural soil". In: *Scientific Reports* 13.1 (June 8, 2023). Publisher: Nature Publishing Group, p. 9323. ISSN: 2045-2322. DOI: 10.1038/s41598-023-36101-8.

[100] Michael L. Metzker. "Sequencing technologies — the next generation". In: *Nature Reviews Genetics* 11.1 (Jan. 2010). Publisher: Nature Publishing Group, pp. 31–46. ISSN: 1471-0064. DOI: 10.1038/nrg2626.

[101] Ezra J. Barzilay et al. "Cholera Surveillance during the Haiti Epidemic — The First 2 Years". In: *New England Journal of Medicine* 368.7 (Feb. 14, 2013). Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1204927, pp. 599–609. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1204927.

[102] Lisa A. King et al. "Outbreak of Shiga toxin-producing Escherichia coli O104:H4 associated with organic fenugreek sprouts, France, June 2011". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 54.11 (June 2012), pp. 1588–1594. ISSN: 1537-6591. DOI: 10.1093/cid/cis255.

[103] Alexander Mellmann et al. "Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology". In: *PloS One* 6.7 (2011), e22751. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022751.

[104] Andrey Prjibelski et al. "Using SPAdes De Novo Assembler". In: *Current Protocols in Bioinformatics* 70.1 (2020). _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.102, e102. ISSN: 1934-340X. DOI: 10.1002/cpbi.102.

[105] Bruce J. Walker et al. "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In: *PLOS ONE* 9.11 (Nov. 19, 2014). Publisher: Public Library of Science, e112963. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112963.

[106] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 1, 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170.

[107] Robert A. Petit and Timothy D. Read. "Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes". In: *mSystems* 5.4 (Aug. 4, 2020). Publisher: American Society for Microbiology, 10.1128/msystems.00190–20. DOI: 10.1128/msystems.00190-20.

[108] J. Besser et al. "Next-generation sequencing technologies and their application to the study and control of bacterial infections". In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 24.4 (Apr. 2018), pp. 335–341. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2017.10.013.

[109] Ruud H. Deurenberg et al. "Application of next generation sequencing in clinical microbiology and infection prevention". In: *Journal of Biotechnology* 243 (Feb. 10, 2017), pp. 16–24. ISSN: 0168-1656. DOI: 10.1016/j.jbiotec.2016.12.022.

[110] *WHO global strategy for food safety 2022-2030: towards stronger food safety systems and global cooperation*.

[111] Brendan R. Jackson et al. "Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation". In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 63.3 (Aug. 1, 2016), pp. 380–386. ISSN: 1537-6591. DOI: 10.1093/cid/ciw242.

[112] Efrain M. Ribot et al. "PulseNet: Entering the Age of Next-Generation Sequencing". In: *Foodborne Pathogens and Disease* 16.7 (July 1, 2019), pp. 451–456. ISSN: 1535-3141. DOI: 10.1089/fpd.2019.2634.

[113] European Centre for Disease Prevention and Control. *ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations: 2019–2021*. LU: Publications Office, 2019.

[114] European Food Safety Authority (EFSA) et al. "Guidelines for reporting Whole Genome Sequencing-based typing data through the EFSA One Health WGS System". In: *EFSA Supporting Publications* 19.6 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2022.EN-7413, 7413E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2022.EN-7413.

[115] "Prolonged multi-country outbreak of Listeria monocytogenes ST1607 linked to smoked salmon products". In: *EFSA Supporting Publications* 21.5 (2024). _eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2024.EN-8810, 8810E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2024.EN-8810.

[116] European Food Safety Authority. "Prolonged multi-country outbreak of Listeria monocytogenes ST173 linked to consumption of fish products". In: *EFSA Supporting Publications* 21.6 (2024). _eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2024.EN-8885, 8885E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2024.EN-8885.

[117] European Centre for Disease Prevention and Control, European Food Safety Authority. "Multi-country outbreak of Salmonella Virchow ST16 infections linked to the consumption of meat products containing chicken meat". In: *EFSA Supporting Publications* 20.4 (2023). _eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2023.EN-7983, 7983E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2023.EN-7983.

[118] European Centre for Disease Prevention and Control, European Food Safety Authority. "Multi-country outbreak of monophasic Salmonella Typhimurium sequence type 34 linked to chocolate products – first update – 18 May 2022". In: *EFSA Supporting Publications* 19.6 (2022). _eprint: https://efsa.onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2022.EN-7352, 7352E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2022.EN-7352.

[119] Verónica Mixão et al. "Multi-country and intersectoral assessment of cluster congruence between pipelines for genomics surveillance of foodborne pathogens". In: *Nature Communications* 16.1 (Apr. 28, 2025). Publisher: Nature Publishing Group, p. 3961. ISSN: 2041-1723. DOI: 10.1038/s41467-025-59246-8.

[120] Rachel Urwin and Martin C. J. Maiden. "Multi-locus sequence typing: a tool for global epidemiology". In: *Trends in Microbiology* 11.10 (Oct. 1, 2003). Publisher: Elsevier, pp. 479–487. ISSN: 0966-842X, 1878-4380. DOI: 10.1016/j.tim.2003.08.006.

[121] Martin C. J. Maiden et al. "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms". In: *Proceedings of the National Academy of Sciences* 95.6 (Mar. 17, 1998). Publisher: Proceedings of the National Academy of Sciences, pp. 3140–3145. DOI: 10.1073/pnas.95.6.3140.

[122] Keith A. Jolley et al. "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain". In: *Microbiology (Reading, England)* 158 (Pt 4 Apr. 2012), pp. 1005–1015. ISSN: 1465-2080. DOI: 10.1099/mic.0.055459-0.

[123] Mark C. Enright and Brian G. Spratt. "Multilocus sequence typing". In: *Trends in Microbiology* 7.12 (Dec. 1, 1999), pp. 482–487. ISSN: 0966-842X. DOI: 10.1016/S0966-842X(99)01609-1.

[124] Mark C. Enright and Brian G. Spratt. "A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease". In: *Microbiology* 144.11 (1998). Publisher: Microbiology Society, pp. 3049–3060. ISSN: 1465-2080. DOI: 10.1099/00221287-144-11-3049.

[125] Martin C. J. Maiden et al. "MLST revisited: the gene-by-gene approach to bacterial genomics". In: *Nature Reviews Microbiology* 11.10 (Oct. 2013). Publisher: Nature Publishing Group, pp. 728–736. ISSN: 1740-1534. DOI: 10.1038/nrmicro3093.

[126] Lavin A. Joseph et al. “Evaluation of core genome and whole genome multilocus sequence typing schemes for Campylobacter jejuni and Campylobacter coli outbreak detection in the USA”. In: *Microbial Genomics* 9.5 (2023). Publisher: Microbiology Society, p. 001012. ISSN: 2057-5858. DOI: 10.1099/mgen.0.001012.

[127] Molly M. Leeper et al. “Evaluation of whole and core genome multilocus sequence typing allele schemes for Salmonella enterica outbreak detection in a national surveillance network, PulseNet USA”. In: *Frontiers in Microbiology* 14 (Sept. 21, 2023). Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2023.1254777.

[128] Molly M. Leeper et al. “Validation of Core and Whole-Genome Multi-Locus Sequence Typing Schemes for Shiga-Toxin-Producing E. coli (STEC) Outbreak Detection in a National Surveillance Network, PulseNet 2.0, USA”. In: *Microorganisms* 13.6 (June 2025). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 1310. ISSN: 2076-2607. DOI: 10.3390/microorganisms13061310.

[129] Keith A. Jolley and Martin CJ Maiden. “BIGSdb: Scalable analysis of bacterial genome variation at the population level”. In: *BMC Bioinformatics* 11.1 (Dec. 10, 2010), p. 595. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-595.

[130] Keith A. Jolley, James E. Bray, and Martin C. J. Maiden. “Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications”. In: *Wellcome Open Research* 3 (2018), p. 124. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.14826.1.

[131] Timothy Dallman et al. “SnapperDB: a database solution for routine sequencing analysis of bacterial isolates”. In: *Bioinformatics* 34.17 (Sept. 1, 2018), pp. 3028–3029. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty212.

[132] Michael Roberts et al. “Reducing storage requirements for biological sequence comparison”. In: *Bioinformatics (Oxford, England)* 20.18 (Dec. 12, 2004), pp. 3363–3369. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth408.

[133] Kristoffer Sahlin. “Effective sequence similarity detection with strobemers”. In: *Genome Research* 31.11 (Jan. 11, 2021). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 2080–2094. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.275648.121.

[134] Malick Ndiaye et al. “When less is more: sketching with minimizers in genomics”. In: *Genome Biology* 25.1 (Oct. 14, 2024), p. 270. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03414-4.

[135] Moein Karami et al. “Designing efficient randstrobes for sequence similarity analyses”. In: *Bioinformatics* (Apr. 5, 2024), btae187. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btae187.

[136] Bryce Kille et al. “Minmers are a generalization of minimizers that enable unbiased local Jaccard estimation”. In: *Bioinformatics* 39.9 (Sept. 1, 2023), btad512. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad512.

[137] Jim Shaw and Yun William Yu. “Theory of local k-mer selection with applications to long-read alignment”. In: *Bioinformatics* 38.20 (Oct. 15, 2022), pp. 4659–4669. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab790.

[138] Paul Medvedev and Mihai Pop. “What do Eulerian and Hamiltonian cycles have to do with genome assembly?” In: *PLOS Computational Biology* 17.5 (May 20, 2021). Publisher: Public Library of Science, e1008928. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008928.

[139] Derrick E. Wood and Steven L. Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments”. In: *Genome Biology* 15.3 (Mar. 3, 2014), R46. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-3-r46.

[140] Derrick E. Wood, Jennifer Lu, and Ben Langmead. “Improved metagenomic analysis with Kraken 2”. In: *Genome Biology* 20.1 (Nov. 28, 2019), p. 257. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1891-0.

[141] Romain Derelle et al. “Seamless, rapid, and accurate analyses of outbreak genomic data using split k-mer analysis”. In: *Genome Research* 34.10 (Jan. 10, 2024). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1661–1673. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.279449.124.

[142] John A. Lees et al. “Fast and flexible bacterial genomic epidemiology with PopPUNK”. In: *Genome Research* 29.2 (Jan. 2, 2019). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 304–316. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.241455.118.

[143] Mickael Silva et al. “chewBBACA: A complete suite for gene-by-gene schema creation and strain identification”. In: *Microbial Genomics* 4.3 (Mar. 15, 2018), e000166. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000166.

[144] European Food Safety Authority (EFSA). “EFSA statement on the requirements for whole genome sequence analysis of microorganisms intentionally used in the food chain”. In: *EFSA Journal* 22.8 (2024). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/j.efsa.2024.8912, e8912. ISSN: 1831-4732. DOI: 10.2903/j.efsa.2024.8912.

[145] Marc J. Struelens et al. “Real-time genomic surveillance for enhanced control of infectious diseases and antimicrobial resistance”. In: *Frontiers in Science* 2 (Apr. 25, 2024). Publisher: Frontiers. ISSN: 2813-6330. DOI: 10.3389/fsci.2024.1298248.

[146] Laura Uelze et al. “Typing methods based on whole genome sequencing data”. In: *One Health Outlook* 2.1 (Feb. 18, 2020), p. 3. ISSN: 2524-4655. DOI: 10.1186/s42522-020-0010-1.

[147] Stephen J Bush et al. “Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines”. In: *GigaScience* 9.2 (Feb. 1, 2020), giaa007. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa007.

[148] Carlos Valiente-Mullor et al. “One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads”. In: *PLOS Computational Biology* 17.1 (Jan. 27, 2021). Publisher: Public Library of Science, e1008678. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008678.

[149] Mahdi Belbasi et al. “The minimizer Jaccard estimator is biased and inconsistent”. In: *Bioinformatics* 38 (Supplement_1 June 24, 2022), pp. i169–i176. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac244.

[150] Alannah C. King et al. “Comparison of gene-by-gene and genome-wide short nucleotide sequence-based approaches to define the global population structure of Streptococcus pneumoniae”. In: *Microbial Genomics* 10.8 (2024). Publisher: Microbiology Society, p. 001278. ISSN: 2057-5858. DOI: 10.1099/mgen.0.001278.

[151] Zhemin Zhou et al. “The EnteroBase user’s guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity”. In: *Genome Research* 30.1 (Jan. 2020), pp. 138–152. ISSN: 1549-5469. DOI: 10.1101/gr.251678.119.

[152] Rafael Mamede et al. “Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas”. In: *Nucleic Acids Research* 49 (D1 Jan. 8, 2021), pp. D660–D666. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa889.

[153] *cgMLST.org Nomenclature Server (h25)*.

[154] David A. Rasko, Garry SA Myers, and Jacques Ravel. “Visualization of comparative genomic analyses by BLAST score ratio”. In: *BMC Bioinformatics* 6.1 (Jan. 5, 2005), p. 2. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-2.

[155] Christiam Camacho et al. “BLAST+: architecture and applications”. In: *BMC Bioinformatics* 10.1 (Dec. 15, 2009), p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421.

[156] The UniProt Consortium. “UniProt: the Universal Protein Knowledgebase in 2025”. In: *Nucleic Acids Research* 53 (D1 Jan. 6, 2025), pp. D609–D617. ISSN: 0305-1048. DOI: 10.1093/nar/gkae1010.

[157] *React*.

[158] Martin Larralde. “Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes”. In: *Journal of Open Source Software* 7.72 (Apr. 25, 2022), p. 4296. ISSN: 2475-9066. DOI: 10.21105/joss.04296.

[159] Doug Hyatt et al. “Prodigal: prokaryotic gene recognition and translation initiation site identification”. In: *BMC Bioinformatics* 11.1 (Mar. 8, 2010), p. 119. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119.

[160] Alexandre Lomsadze et al. “Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes”. In: *Genome Research* 28.7 (July 2018), pp. 1079–1089. ISSN: 1549-5469. DOI: 10.1101/gr.230615.117.

[161] Markus J. Sommer and Steven L. Salzberg. “Balrog: A universal protein model for prokaryotic gene prediction”. In: *PLOS Computational Biology* 17.2 (Feb. 26, 2021). Publisher: Public Library of Science, e1008727. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008727.

[162] Saul Schleimer, Daniel S Wilkerson, and Alex Aiken. “Winnowing: Local Algorithms for Document Fingerprinting”. In: ().

[163] Guillaume Marçais et al. “Improving the performance of minimizers and winnowing schemes”. In: *Bioinformatics* 33.14 (July 15, 2017), pp. i110–i117. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx235.

[164] Hongyu Zheng, Carl Kingsford, and Guillaume Marçais. “Improved design and analysis of practical minimizers”. In: *Bioinformatics* 36 (Supplement_1 July 1, 2020), pp. i119–i127. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa472.

[165] *chewBBACA 3 documentation*.

[166] Eric W Sayers et al. “Database resources of the national center for biotechnology information”. In: *Nucleic Acids Research* 50 (D1 Jan. 7, 2022), pp. D20–D26. ISSN: 0305-1048. DOI: 10.1093/nar/gkab1112.

[167] Adrien Biguenet et al. “Introduction and benchmarking of pyMLST: open-source software for assessing bacterial clonality using core genome MLST”. In: *Microbial Genomics* 9.11 (Nov. 15, 2023), p. 001126. DOI: 10.1099/mgen.0.001126.

[168] Nicola N Lynskey et al. “Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study”. In: *The Lancet Infectious Diseases* 19.11 (Nov. 1, 2019), pp. 1209–1218. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(19)30446-3.

[169] Thor Bech Johannesen et al. “Increase in invasive group A streptococcal infections and emergence of novel, rapidly expanding sub-lineage of the virulent Streptococcus pyogenes M1 clone, Denmark, 2023”. In: *Eurosurveillance* 28.26 (June 29, 2023), p. 2300291. DOI: 10.2807/1560-7917.ES.2023.28.26.2300291.

[170] Rafael Mamede et al. *Supplementary material of "chewBBACA 3: lowering the barrier for scalable and detailed whole- and core-genome multilocus sequence typing"*. Version 1.0.0. Jan. 13, 2025. DOI: 10.5281/zenodo.14637859.

[171] Kazutaka Katoh and Daron M. Standley. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. In: *Molecular Biology and Evolution* 30.4 (Jan. 16, 2013), p. 772. DOI: 10.1093/molbev/mst010.

[172] Khalil Abudahab et al. *Phylocanvas.gl: A WebGL-powered JavaScript library for large tree visualisation*. July 2, 2021. DOI: 10.31219/osf.io/nfv6m.

[173] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". In: *PLOS ONE* 5.3 (Mar. 10, 2010). Publisher: Public Library of Science, e9490. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490.

[174] Nuala A. O'Leary et al. "Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets". In: *Scientific Data* 11.1 (July 5, 2024). Publisher: Nature Publishing Group, p. 732. ISSN: 2052-4463. DOI: 10.1038/s41597-024-03571-y.

[175] Grace A. Blackwell et al. "Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences". In: *PLoS biology* 19.11 (Nov. 2021), e3001421. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001421.

[176] Torsten Seemann. *mlst*.

[177] Jim Shaw and Yun William Yu. "Fast and robust metagenomic sequence comparison through sparse chaining with skani". In: *Nature Methods* 20.11 (Nov. 2023). Publisher: Nature Publishing Group, pp. 1661–1665. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02018-3.

[178] Sebastian Deorowicz, Agnieszka Danek, and Heng Li. "AGC: compact representation of assembled genomes with fast queries and updates". In: *Bioinformatics* 39.3 (Mar. 1, 2023), btad097. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad097.

[179] A. Friães et al. "Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of Streptococcus pyogenes". In: *Journal of Clinical Microbiology* 60.6 (May 9, 2022). Publisher: American Society for Microbiology, e00315–22. DOI: 10.1128/jcm.00315-22.

[180] Paul Sumby et al. "Evolutionary Origin and Emergence of a Highly Successful Clone of Serotype M1 Group A Streptococcus Involved Multiple Horizontal Gene Transfer Events". In: *The Journal of Infectious Diseases* 192.5 (Sept. 1, 2005), pp. 771–782. ISSN: 0022-1899. DOI: 10.1086/432514.

[181] Allison Black et al. "Ten recommendations for supporting open pathogen genomic analysis in public health". In: *Nature Medicine* 26.6 (June 2020), pp. 832–841. ISSN: 1546-170X. DOI: 10.1038/s41591-020-0935-z.

[182] Xianding Deng et al. "Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California". In: *Science (New York, N.Y.)* 369.6503 (July 31, 2020), pp. 582–587. ISSN: 1095-9203. DOI: 10.1126/science.abb9263.

[183] Joana Isidro et al. "Virulence and antibiotic resistance plasticity of *Arcobacter butzleri*: Insights on the genomic diversity of an emerging human pathogen". In: *Infection, Genetics and Evolution* 80 (June 1, 2020), p. 104213. ISSN: 1567-1348. DOI: 10.1016/j.meegid.2020.104213.

[184] Ann-Katrin Llarena et al. “INNUENDO: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens”. In: *EFSA Supporting Publications* 15.11 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2018.EN-1498, 1498E. ISSN: 2397-8325. DOI: 10.2903/sp.efsa.2018.EN-1498.

[185] J. A. Carriço et al. “Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution”. In: *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 18.4 (Jan. 24, 2013), p. 20382. ISSN: 1560-7917. DOI: 10.2807/ese.18.04.20382-en.

[186] Cátia Vaz et al. “TypOn: the microbial typing ontology”. In: *Journal of Biomedical Semantics* 5.1 (Oct. 18, 2014), p. 43. ISSN: 2041-1480. DOI: 10.1186/2041-1480-5-43.

[187] Mark van der Linden et al. “Heterogeneity of penicillin-non-susceptible group B streptococci isolated from a single patient in Germany”. In: *The Journal of Antimicrobial Chemotherapy* 75.2 (Feb. 1, 2020), pp. 296–299. ISSN: 1460-2091. DOI: 10.1093/jac/dkz465.

[188] UniProt Consortium. “UniProt: a worldwide hub of protein knowledge”. In: *Nucleic Acids Research* 47 (D1 Jan. 8, 2019), pp. D506–D515. ISSN: 1362-4962. DOI: 10.1093/nar/gky1049.

[189] Andrew J. Page et al. “Comparison of classical multi-locus sequence typing software for next-generation sequencing data”. In: *Microbial Genomics* 3.8 (2017). Publisher: Microbiology Society, e000124. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000124.

[190] S. F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (Oct. 5, 1990), pp. 403–410. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(05)80360-2.

[191] Yuan Li et al. “Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis”. In: *Nature Communications* 10.1 (Jan. 14, 2019), p. 178. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07997-y.

[192] Masaya Yamaguchi et al. “Identification of evolutionarily conserved virulence factor by selective pressure analysis of Streptococcus pneumoniae”. In: *Communications Biology* 2 (2019), p. 96. ISSN: 2399-3642. DOI: 10.1038/s42003-019-0340-7.

[193] Richard Moxon, Pedro A. Reche, and Rino Rappuoli. “Editorial: Reverse Vaccinology”. In: *Frontiers in Immunology* 10 (2019), p. 2776. ISSN: 1664-3224. DOI: 10.3389/fimmu.2019.02776.

[194] Martin Christopher James Maiden. “The Impact of Nucleotide Sequence Analysis on Meningococcal Vaccine Development and Assessment”. In: *Frontiers in Immunology* 9 (2018), p. 3151. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.03151.

[195] Can Firtina et al. “Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm”. In: *Bioinformatics (Oxford, England)* 36.12 (June 1, 2020), pp. 3669–3679. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa179.

[196] Paul G. Higgins et al. “Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of Acinetobacter baumannii”. In: *PloS One* 12.6 (2017), e0179228. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0179228.

[197] Jonathan R. Carapetis et al. “The global burden of group A streptococcal diseases”. In: *The Lancet. Infectious Diseases* 5.11 (Nov. 2005), pp. 685–694. ISSN: 1473-3099. DOI: 10.1016/S1473-3099(05)70267-X.

[198] Johan Vekemans et al. “The Path to Group A Streptococcus Vaccines: World Health Organization Research and Development Technology Roadmap and Preferred Product Characteristics”. In: *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 69.5 (Aug. 16, 2019), pp. 877–883. ISSN: 1537-6591. DOI: 10.1093/cid/ciy1143.

[199] B. Beall, R. Facklam, and T. Thompson. “Sequencing emm-specific PCR products for routine and accurate typing of group A streptococci”. In: *Journal of Clinical Microbiology* 34.4 (Apr. 1996), pp. 953–958. ISSN: 0095-1137. DOI: 10.1128/jcm.34.4.953-958.1996.

[200] J. A. Carriço et al. “Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes”. In: *Journal of Clinical Microbiology* 44.7 (July 2006), pp. 2524–2532. ISSN: 0095-1137. DOI: 10.1128/JCM.02536-05.

[201] A. Friães et al. “Superantigen gene complement of Streptococcus pyogenes–relationship with other typing methods and short-term stability”. In: *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* 32.1 (Jan. 2013), pp. 115–125. ISSN: 1435-4373. DOI: 10.1007/s10096-012-1726-3.

[202] M. C. Enright et al. “Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone”. In: *Infection and Immunity* 69.4 (Apr. 2001), pp. 2416–2427. ISSN: 0019-9567. DOI: 10.1128/IAI.69.4.2416-2427.2001.

[203] Ana Friães et al. “Emergence of the Same Successful Clade among Distinct Populations of emm89 Streptococcus pyogenes in Multiple Geographic Regions”. In: *mBio* 6.6 (Dec. 1, 2015), e01780–01715. ISSN: 2150-7511. DOI: 10.1128/mBio.01780-15.

[204] Claire E. Turner et al. “Emergence of a New Highly Successful Acapsular Group A Streptococcus Clade of Genotype emm89 in the United Kingdom”. In: *mBio* 6.4 (July 14, 2015), e00622. ISSN: 2150-7511. DOI: 10.1128/mBio.00622-15.

[205] Luchang Zhu et al. “A molecular trigger for intercontinental epidemics of group A Streptococcus”. In: *The Journal of Clinical Investigation* 125.9 (Sept. 2015), pp. 3545–3559. ISSN: 1558-8238. DOI: 10.1172/JCI82478.

[206] Luchang Zhu et al. “Trading Capsule for Increased Cytotoxin Production: Contribution to Virulence of a Newly Emerged Clade of emm89 Streptococcus pyogenes”. In: *mBio* 6.5 (Oct. 6, 2015), e01378–01315. ISSN: 2150-7511. DOI: 10.1128/mBio.01378-15.

[207] Lidewij W. Rümke et al. “Dominance of M1UK clade among Dutch M1 Streptococcus pyogenes”. In: *The Lancet. Infectious Diseases* 20.5 (May 2020), pp. 539–540. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(20)30278-4.

[208] Yuan Li et al. “M1UK lineage in invasive group A streptococcus isolates from the USA”. In: *The Lancet. Infectious Diseases* 20.5 (May 2020), pp. 538–539. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(20)30279-6.

[209] Walter Demczuk et al. “Identification of Streptococcus pyogenes M1UK clone in Canada”. In: *The Lancet. Infectious Diseases* 19.12 (Dec. 2019), pp. 1284–1285. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(19)30622-X.

[210] Waleed Nasser et al. “Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.17 (Apr. 29, 2014), E1768–1776. ISSN: 1091-6490. DOI: 10.1073/pnas.1403138111.

[211] Stephen B. Beres et al. “Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.9 (Mar. 2, 2010), pp. 4371–4376. ISSN: 1091-6490. DOI: 10.1073/pnas.0911295107.

[212] Claire E. Turner et al. “Community outbreaks of group A Streptococcus revealed by genome sequencing”. In: *Scientific Reports* 7.1 (Aug. 17, 2017), p. 8554. ISSN: 2045-2322. DOI: 10.1038/s41598-017-08914-x.

[213] J. M. Coelho et al. “Genomic sequence investigation Streptococcus pyogenes clusters in England (2010-2015)”. In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 25.1 (Jan. 2019), pp. 96–101. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2018.04.011.

[214] Mark R. Davies et al. “Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics”. In: *Nature Genetics* 51.6 (June 2019), pp. 1035–1043. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0417-8.

[215] Stephen B. Beres et al. “Integrative Reverse Genetic Analysis Identifies Polymorphisms Contributing to Decreased Antimicrobial Agent Susceptibility in Streptococcus pyogenes”. In: *mBio* 13.1 (Jan. 18, 2022). Publisher: American Society for Microbiology, e03618–21. DOI: 10.1128/mbio.03618-21.

[216] Bernd Neumann et al. “A Core Genome Multilocus Sequence Typing Scheme for Enterococcus faecalis”. In: *Journal of Clinical Microbiology* 57.3 (Mar. 2019), e01686–18. ISSN: 1098-660X. DOI: 10.1128/JCM.01686-18.

[217] Karen F. McGregor et al. “Multilocus sequence typing of Streptococcus pyogenes representing most known emm types and distinctions among subpopulation genetic structures”. In: *Journal of Bacteriology* 186.13 (July 2004), pp. 4285–4294. ISSN: 0021-9193. DOI: 10.1128/JB.186.13.4285-4294.2004.

[218] J. A. Carriço et al. “A primer on microbial bioinformatics for nonbioinformaticians”. In: *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 24.4 (Apr. 2018), pp. 342–349. ISSN: 1469-0691. DOI: 10.1016/j.cmi.2017.12.015.

[219] Shana R. Leopold et al. “Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes”. In: *Journal of Clinical Microbiology* 52.7 (July 2014), pp. 2365–2370. ISSN: 1098-660X. DOI: 10.1128/JCM.00262-14.

[220] Stefan Bletz et al. “Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of Clostridium difficile”. In: *Journal of Clinical Microbiology* 56.6 (June 2018), e01987–17. ISSN: 1098-660X. DOI: 10.1128/JCM.01987-17.

[221] Mostafa Y. Abdel-Glil et al. “A Whole-Genome-Based Gene-by-Gene Typing System for Standardized High-Resolution Strain Typing of Bacillus anthracis”. In: *Journal of Clinical Microbiology* 59.7 (June 18, 2021), e0288920. ISSN: 1098-660X. DOI: 10.1128/JCM.02889-20.

[222] M. Pinto et al. “Insights into the population structure and pan-genome of Haemophilus influenzae”. In: *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 67 (Jan. 2019), pp. 126–135. ISSN: 1567-7257. DOI: 10.1016/j.meegid.2018.10.025.

[223] Svetlana Bardenstein et al. “Brucellosis Outbreak Traced to Commercially Sold Camel Milk through Whole-Genome Sequencing, Israel”. In: *Emerging Infectious Diseases* 27.6 (June 2021), pp. 1728–1731. ISSN: 1080-6059. DOI: 10.3201/eid2706.204902.

[224] Ana Friães et al. *Supplemental material of "An annotated whole-genome multilocus sequence typing schema for scalable high resolution typing of Streptococcus pyogenes"*. Version 3.0.0. Aug. 3, 2023. DOI: 10.5281/zenodo.8211298.

[225] Ana Friães et al. “Group A streptococci clones associated with invasive infections and pharyngitis in Portugal present differences in emm types, superantigen gene content and antimicrobial resistance”. In: *BMC microbiology* 12 (Nov. 27, 2012), p. 280. ISSN: 1471-2180. DOI: 10.1186/1471-2180-12-280.

[226] Catarina Pato et al. “Streptococcus pyogenes Causing Skin and Soft Tissue Infections Are Enriched in the Recently Emerged emm89 Clade 3 and Are Not Associated With Abrogation of CovRS”. In: *Frontiers in Microbiology* 9 (2018), p. 2372. ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.02372.

[227] Ana Friães et al. “Changes in Streptococcus pyogenes causing invasive disease in Portugal: evidence for superantigen gene loss and acquisition”. In: *International journal of medical microbiology: IJMM* 303.8 (Dec. 2013), pp. 505–513. ISSN: 1618-0607. DOI: 10.1016/j.ijmm.2013.07.004.

[228] *Streptococcus pyogenes como agente de infecção da pele e tecidos moles - ProQuest*.

[229] Nuala A. O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D733–D745. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1189.

[230] *Release INNUca v4.2.2 - SPAdes v3.14.0 · B-UMMI/INNUca*. GitHub.

[231] *Release v0.2.0 · MDU-PHL/emmtyper*. GitHub.

[232] *Release Play it again Sam, but double the tempo · tseemann/snippy*. GitHub.

[233] *B-UMMI/seq_typing*. original-date: 2018-01-10T15:51:04Z. Apr. 15, 2025.

[234] *Release Initial development · B-UMMI/Schema_Refinery*. GitHub.

[235] Marta Nascimento et al. “PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods”. In: *Bioinformatics (Oxford, England)* 33.1 (Jan. 1, 2017), pp. 128–129. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw582.

[236] Bruno Ribeiro-Gonçalves et al. “PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees”. In: *Nucleic Acids Research* 44 (W1 July 8, 2016), W246–W251. ISSN: 0305-1048. DOI: 10.1093/nar/gkw359.

[237] Ana Severiano et al. “Adjusted Wallace coefficient as a measure of congruence between typing methods”. In: *Journal of Clinical Microbiology* 49.11 (Nov. 2011), pp. 3997–4000. ISSN: 1098-660X. DOI: 10.1128/JCM.00624-11.

[238] C. Silva-Costa et al. “Differences between macrolide-resistant and -susceptible Streptococcus pyogenes: importance of clonal properties in addition to antibiotic consumption”. In: *Antimicrobial Agents and Chemotherapy* 56.11 (Nov. 2012), pp. 5661–5666. ISSN: 1098-6596. DOI: 10.1128/AAC.01133-12.

[239] Catarina Silva-Costa et al. “Macrolide-resistant Streptococcus pyogenes: prevalence and treatment strategies”. In: *Expert Review of Anti-Infective Therapy* 13.5 (May 2015), pp. 615–628. ISSN: 1744-8336. DOI: 10.1586/14787210.2015.1023292.

[240] Francesco Iannelli et al. “Type M Resistance to Macrolides Is Due to a Two-Gene Efflux Transport System of the ATP-Binding Cassette (ABC) Superfamily”. In: *Frontiers in Microbiology* 9 (2018), p. 1670. ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.01670.

[241] Andrew C. Steer et al. “Global emm type distribution of group A streptococci: systematic review and implications for vaccine development”. In: *The Lancet. Infectious Diseases* 9.10 (Oct. 2009), pp. 611–616. ISSN: 1474-4457. DOI: 10.1016/S1473-3099(09)70178-1.

[242] T. C. Barnett, A. C. Bowen, and J. R. Carapetis. “The fall and rise of Group A Streptococcus diseases”. In: *Epidemiology and Infection* 147 (Aug. 15, 2018), e4. ISSN: 1469-4409. DOI: 10.1017/S0950268818002285.

[243] A. J. Sabat et al. “Overview of molecular typing methods for outbreak detection and epidemiological surveillance”. In: *Eurosurveillance* 18.4 (Jan. 24, 2013). Publisher: European Centre for Disease Prevention and Control, p. 20380. ISSN: 1560-7917. DOI: 10.2807/ese.18.04.20380-en.

[244] Rafael C. Jiménez et al. “Four simple recommendations to encourage best practices in research software”. In: *F1000Research* 6 (2017), ELIXIR–876. ISSN: 2046-1402. DOI: 10.12688/f1000research.11407.1.

[245] Luis Pedro Coelho. “For long-term sustainable software in bioinformatics”. In: *PLOS Computational Biology* 20.3 (Mar. 15, 2024). Publisher: Public Library of Science, e1011920. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011920.

[246] Paul P. Gardner et al. “Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software”. In: *Genome Biology* 23.1 (Feb. 16, 2022), p. 56. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02625-x.

[247] Mehran Karimzadeh and Michael M. Hoffman. “Top considerations for creating bioinformatics software documentation”. In: *Briefings in Bioinformatics* 19.4 (July 20, 2018), pp. 693–699. ISSN: 1477-4054. DOI: 10.1093/bib/bbw134.

[248] Torsten Seemann. “Ten recommendations for creating usable bioinformatics command line software”. In: *GigaScience* 2.1 (Nov. 13, 2013), p. 15. ISSN: 2047-217X. DOI: 10.1186/2047-217X-2-15.

[249] Mohammed Alser et al. “Packaging and containerization of computational methods”. In: *Nature Protocols* 19.9 (Sept. 2024). Publisher: Nature Publishing Group, pp. 2529–2539. ISSN: 1750-2799. DOI: 10.1038/s41596-024-00986-0.

[250] Björn Grüning et al. “Bioconda: sustainable and comprehensive software distribution for the life sciences”. In: *Nature Methods* 15.7 (July 2018), pp. 475–476. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0046-7.

[251] Bjorn Gruening et al. “Recommendations for the packaging and containerizing of bioinformatics software”. In: *F1000Research* 7 (2018), ISCB Comm J–742. ISSN: 2046-1402. DOI: 10.12688/f1000research.15140.2.

[252] Daniel Nüst et al. “Ten simple rules for writing Dockerfiles for reproducible data science”. In: *PLoS computational biology* 16.11 (Nov. 2020), e1008316. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008316.

[253] Carl Boettiger. “An introduction to Docker for reproducible research”. In: *SIGOPS Oper. Syst. Rev.* 49.1 (Jan. 20, 2015), pp. 71–79. ISSN: 0163-5980. DOI: 10.1145/2723872.2723882.

[254] Sabah Kadri et al. “Containers in Bioinformatics: Applications, Practical Considerations, and Best Practices in Molecular Pathology”. In: *The Journal of Molecular Diagnostics* 24.5 (May 1, 2022), pp. 442–454. ISSN: 1525-1578. DOI: 10.1016/j.jmoldx.2022.01.006.

[255] Laura Wratten, Andreas Wilm, and Jonathan Göke. “Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers”. In: *Nature Methods* 18.10 (Oct. 2021), pp. 1161–1168. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01254-9.

[256] Francesco Strozzi et al. “Scalable Workflows and Reproducible Data Analysis for Genomics”. In: *Methods in Molecular Biology (Clifton, N.J.)* 1910 (2019), pp. 723–745. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-9074-0_24.

[257] Matthew Krafczyk et al. “Scientific Tests and Continuous Integration Strategies to Enhance Reproducibility in the Scientific Software Context”. In: *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*. P-RECS ’19. New York, NY, USA: Association for Computing Machinery, June 17, 2019, pp. 23–28. ISBN: 978-1-4503-6756-1. DOI: 10.1145/3322790.3330595.

[258] Boas C.L. van der Putten et al. “Software testing in microbial bioinformatics: a call to action”. In: *Microbial Genomics* 8.3 (2022). Publisher: Microbiology Society, p. 000790. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000790.

[259] Nicholas J Dimonaco et al. “No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study”. In: *Bioinformatics* 38.5 (Mar. 1, 2022), pp. 1198–1207. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab827.

[260] Torsten Seemann. “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* 30.14 (July 15, 2014), pp. 2068–2069. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu153.

[261] Oliver Schwengers et al. “Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification”. In: *Microbial Genomics* 7.11 (2021). Publisher: Microbiology Society, p. 000685. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000685.

[262] Wenjun Li et al. “RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation”. In: *Nucleic Acids Research* 49 (D1 Jan. 8, 2021), pp. D1020–D1028. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa1105.

[263] Nick J. B. Isaac and Michael J. O. Pocock. “Bias and information in biological records”. In: *Biological Journal of the Linnean Society* 115.3 (July 1, 2015), pp. 522–531. ISSN: 0024-4066. DOI: 10.1111/bij.12532.

[264] Sage Albright and Stilianos Louca. “Trait biases in microbial reference genomes”. In: *Scientific Data* 10.1 (Feb. 9, 2023). Publisher: Nature Publishing Group, p. 84. ISSN: 2052-4463. DOI: 10.1038/s41597-023-01994-7.

[265] Michael G. Ross et al. “Characterizing and measuring bias in sequence data”. In: *Genome Biology* 14.5 (May 29, 2013), R51. ISSN: 1474-760X. DOI: 10.1186/gb-2013-14-5-r51.

[266] Stefan A. Boers, Ruud Jansen, and John P. Hays. “Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory”. In: *European Journal of Clinical Microbiology & Infectious Diseases* 38.6 (June 1, 2019), pp. 1059–1070. ISSN: 1435-4373. DOI: 10.1007/s10096-019-03520-3.

[267] Luis M. Rodriguez-R et al. “An ANI gap within bacterial species that advances the definitions of intra-species units”. In: *mBio* 15.1 (Dec. 12, 2023). Publisher: American Society for Microbiology, e02696–23. DOI: 10.1128/mbio.02696-23.

[268] Chirag Jain et al. “High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries”. In: *Nature Communications* 9.1 (Nov. 30, 2018). Publisher: Nature Publishing Group, p. 5114. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07641-9.

[269] Santiago Castillo-Ramírez. “On the road to genomically defining bacterial intra-species units”. In: *mSystems* 9.7 (), e00584–24. ISSN: 2379-5077. DOI: 10.1128/msystems.00584-24.

[270] Konstantinos T. Konstantinidis. “Sequence-discrete species for prokaryotes and other microbes: A historical perspective and pending issues”. In: *mLife* 2.4 (Dec. 2023), pp. 341–349. ISSN: 2770-100X. DOI: 10.1002/mlf2.12088.

[271] Ramon Rosselló-Móra and Rudolf Amann. “Past and future species definitions for Bacteria and Archaea”. In: *Systematic and Applied Microbiology* 38.4 (June 2015), pp. 209–216. ISSN: 1618-0984. DOI: 10.1016/j.syapm.2015.02.001.

[272] Connor S. Murray, Yingnan Gao, and Martin Wu. “Re-evaluating the evidence for a universal genetic boundary among microbial species”. In: *Nature Communications* 12 (July 7, 2021), p. 4059. ISSN: 2041-1723. DOI: 10.1038/s41467-021-24128-2.

[273] Mostafa Y. Abdel-Glil et al. “Core Genome Multilocus Sequence Typing Scheme for Improved Characterization and Epidemiological Surveillance of Pathogenic Brucella”. In: *Journal of Clinical Microbiology* 60.8 (Aug. 17, 2022), e0031122. ISSN: 1098-660X. DOI: 10.1128/jcm.00311-22.

[274] Julien Guglielmini et al. “Genus-wide Leptospira core genome multilocus sequence typing for strain taxonomy and global surveillance”. In: *PLOS Neglected Tropical Diseases* 13.4 (Apr. 26, 2019). Publisher: Public Library of Science, e0007374. ISSN: 1935-2735. DOI: 10.1371/journal.pntd.0007374.

[275] Madison E. Pearce et al. “A proposed core genome scheme for analyses of the Salmonella genus”. In: *Genomics* 112.1 (Jan. 2020), pp. 371–378. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2019.02.016.

[276] Cyril Savin et al. “Genus-wide Yersinia core-genome multilocus sequence typing for species identification and strain characterization”. In: *Microbial Genomics* 5.10 (2019). Publisher: Microbiology Society, e000301. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000301.

[277] Ouli Xie et al. “Inter-species gene flow drives ongoing evolution of Streptococcus pyogenes and Streptococcus dysgalactiae subsp. equisimilis”. In: *Nature Communications* 15.1 (Mar. 13, 2024). Publisher: Nature Publishing Group, p. 2286. ISSN: 2041-1723. DOI: 10.1038/s41467-024-46530-2.

[278] Johanna Dabernig-Heinz et al. “Core genome multilocus sequence typing (cgMLST) applicable to the monophyletic Klebsiella oxytoca species complex”. In: *Journal of Clinical Microbiology* 62.6 (May 23, 2024). Publisher: American Society for Microbiology, e01725–23. DOI: 10.1128/jcm.01725-23.

[279] Samuel D. Chorlton. “Ten common issues with reference sequence databases and how to mitigate them”. In: *Frontiers in Bioinformatics* 4 (Mar. 15, 2024). Publisher: Frontiers. ISSN: 2673-7647. DOI: 10.3389/fbinf.2024.1278228.

[280] Andrew J. McArdle and Myrsini Kaforou. “Sensitivity of shotgun metagenomics to host DNA: abundance estimates depend on bioinformatic tools and contamination is the main issue”. In: *Access Microbiology* 2.4 (2020). Publisher: Microbiology Society, e000104. ISSN: 2516-8290. DOI: 10.1099/acmi.0.000104.

[281] Yunyun Gao et al. “Benchmarking short-read metagenomics tools for removing host contamination”. In: *GigaScience* 14 (Jan. 1, 2025), giaf004. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaf004.

[282] Kumeren N. Govender and David W. Eyre. “Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications”. In: *Microbial Genomics* 8.10 (2022). Publisher: Microbiology Society, p. 000886. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000886.

[283] Bede Constantinides, Martin Hunt, and Derrick W Crook. “Hostile: accurate decontamination of microbial host sequences”. In: *Bioinformatics* 39.12 (Dec. 1, 2023), btad728. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad728.

[284] Craig Billington, Joanne M. Kingsbury, and Lucia Rivas. “Metagenomics Approaches for Improving Food Safety: A Review”. In: *Journal of Food Protection* 85.3 (Mar. 1, 2022), pp. 448–464. ISSN: 0362-028X. DOI: 10.4315/JFP-21-301.

[285] Sarah Buddle et al. “Evaluating metagenomics and targeted approaches for diagnosis and surveillance of viruses”. In: *Genome Medicine* 16.1 (Sept. 9, 2024), p. 111. ISSN: 1756-994X. DOI: 10.1186/s13073-024-01380-x.

[286] Jyoti Kant Chaudhari et al. “Biological big-data sources, problems of storage, computational issues, and applications: a comprehensive review”. In: *Knowledge and Information Systems* 66.6 (June 1, 2024), pp. 3159–3209. ISSN: 0219-3116. DOI: 10.1007/s10115-023-02049-4.

[287] Raphael O. Betschart et al. “A benchmark study of compression software for human short-read sequence data”. In: *Scientific Reports* 15.1 (May 2, 2025). Publisher: Nature Publishing Group, p. 15358. ISSN: 2045-2322. DOI: 10.1038/s41598-025-00491-8.

[288] Foad Nazari et al. “Lossless and reference-free compression of FASTQ/A files using GeneSqueeze”. In: *Scientific Reports* 15.1 (Jan. 2, 2025). Publisher: Nature Publishing Group, p. 322. ISSN: 2045-2322. DOI: 10.1038/s41598-024-79258-6.

[289] Karel Břinda et al. “Efficient and robust search of microbial genomes via phylogenetic compression”. In: *Nature Methods* 22.4 (Apr. 2025). Publisher: Nature Publishing Group, pp. 692–697. ISSN: 1548-7105. DOI: 10.1038/s41592-025-02625-2.

[290] Andrzej Zielezinski et al. “Alignment-free sequence comparison: benefits, applications, and tools”. In: *Genome Biology* 18.1 (Oct. 3, 2017), p. 186. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1319-7.

[291] Andrzej Zielezinski et al. “Benchmarking of alignment-free sequence comparison methods”. In: *Genome Biology* 20.1 (July 25, 2019), p. 144. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1755-7.

[292] A. A. Schäffer et al. “Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements”. In: *Nucleic Acids Research* 29.14 (July 15, 2001), pp. 2994–3005. ISSN: 1362-4962. DOI: 10.1093/nar/29.14.2994.

[293] Yi-Kuo Yu and Stephen F. Altschul. “The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions”. In: *Bioinformatics* 21.7 (Apr. 1, 2005), pp. 902–911. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bti070.

[294] Martin T Swain and Martin Vickers. “Interpreting alignment-free sequence comparison: what makes a score a good score?” In: *NAR Genomics and Bioinformatics* 4.3 (Sept. 1, 2022), lqac062. ISSN: 2631-9268. DOI: 10.1093/nargab/lqac062.

[295] Karel Břinda, Maciej Sykulski, and Gregory Kucherov. “Spaced seeds improve k-mer-based metagenomic classification”. In: *Bioinformatics* 31.22 (Nov. 15, 2015), pp. 3584–3592. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv419.

[296] Hartmut Häntze and Paul Horton. "Effects of spaced k-mers on alignment-free genotyping". In: *Bioinformatics* 39 (Supplement_1 June 1, 2023), pp. i213–i221. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad202.

[297] Robert Edgar. "Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences". In: *PeerJ* 9 (Feb. 5, 2021), e10805. ISSN: 2167-8359. DOI: 10.7717/peerj.10805.

[298] Abhinav Dutta, David Pellow, and Ron Shamir. "Parameterized syncmer schemes improve long-read mapping". In: *PLOS Computational Biology* 18.10 (Oct. 28, 2022). Publisher: Public Library of Science, e1010638. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010638.

[299] Camille Moeckel et al. "A survey of k-mer methods and applications in bioinformatics". In: *Computational and Structural Biotechnology Journal* 23 (Dec. 1, 2024), pp. 2289–2303. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2024.05.025.

[300] Mohammed Alser et al. "Technology dictates algorithms: recent developments in read alignment". In: *Genome Biology* 22.1 (Aug. 26, 2021), p. 249. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02443-7.

[301] Md. Vasimuddin et al. "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems". In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). ISSN: 1530-2075. May 2019, pp. 314–324. DOI: 10.1109/IPDPS.2019.00041.

[302] Heng Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (Sept. 15, 2018), pp. 3094–3100. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty191.

[303] Kristoffer Sahlin et al. "A survey of mapping algorithms in the long-reads era". In: *Genome Biology* 24.1 (June 1, 2023), p. 133. ISSN: 1474-760X. DOI: 10.1186/s13059-023-02972-3.

[304] Jeffry M. Gaston, Eric J. Alm, and An-Ni Zhang. "X-Mapper: fast and accurate sequence alignment via gapped x-mers". In: *Genome Biology* 26.1 (Jan. 22, 2025), p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03473-7.

[305] Philip T. L. C. Clausen, Frank M. Aarestrup, and Ole Lund. "Rapid and precise alignment of raw reads against redundant databases with KMA". In: *BMC Bioinformatics* 19.1 (Aug. 29, 2018), p. 307. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2336-6.

[306] Peter Menzel, Kim Lee Ng, and Anders Krogh. "Fast and sensitive taxonomic classification for metagenomics with Kaiju". In: *Nature Communications* 7.1 (Apr. 13, 2016). Publisher: Nature Publishing Group, p. 11257. ISSN: 2041-1723. DOI: 10.1038/ncomms11257.

[307] Li Song and Ben Langmead. "Centrifuger: lossless compression of microbial genomes for efficient and accurate metagenomic sequence classification". In: *Genome Biology* 25.1 (Apr. 25, 2024), p. 106. ISSN: 1474-760X. DOI: 10.1186/s13059-024-03244-4.

[308] Daehwan Kim et al. "Centrifuge: rapid and sensitive classification of metagenomic sequences". In: *Genome Research* 26.12 (Jan. 12, 2016). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1721–1729. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.210641.116.

[309] Daniel M. Portik, C. Titus Brown, and N. Tessa Pierce-Ward. "Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets". In: *BMC Bioinformatics* 23.1 (Dec. 13, 2022), p. 541. ISSN: 1471-2105. DOI: 10.1186/s12859-022-05103-0.

[310] Jennifer Lu et al. "Metagenome analysis using the Kraken software suite". In: *Nature Protocols* 17.12 (Dec. 2022). Publisher: Nature Publishing Group, pp. 2815–2839. ISSN: 1750-2799. DOI: 10.1038/s41596-022-00738-y.

[311] Haitao Han, Ziye Wang, and Shanfeng Zhu. "Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes". In: *Nature Communications* 16.1 (Mar. 24, 2025). Publisher: Nature Publishing Group, p. 2865. ISSN: 2041-1723. DOI: 10.1038/s41467-025-57957-6.

[312] Yu-Wei Wu, Blake A. Simmons, and Steven W. Singer. "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets". In: *Bioinformatics* 32.4 (Feb. 1, 2016), pp. 605–607. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv638.

[313] Shaojun Pan, Xing-Ming Zhao, and Luis Pedro Coelho. "SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing". In: *Bioinformatics* 39 (Supplement_1 June 1, 2023), pp. i21–i29. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad209.

[314] Zhemin Zhou, Jane Charlesworth, and Mark Achtman. "Accurate reconstruction of bacterial pan- and core genomes with PEPPAN". In: *Genome Research* 30.11 (Jan. 11, 2020). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1667–1679. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.260828.120.

[315] Ivan Borozan, Stuart Watt, and Vincent Ferretti. "Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification". In: *Bioinformatics* 31.9 (May 1, 2015), pp. 1396–1404. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv006.

[316] Hagay Enav, Inbal Paz, and Ruth E. Ley. "Strain tracking in complex microbiomes using synteny analysis reveals per-species modes of evolution". In: *Nature Biotechnology* 43.5 (May 2025). Publisher: Nature Publishing Group, pp. 773–783. ISSN: 1546-1696. DOI: 10.1038/s41587-024-02276-2.

[317] Matthew P. Moore et al. "KmerAperture: Retaining k-mer synteny for alignment-free extraction of core and accessory differences between bacterial genomes". In: *PLOS Genetics* 20.4 (Apr. 29, 2024). Publisher: Public Library of Science, e1011184. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1011184.

[318] Zhemin Zhou et al. "GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens". In: *Genome Research* 28.9 (Jan. 9, 2018). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1395–1404. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.232397.117.

[319] Verónica Mixão et al. "ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data". In: *Genome Medicine* 15.1 (June 15, 2023), p. 43. ISSN: 1756-994X. DOI: 10.1186/s13073-023-01196-1.

[320] Andrea De Ruvo et al. "SPREAD: Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard". In: *Veterinaria Italiana* 60.4 (Dec. 31, 2024). ISSN: 1828-1427. DOI: 10.12834/VetIt.3476.23846.1.

[321] Alexander Klimka et al. "Epitope-specific immunity against Staphylococcus aureus coproporphyrinogen III oxidase". In: *npj Vaccines* 6.1 (Jan. 18, 2021). Publisher: Nature Publishing Group, pp. 1–12. ISSN: 2059-0105. DOI: 10.1038/s41541-020-00268-2.

[322] Vimbai Irene Machimbirike et al. "Comparative genomics of *Edwardsiella ictaluri* revealed four distinct host-specific genotypes and thirteen potential vaccine candidates". In: *Genomics* 113.4 (July 1, 2021), pp. 1976–1987. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2021.04.016.

[323] Paul Stothard, Jason R Grant, and Gary Van Domselaar. "Visualizing and comparing circular genomes using the CGView family of tools". In: *Briefings in Bioinformatics* 20.4 (July 19, 2019), pp. 1576–1582. ISSN: 1477-4054. DOI: 10.1093/bib/bbx081.

[324] Nigel P Dyer et al. "EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens". In: *Nucleic Acids Research* 53 (D1 Jan. 6, 2025), pp. D757–D762. ISSN: 1362-4962. DOI: 10.1093/nar/gkae902.

[325] Nicolas J. Tourasse et al. "Core genome multilocus sequence typing scheme for *Bacillus cereus* group bacteria". In: *Research in Microbiology*. Insights into the Bacillus anthracis cereus thuringiensis 2022 conference 174.6 (July 1, 2023), p. 104050. ISSN: 0923-2508. DOI: 10.1016/j.resmic.2023.104050.

[326] Romário Oliveira de Sales et al. "A Core Genome Multilocus Sequence Typing Scheme for Pseudomonas aeruginosa". In: *Frontiers in Microbiology* 11 (May 26, 2020). Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2020.01049.

[327] Yizhak Hershko et al. "Construction of core genome multi-locus sequence typing schemes for population structure analyses of *Nocardia* species". In: *Research in Microbiology* 175.8 (Nov. 1, 2024), p. 104246. ISSN: 0923-2508. DOI: 10.1016/j.resmic.2024.104246.

[328] Sofia Kozak et al. "Core genome multilocus sequence typing schemes for epidemiological investigation of *Taylorella equigenitalis* and *Taylorella asinigenitalis*". In: *Veterinary Microbiology* 302 (Mar. 1, 2025), p. 110419. ISSN: 0378-1135. DOI: 10.1016/j.vetmic.2025.110419.

[329] Chiara Crestani et al. "Microevolution and genomic epidemiology of the diphtheria-causing zoonotic pathogen Corynebacterium ulcerans". In: *Nature Communications* 16 (May 24, 2025), p. 4843. ISSN: 2041-1723. DOI: 10.1038/s41467-025-60065-0.

[330] Sofia Carneiro et al. "Genome-Scale Characterization of Mycobacterium abscessus Complex Isolates from Portugal". In: *International Journal of Molecular Sciences* 24.20 (Oct. 20, 2023), p. 15402. ISSN: 1422-0067. DOI: 10.3390/ijms242015402.

[331] Kaisa Thorell et al. "The Helicobacter pylori Genome Project: insights into H. pylori population structure from analysis of a worldwide collection of complete genomes". In: *Nature Communications* 14.1 (Dec. 11, 2023). Publisher: Nature Publishing Group, p. 8184. ISSN: 2041-1723. DOI: 10.1038/s41467-023-43562-y.

[332] Yu Feng et al. "Population genomics uncovers global distribution, antimicrobial resistance, and virulence genes of the opportunistic pathogen *Klebsiella aerogenes*". In: *Cell Reports* 43.8 (Aug. 27, 2024), p. 114602. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2024.114602.

[333] Melissa J. Jansen van Rensburg et al. "Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure". In: *Microbial Genomics* 10.8 (2024). Publisher: Microbiology Society, p. 001280. ISSN: 2057-5858. DOI: 10.1099/mgen.0.001280.

[334] Duccio Medini et al. "The microbial pan-genome". In: *Current Opinion in Genetics & Development*. Genomes and evolution 15.6 (Dec. 1, 2005), pp. 589–594. ISSN: 0959-437X. DOI: 10.1016/j.gde.2005.09.006.

[335] Hervé Tettelin et al. "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"". In: *Proceedings of the National Academy of Sciences* 102.39 (Sept. 27, 2005). Publisher: Proceedings of the National Academy of Sciences, pp. 13950–13955. DOI: 10.1073/pnas.0506758102.

[336] Sávio Souza Costa et al. “First Steps in the Analysis of Prokaryotic Pan-Genomes”. In: *Bioinformatics and Biology Insights* 14 (Aug. 7, 2020), p. 1177932220938064. ISSN: 1177-9322. DOI: 10.1177/1177932220938064.

[337] Déborah Merda et al. “Unraveling the impact of genome assembly on bacterial typing: a one health perspective”. In: *BMC Genomics* 25.1 (Nov. 8, 2024), p. 1059. ISSN: 1471-2164. DOI: 10.1186/s12864-024-10982-z.

[338] Ryan R. Wick et al. “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads”. In: *PLOS Computational Biology* 13.6 (June 8, 2017). Publisher: Public Library of Science, e1005595. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005595.

[339] Clémentine Henri et al. “An Assessment of Different Genomic Approaches for Inferring Phylogeny of Listeria monocytogenes”. In: *Frontiers in Microbiology* 8 (Nov. 29, 2017). Publisher: Frontiers. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02351.

[340] Madison E. Pearce et al. “Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak”. In: *International Journal of Food Microbiology* 274 (June 2, 2018), pp. 1–11. ISSN: 1879-3460. DOI: 10.1016/j.ijfoodmicro.2018.02.023.

[341] Caroline Vincent et al. “Comparison of advanced whole genome sequence-based methods to distinguish strains of Salmonella enterica serovar Heidelberg involved in foodborne outbreaks in Québec”. In: *Food Microbiology* 73 (Aug. 2018), pp. 99–110. ISSN: 1095-9998. DOI: 10.1016/j.fm.2018.01.004.

[342] Dominique S. Blanc et al. “Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa”. In: *Frontiers in Microbiology* 11 (July 22, 2020), p. 1729. ISSN: 1664-302X. DOI: 10.3389/fmicb.2020.01729.

[343] A. Baktash et al. “Comparison of Whole-Genome Sequence-Based Methods and PCR Ribotyping for Subtyping of Clostridioides difficile”. In: *Journal of Clinical Microbiology* 60.2 (), e01737–21. ISSN: 0095-1137. DOI: 10.1128/jcm.01737-21.

[344] Kathryn E. Holt et al. “Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health”. In: *Proceedings of the National Academy of Sciences* 112.27 (July 7, 2015). Publisher: Proceedings of the National Academy of Sciences, E3574–E3581. DOI: 10.1073/pnas.1501049112.

[345] Robert W Jackson et al. “The influence of the accessory genome on bacterial pathogen evolution”. In: *Mobile Genetic Elements* 1.1 (2011), pp. 55–65. ISSN: 2159-2543. DOI: 10.4161/mge.1.1.16432.

[346] Helena Darmancier et al. “Are Virulence and Antibiotic Resistance Genes Linked? A Comprehensive Analysis of Bacterial Chromosomes and Plasmids”. In: *Antibiotics* 11.6 (June 2022). Number: 6 Publisher: Multidisciplinary Digital Publishing Institute, p. 706. ISSN: 2079-6382. DOI: 10.3390/antibiotics11060706.

[347] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (Aug. 2021). Publisher: Nature Publishing Group, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

[348] Josh Abramson et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. In: *Nature* 630.8016 (June 2024). Publisher: Nature Publishing Group, pp. 493–500. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07487-w.

[349] Guillaume Holley and Páll Melsted. “Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs”. In: *Genome Biology* 21.1 (Sept. 17, 2020), p. 249. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02135-8.

[350] Joshua D. Harling-Lee et al. “A graph-based approach for the visualisation and analysis of bacterial pangenomes”. In: *BMC Bioinformatics* 23.1 (Oct. 8, 2022), p. 416. ISSN: 1471-2105. DOI: 10.1186/s12859-022-04898-2.

[351] Nicholas Noll et al. “PanGraph: scalable bacterial pan-genome graph construction”. In: *Microbial Genomics* 9.6 (2023). Publisher: Microbiology Society, p. 001034. ISSN: 2057-5858. DOI: 10.1099/mgen.0.001034.

[352] Lauren Maxwell et al. “FAIR, ethical, and coordinated data sharing for COVID-19 response: a scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries”. In: *The Lancet. Digital Health* 5.10 (Sept. 27, 2023), e712–e736. ISSN: 2589-7500. DOI: 10.1016/S2589-7500(23)00129-2.

[353] Emma J Griffiths et al. “Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package”. In: *GigaScience* 11 (Jan. 1, 2022), giac003. ISSN: 2047-217X. DOI: 10.1093/gigascience/giac003.

Appendix

D660–D666 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 17 October 2020*
doi: 10.1093/nar/gkaa889

Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas

Rafael Mamede†, Pedro Vila-Cerqueira†, Mickael Silva, João A. Carriço and Mário Ramirez*

Instituto de Microbiologia and Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal

Received August 07, 2020; Revised September 18, 2020; Editorial Decision September 28, 2020; Accepted October 02, 2020

ABSTRACT

Chewie Nomenclature Server (chewie-NS, https://chewbbaca.online/) allows users to share genome-based gene-by-gene typing schemas and to maintain a common nomenclature, simplifying the comparison of results. The combination between local analyses and a public repository of allelic data strikes a balance between potential confidentiality issues and the need to compare results. The possibility of deploying private instances of chewie-NS facilitates the creation of nomenclature servers with a restricted user base to allow compliance with the strictest data policies. Chewie-NS allows users to easily share their own schemas and to explore publicly available schemas, including informative statistics on schemas and loci presented in interactive charts and tables. Users can retrieve all the information necessary to run a schema locally or all the alleles identified at a particular locus. The integration with the chewBBACA suite enables users to directly upload new schemas to chewie-NS, download existing schemas and synchronize local and remote schemas from chewBBACA command line version, allowing an easier integration into high-throughput analysis pipelines. The same REST API linking chewie-NS and the chewBBACA suite supports the interaction of other interfaces or pipelines with the databases available at chewie-NS, facilitating the reusability of the stored data.

INTRODUCTION

The importance of distinguishing strains within the same microbial species has been proven critical for identifying chains of transmission and understanding pathogen evolution, as recently illustrated by the SARS-CoV-2 pandemic (1,2). The advent and widespread adoption of high-throughput sequencing allowed leveraging genomic information for this purpose (1,2). One of the most common approaches in bacterial typing is gene-by-gene methods, which extend the concept of multilocus sequence typing (MLST) to include all genes present in the core genome of a given species (cgMLST) or, trying to cover a significant fraction of a species' pan-genome, in whole genome (wgMLST) (3). Current software approaches implementing these wg/cgMLST typing methods suffer from standardization issues when comparing results between different tools and between different laboratories or users (4).

We have previously developed a suite, chewBBACA (5), allowing the creation of gene-by-gene schemas and performing allele calls on assembled draft genomes. Since chewBBACA was designed to perform local analysis to address concerns over data privacy and scalability, it has the drawback that small adjustments in parameters may lead to inconsistencies between runs. Moreover, the software allows users to create their own wg/cgMLST schemas but currently no tool is available for the easy sharing of schemas, which potentially hampers long-term and multinational studies, as well as the reusability of already published schemas (6,7).

There are well-established websites for performing gene-by-gene analyses, such as PubMLST (https://pubmlst.org/) (8) and EnteroBase (https://enterobase.warwick.ac.uk/) (9), that centralize analysis and hosting of public and private schemas. chewBBACA does not depend on a web server and by enabling local analyses and schema creation allows for scalable and private analyses of genomes, but the existing implementation lacked an easy way to share schemas and the associated allelic information, which is possible in a centralized solution.

In order to allow users to share gene-by-gene typing schemas and for a common allelic nomenclature to be maintained (10), we developed chewie-NS, a nomenclature server based on the TypOn ontology (11) offering a web interface that also integrates directly with local instances of chewB-

*To whom correspondence should be addressed. Tel: +351 217999460; Fax: +351 217999459; Email: ramirez@fm.ul.pt
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

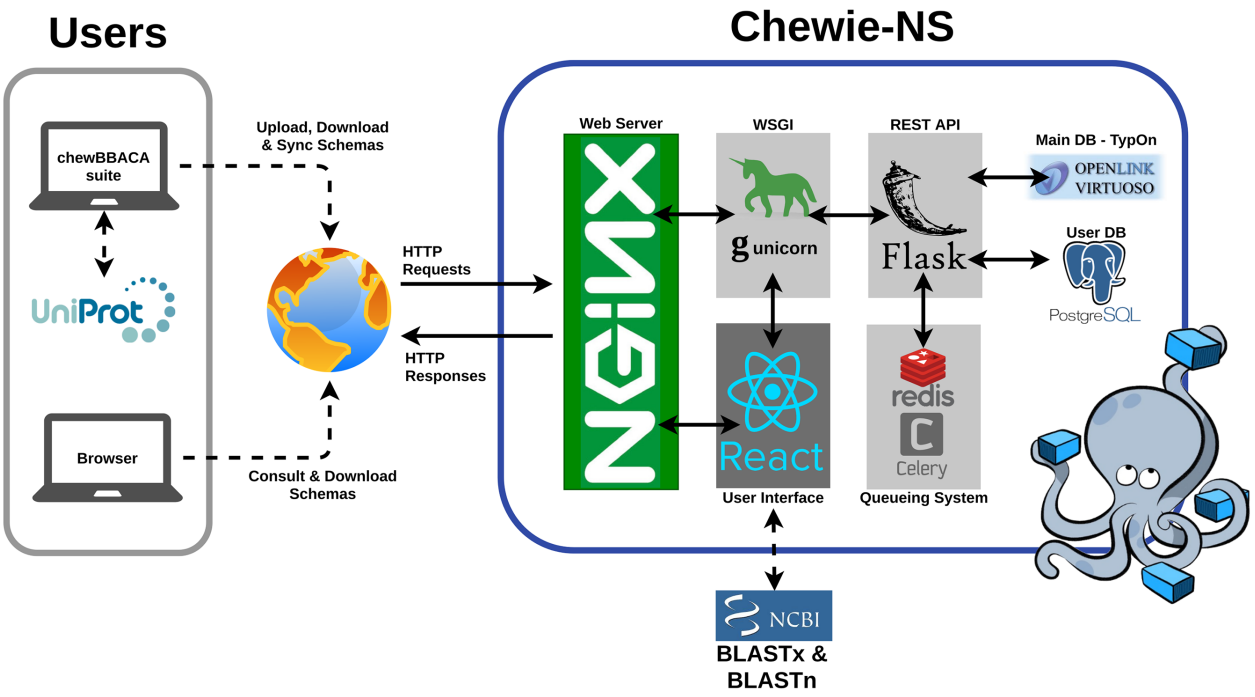


Figure 1. The chewie-NS service: global overview of the technologies used and API connectivity.

BACA and can be programmatically accessed by external resources. Chewie-NS aims to complement the private local analysis of strains by also allowing the simple communication of results while providing an interface for users to easily explore the allelic diversity within species. The importance of the latter is becoming increasingly clear with the recognition that bacterial phenotypes can be profoundly altered by allelic variants (12,13). Other publicly available web services require submission of raw data, something that may raise privacy and ownership concerns, while our approach of enabling local analyses is more flexible and scalable, and respects data privacy concerns. Current wg/cgMLST typing methods suffer from standardization difficulties or issues that manifest not only when trying to reconcile results from different tools, but also when the same tool is run at different times with small adjustments in parameter values or database modifications that may lead to inconsistencies. This is an even more complex problem than it was for classical MLST methods (14). Having a repository of schemas, their associated parameters and the allelic diversity identified will allow the consistent use of gene-by-gene typing schemas by different groups and to build upon the results of different studies to monitor microbial populations and study outbreaks.

Chewie-NS is available at https://chewbbaca.online and its source code is available at https://github.com/B-UMMI/Chewie-NS. Detailed documentation, including a descriptive tutorial on how to deploy and use the server, can be found at https://chewie-ns.readthedocs.io. Additionally, a tutorial version of the server aiming at familiarizing users with the integration between the chewBBACA suite and chewie-NS, which allows users to perform mock submissions of schemas and synchronizations without the need to register and with a much reduced database, is available at https://tutorial.chewbbaca.online/.

DATABASE CREATION

Backend

The architecture of chewie-NS is shown schematically in Figure 1. The backend component of chewie-NS makes use of the Virtuoso triple store (v. 7.2.6) (https://virtuoso.openlinksw.com/). This database management system allows the integration of a Resource Description Framework to implement the TypOn ontology (11) structure to store schema data. Additionally, a PostgreSQL database (v. 10) (https://www.postgresql.org/) was adopted for user management. These databases are accessible through a Python 3 REST API developed in the Flask (v. 1.1.0) (https://flask.palletsprojects.com/en/1.1.x/) web development microframework, which allows requests through defined endpoints and facilitates programmatic access to the nomenclature server. Requests and HTTPS connections are handled by a web server, NGINX (v. 1.17) (https://www.nginx.com/), that communicates with Gunicorn (v. 20.0.4) (https://gunicorn.org/), a WSGI application server capable of running multiple processes of the web application and distributing incoming requests to ensure scalability and load balancing. A queueing system was implemented to manage all tasks with possible concurrent user access through Redis (v. 5.0.6) (https://redis.io/) and Celery (v. 4.4.0rc2) (https://docs.celeryproject.org/en/stable/getting-started/introduction.html).

Frontend

The user interface (UI) for chewie-NS was built with the JavaScript frameworks React (v. 16.12.0) (https://reactjs.org/) and Material-UI (v. 4.9.14) (https://material-ui.com/). The UI provides a list of available schemas and displays relevant schema and locus statistics in a responsive and in-

Figure 2. The schemas overview page of chewie-NS.

teractive manner. Access to daily updated compressed files of the schemas for download and local use is also provided. All interactive charts were rendered with the graph visualization library Plotly.js (v. 1.52.1) (https://plotly.com/javascript/) through its React component, react-plotly (v. 2.4.0) (https://plotly.com/javascript/react/).

Chewie-NS usage

Local installation. Deployment of local instances can be easily achieved through Docker Compose (https://www.docker.com/) (available at https://github.com/B-UMMI/Chewie-NS). The use of a container orchestrator (https://docs.docker.com/compose/) supports the easy deployment of local instances independently of the hardware available, allowing the creation of private trusted databases if public access is not possible. Instructions on how to achieve this can be found at https://github.com/B-UMMI/Chewie-NS. This can be particularly important for national public health institutions in the context of restrictive or ambiguous data sharing laws because it allows stricter user access control.

Application programming interface. A RESTful API also referred to as a RESTful web service or REST API, i.e. based on representational state transfer (REST), is available. The user can interact with chewie-NS's API through the web interface, by clicking on the 'API' button on the menu. This will open a page with Swagger UI (https://swagger.io/tools/swagger-ui/), a user-friendly tool for the user to interact directly with the REST API. Programmatic access is also possible through command line applications such as curl or tools such as Postman (https://www.postman.com/). Chewie-NS's REST API allows interaction with the PostgreSQL database to manage user registrations on local instances. Through the API, users are also able to query the Virtuoso database to download compressed schemas, search for specific alleles and query data about specific species, loci or alleles.

Web interface

Schemas overview. A table summarizes the species and number of schemas available for each species in chewie-NS. Selecting a species leads to another table (Figure 2) with a list of relevant information about each available schema, namely the schema internal identifier, the user provided schema name, the username of the creator, the number of loci in the schema, the number of alleles, the software and version used to create the schema, the date of creation, the date of the last modification, the BLAST (15) score ratio selected, the translation table used, the minimum locus length and size threshold. In the table, there is a link to download the compressed file of the schema and the training file used to create it, both necessary to use the schema locally with the chewBBACA suite. Each table entry has also a link to a page containing more details about the schema. Below this table, an interactive bar chart displays the number of alleles per locus for each schema. The user can zoom in on the chart to obtain a better view of a given set of loci and can click on a bar to go to a page with more details on that particular locus.

Schema details. The schema evaluation and annotation page contains a description of the schema provided by the schema creator. During the schema upload, this information can be provided in a file using markdown, a simple plain-text-formatting syntax that allows the easy integration of hyperlinks, tables and images, allowing for a rich use of data for the description of the schema. Below this table are four charts in different tabs. Two charts (Figure 3) display characteristics of the schema: the distribution of the number of alleles per locus and of locus size. Two interactive charts represent for each locus its size summary statistics versus the number of alleles, and another a box plot of the size distribution of each locus. In all charts, the user can zoom in on particular regions for more detailed inspection and, on the latter two, clicking on the chart element opens a page with more information on that particular locus. Below the charts is a table of all the loci in the schema, including relevant information for each locus. This table, as all other tables of chewie-NS, is searchable, facilitating finding loci with particular characteristics (Figure 4). Similarly to other tables, the table can also be exported in comma-separated values format.

Locus details. Already in the schema evaluation and annotation page is shown most of the information of each locus. This includes the internal locus identification and label, the automated annotation created by chewBBACA including a link to the relevant UniProt (13) page, a user locus name and user custom annotation (supporting markdown syntax), number of alleles and allele size information. The possibility of schema creators offering their own annotation allows for domain-specific information to be added to the schema, including potentially richer complementary data and links to relevant external resources. Two charts are offered, one summarizing the size distribution of the alleles (frequency of binned sizes) and the other representing the sizes of each allele. Direct links to perform BLAST searches using allele 1 of that locus (BLASTn and BLASTx) are available at the bottom of the page, allowing the user to check for similarities that could offer further insights into the origin or likely function of the protein potentially encoded by the locus. A multi-fasta file containing the alleles of the locus can be downloaded by the users from this page. On a different page, a simple query feature allows finding exact matches to specific sequences stored in any of chewie-NS's databases, returning the loci and schemas where the sequence is found.

Integration with the chewBBACA suite and use of the API. By taking advantage of chewie-NS's API, chewBBACA is capable of handling not only the schema creation, but also its upload, synchronization and download. Users of chewBBACA registered in chewie-NS and with contributor privileges will be able to automatically upload a novel schema, making it available in chewie-NS. Any authorized registered user can also contribute novel alleles identified in local analyses to chewie-NS, contributing to the incremental development of the schema. This involves only allele information without the need to share a complete allelic profile with chewie-NS. On the other hand, one does not have to be registered to download any of the data stored in chewie-NS, including the compressed schemas or the novel alleles submitted to the chewie-NS database and that were not present on the compressed file to update the local schema. Detailed instructions on the chewBBACA commands to achieve this can be found at https://chewie-ns.readthedocs.io/en/latest/user/chewbbaca.html.

If a user creates a schema through a software other than chewBBACA, the API can still be used to submit this new schema to chewie-NS or to add novel alleles to non-chewBBACA schemas. A user would have to register with chewie-NS and would have to take on the responsibility for making the correct API calls for schema submission. Furthermore, it would be up to each user submitting new alleles to guarantee the consistency of these novel alleles with the schema originally deposited. These functions are handled transparently by chewBBACA in its interaction with chewie-NS. Anyone can use the API to download any of the schemas deposited in chewie-NS to be run locally with chewBBACA or any other allele-calling algorithm. These multiple ways in which a user can interact with chewie-NS allows tailoring the sharing of information to user preferences or to restrictions imposed on particular users.

In order to facilitate familiarization of the interaction between chewBBACA and chewie-NS, a tutorial website was created (https://tutorial.chewbbaca.online/), together with step-by-step instructions on how to perform mock operations with a small size schema (https://chewie-ns.readthedocs.io/en/latest/user/tutorial.html). This allows users to perform submissions and synchronization of schemas without the need for registering. The schemas submitted to the tutorial site are not permanent and are removed automatically 48 h after creation.

For reproducibility and traceability purposes, a feature to retrieve database snapshots at specific dates is available, allowing a user to be able to recover the exact schema, as it was available on a given date. Full documentation of the schemas allows for traceability, which is critical in public health applications. These various options will continue to allow data privacy, while striving for a common nomenclature. The detailed parameterization associated with each schema created with chewBBACA and the consistency checks implemented in chewie-NS mean that no human curation is necessary after the schema creation step, contributing to the rapid update of the database and exchange of information. However, although chewie-NS can be used to store and retrieve information of schemas not created with chewBBACA, it does not currently automatically guarantee the consistency of newly submitted alleles since each allele-calling algorithm will have specific parametrization requirements. Nevertheless, these can be implemented in the future as other allele-calling algorithms make use of the chewie-NS platform.

DISCUSSION

Chewie-NS accomplishes four important goals. First, it stores all the information required to define a chewBBACA schema, facilitating accessibility of schemas so that different schemas can be easily compared and evaluated. Second, it maintains a public compendium of the variability in each locus. Since chewBBACA loci are open-reading frames

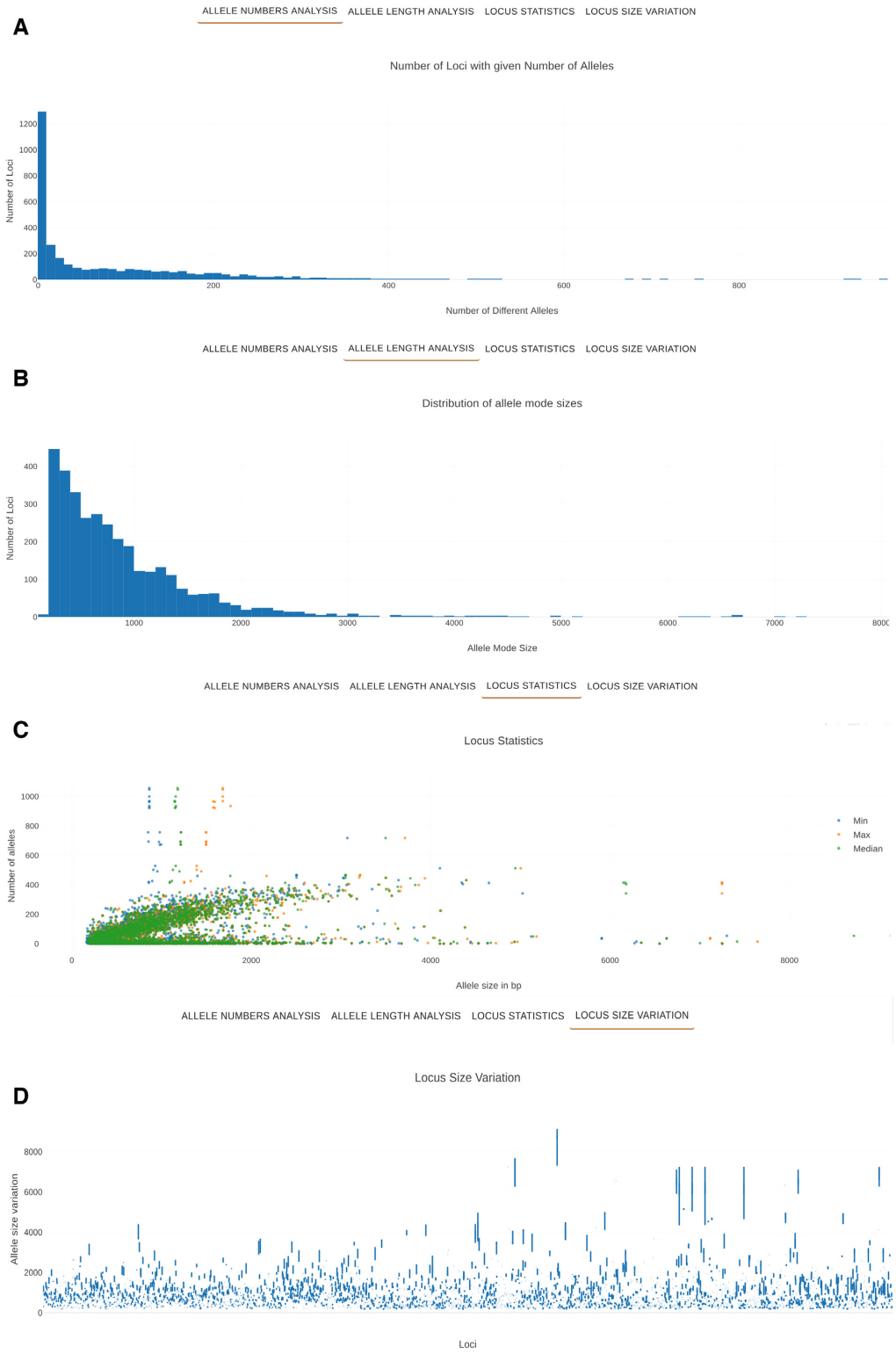


Figure 3. Summary charts displaying relevant information on a given schema. (**A**) Distribution of loci by number of alleles. (**B**) Distribution of loci by allele mode size. (**C**) Representation of summary statistics (minimum allele size in blue, maximum allele size in orange and median allele size in green) for each locus. (**D**) Box plots of loci size distribution; the loci in the *x*-axis are ordered by locus ID.

Annotations

speB speI speJ speQ speR

Uniprot Label	Uniprot URI	User locus name	Custom Annotation	Locus ID	Locus Label	Total Number of Alleles	Alleles Mode	Size Range (bp)
Streptococcal cysteine protease (Streptopain) / Streptococcal pyrogenic exotoxin B (SpeB)	http://purl.uniprot.org/uniparc/UPI000181F3F3	speB	Cysteine protease exotoxin SpeB (https://doi.org/10.1111/j.1365-2958.2011.07709.x; https://doi.org/10.1515/BC.2011.208)	86156	wgMLST-00086156	204	1197	966-1254
Exotoxin J	http://purl.uniprot.org/uniprot/Q7BAE3	speJ	Streptococcal pyrogenic exotoxin J (https://doi.org/10.1016/j.molmed.2013.10.004)	86338	wgMLST-00086338	56	699	636-711
N/A	N/A	speI	Streptococcal pyrogenic exotoxin I (https://doi.org/10.1016/j.molmed.2013.10.004)	87087	wgMLST-00087087	10	780	780-780
exotoxin	http://purl.uniprot.org/uniparc/UPI000DA4198B	speR	Streptococcal pyrogenic exotoxin SpeR (https://doi.org/10.1016/j.jinf.2019.02.005)	87813	wgMLST-00087813	8	675	675-675
Pyrogenic exotoxin K	http://purl.uniprot.org/uniparc/UPI0002D8ABCA	speQ	Streptococcal pyrogenic exotoxin SpeQ (https://doi.org/10.1016/j.jinf.2019.02.005)	88901	wgMLST-00088901	43	780	657-822

Figure 4. Schema loci table: search functionality. The figure presents the results of filtering for *speB, speI, speJ, speQ* and *speR* in the *user locus name* field of the *Streptococcus pyogenes* schema 1. Note that the user provided annotations complement and correct some of the annotations retrieved from UniProt.

(ORFs), this will allow monitoring the variability of the proteins potentially encoded by these loci. This is important when studying microbial pathogens because small variations in sequence can lead to dramatic changes in virulence (16) or antimicrobial resistance (12). On the other hand, allelic diversity can also be indicative of stabilizing or diversifying selective pressures, which in turn can be leveraged to obtain insights into pathogen evolution or interaction with the host (17). Allelic diversity is also important in reverse vaccinology (18) and to monitor the continued potential efficacy of some available vaccines (19). Third, through its integration with chewBBACA, it offers a simplified way for the user to control the flow of information between the local instance and chewie-NS. This is important to keep the local instance of chewBBACA updated with the current common nomenclature stored in chewie-NS databases and to contribute new alleles to the common databases, but it also allows for limited sharing of data to comply with any regulations the user may be operating under. Finally, the proposed workflow hopes to stimulate and facilitate data sharing between users using the same schemas, allowing for a faster detection of strain similarity, therefore contributing for genomic epidemiology studies and also faster outbreak detection and investigation by expediting strain comparison between different laboratories or institutions. Although the current integration of chewBBACA with chewie-NS facilitates the user interaction with chewie-NS when using this allele-calling software, the API can be exploited by other allele-calling software to also interact directly with chewie-NS, allowing an easier submission of schemas and of new alleles to schemas not created with chewBBACA.

The definition of a wg/cgMLST schema involves not only the choice of the target loci and of what constitutes a locus [for instance, an ORF as defined by Prodigal (20) in the case of chewBBACA, or a fragment of DNA between two primers in traditional MLST], but also the algorithm and parameters of the allele-calling software. If one uses the same set of target loci defined in the same way, but a different allele-calling software, one cannot guarantee that the alleles called for a given isolate would be the same as with another allele-calling software, a problem that can potentially become all the more acute with the identification of novel alleles. Even when using the same allele-calling software, if the parameters used are different, the alleles identified in a given isolate may also be different. Moreover, the addition of novel alleles to a schema that do not conform with the parameters defined initially may have hard to anticipate consequences on the subsequent allele-calling processes. Upstream of this, we would like to stress the importance of the use of shared assembly pipelines, to ensure that the deposited allele sequences are determined based on standardized procedures, as it has been shown that different assemblers can result in different variants in the assembly (21) and that this variability can introduce artificial allelic variability, even when using the same schema and allele-calling software.

The possibility of setting up local instances of chewie-NS in a simplified way using Docker Compose facilitates creating private services that can cater to trusted groups of users and allow the implementation of chewie-NS in institutions operating under strict privacy rules. In a public health context, this can also be used to deploy services allowing an

easier communication between different agencies operating under distinct mandates.

The databases currently available in the public instance of chewie-NS (https://chewbbaca.online/) include schemas developed within the INNUENDO project (7) for *Salmonella enterica*, *Campylobacter jejuni*, *Escherichia coli* and *Yersinia enterocolitica*, a schema developed for *Arcobacter butzleri* (6), an adaptation of a schema generated using the Ridom SeqSphere+ software for *Acinetobacter baumannii* (22) and in-house developed schemas for *Streptococcus agalactiae* and *Streptococcus pyogenes*. However, we expect that users of chewBBACA and of other allele-calling software will increasingly contribute schemas for these as well as additional species to be deposited in chewie-NS.

DATA AVAILABILITY

Chewie-NS is freely accessible at https://chewbbaca.online/. Its source code is hosted at https://github.com/B-UMMI/Chewie-NS together with instructions on how to deploy it locally using Docker Compose and the documentation can be found at https://chewie-ns.readthedocs.io/. A tutorial version of the server, which allows users to perform mock submissions of schemas and synchronizations with a much reduced database, can be accessed at https://tutorial.chewbbaca.online/.

ACKNOWLEDGEMENTS

The authors would like to thank Catarina Inês Mendes for fruitful discussions and for her multiple design suggestions and Ana Correia for her guidance in solving issues with the chewie-NS implementation.

FUNDING

Fundos Europeus Estruturais e de Investimento (FEEI) and Fundação para a Ciência e a Tecnologia (FCT) [LISBOA-01-0145-FEDER-016417]; FCT and FEDER [01/SAICT/2016 no. 022153]; FCT [PTDC/CCI-BIO/29676/2017]. Funding for open access charge: FEEI and FCT [LISBOA-01-0145-FEDER-016417].
Conflict of interest statement. None declared.

REFERENCES

1. Black,A., MacCannell,D.R., Sibley,T.R. and Bedford,T. (2020) Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.*, **26**, 832–841.
2. Deng,X., Gu,W., Federman,S., du Plessis,L., Pybus,O.G., Faria,N.R., Wang,C., Yu,G., Bushnell,B., Pan,C.-Y. *et al.* (2020) Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science*, **369**, 582–587.
3. Maiden,M.C.J., van Rensburg,M.J.J., Bray,J.E., Earle,S.G., Ford,S.A., Jolley,K.A. and McCarthy,N.D. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.*, **11**, 728–736.
4. Uelze,L., Grützke,J., Borowiak,M., Hammerl,J.A., Juraschek,K., Deneke,C., Tausch,S.H. and Malorny,B. (2020) Typing methods based on whole genome sequencing data. *One Health Outlook*, **2**, 3.
5. Silva,M., Machado,M.P., Silva,D.N., Rossi,M., Moran-Gilad,J., Santos,S., Ramirez,M. and Carriço,J.A. (2018) chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. *Microb. Genomics*, **4**, e000166.
6. Isidro,J., Ferreira,S., Pinto,M., Domingues,F., Oleastro,M., Gomes,J.P. and Borges,V. (2020) Virulence and antibiotic resistance plasticity of *Arcobacter butzleri*: insights on the genomic diversity of an emerging human pathogen. *Infect. Genet. Evol.*, **80**, 104213.
7. Llarena,A.-K., Ribeiro-Gonçalves,B.F., Silva,D.N., Halkilahti,J., Machado,M.P., Silva,M.S.D., Jaakkonen,A., Isidro,J., Hämäläinen,C., Joenperä,J. *et al.* (2018) INNUENDO: a cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. *EFSA Support. Publ.*, **15**, 1498E.
8. Jolley,K.A., Bray,J.E. and Maiden,M.C.J. (2018) Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.*, **3**, 124.
9. Zhou,Z., Alikhan,N.-F., Mohamed,K., Fan,Y., Group,A.S. and Achtman,M. (2019) The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny and *Escherichia* core genomic diversity. *Genome Res.*, **30**, 138–152.
10. Carriço,J., Sabat,A., Friedrich,A. and Ramirez,M. (2013) Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. *Euro Surveill.*, **18**, 20382.
11. Vaz,C., Francisco,A.P., Silva,M., Jolley,K.A., Bray,J.E., Pouseele,H., Rothganger,J., Ramirez,M. and Carriço,J.A. (2014) TypOn: the microbial typing ontology. *J. Biomed. Semant.*, **5**, 43.
12. van der Linden,M., Mamede,R., Levina,N., Helwig,P., Vila-Cerqueira,P., Carriço,J.A., Melo-Cristino,J., Ramirez,M. and Martins,E.R. (2020) Heterogeneity of penicillin-non-susceptible group B streptococci isolated from a single patient in Germany. *J. Antimicrob. Chemother.*, **75**, 296–299.
13. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
14. Page,A.J., Alikhan,N.-F., Carleton,H.A., Seemann,T., Keane,J.A. and Katz,L.S. (2017) Comparison of classical multi-locus sequence typing software for next-generation sequencing data. *Microb. Genomics*, **3**, e000124.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Li,Y., Metcalf,B.J., Chochua,S., Li,Z., Walker,H., Tran,T., Hawkins,P.A., Gierke,R., Pilishvili,T., McGee,L. *et al.* (2019) Genome-wide association analyses of invasive pneumococcal isolates identify a missense bacterial mutation associated with meningitis. *Nat. Commun.*, **10**, 178.
17. Yamaguchi,M., Goto,K., Hirose,Y., Yamaguchi,Y., Sumitomo,T., Nakata,M., Nakano,K. and Kawabata,S. (2019) Identification of evolutionarily conserved virulence factor by selective pressure analysis of *Streptococcus pneumoniae*. *Commun. Biol.*, **2**, 96.
18. Moxon,R., Reche,P.A. and Rappuoli,R. (2019) Editorial: reverse vaccinology. *Front. Immunol.*, **10**, 2776.
19. Maiden,M.C.J. (2019) The impact of nucleotide sequence analysis on meningococcal vaccine development and assessment. *Front. Immunol.*, **9**, 3151.
20. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
21. Firtina,C., Kim,J.S., Alser,M., Senol Cali,D., Cicek,A.E., Alkan,C. and Mutlu,O. (2020) Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. *Bioinformatics*, **36**, 3669–3679.
22. Higgins,P.G., Prior,K., Harmsen,D. and Seifert,H. (2017) Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of *Acinetobacter baumannii*. *PLoS One*, **12**, e0179228.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

EPIDEMIOLOGY

Annotated Whole-Genome Multilocus Sequence Typing Schema for Scalable High-Resolution Typing of *Streptococcus pyogenes*

A. Friães,[a] R. Mamede,[a] M. Ferreira,[a]* J. Melo-Cristino,[a] M. Ramirez[a]

[a]Instituto de Microbiologia, Instituto de Microbiologia Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

A. Friães and R. Mamede contributed equally to this article. Author order was determined in alphabetical order.

ABSTRACT *Streptococcus pyogenes* is a major human pathogen with high genetic diversity, largely created by recombination and horizontal gene transfer, making it difficult to use single nucleotide polymorphism (SNP)-based genome-wide analyses for surveillance. Using a gene-by-gene approach on 208 complete genomes of *S. pyogenes*, a novel whole-genome multilocus sequence typing (wgMLST) schema was developed, comprising 3,044 target loci. The schema was used for core-genome MLST (cgMLST) analyses of previously published data sets and 265 newly sequenced draft genomes with other molecular and phenotypic typing data. Clustering based on cgMLST data supported the genetic heterogeneity of many *emm* types and correlated poorly with pulsed-field gel electrophoresis macrorestriction profiling, superantigen gene profiling, and MLST sequence type, highlighting the limitations of older typing methods. While 763 loci were present in all isolates of a data set representative of *S. pyogenes* genetic diversity, the proposed schema allows scalable cgMLST analysis, which can include more loci for an increased resolution when typing closely related isolates. The cgMLST and PopPUNK clusters were broadly consistent in this diverse population. The cgMLST analyses presented results comparable to those of SNP-based methods in the identification of two recently emerged sublineages of *emm*1 and *emm*89 and the clarification of the genetic relatedness among isolates recovered in outbreak contexts. The schema was thoroughly annotated and made publicly available on the chewie-NS online platform (https://chewbbaca.online/species/1/schemas/1), providing a framework for high-resolution typing and analyzing the genetic variability of loci of particular biological interest.

KEYWORDS outbreak, *Streptococcus pyogenes*, bioinformatics, genomics, group A *Streptococcus*, molecular epidemiology, molecular subtyping, population genetics, surveillance studies, typing

Streptococcus pyogenes (Lancefield group A *Streptococcus* [GAS]) remains a significant cause of global morbidity and ranks among the top 10 infectious causes of death (1). In 2018, the World Health Organization highlighted the importance of developing a GAS vaccine and set out priority activities to reach this goal, including a better characterization of the epidemiology of GAS infections and the identification of appropriate candidate antigens (2).

In recent decades, sequence-based typing of the hypervariable region of the *emm* gene, encoding the M protein, was the most frequently used method to identify GAS lineages (3). However, complementary methods have long been used, which, together with *emm* typing, allow finer discrimination of the circulating strains, including serotyping of the major backbone pilus protein (T antigen), pulsed-field gel electrophoresis (PFGE) macrorestriction profiling, multilocus sequence typing (MLST), and profiling of superantigen (SAg)-coding genes (4–6).

Whole-genome sequencing analysis allowed the identification of emerging intra-*emm* clones with increased fitness or virulence that were otherwise indistinguishable by other

Editor Daniel J. Diekema, University of Iowa College of Medicine

Address correspondence to M. Ramirez, ramirez@fm.ul.pt.

*Present address: M. Ferreira, CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental and ICBAS, Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal.

The authors declare a conflict of interest. J.M.-C. received research grants administered through his university and received honoraria for serving on the speakers bureaus of Pfizer and Merck Sharp and Dohme. M.R. received honoraria for serving on the speakers bureau of Pfizer and Merck Sharp and Dohme and for serving in expert panels of GlaxoSmithKline and Merck Sharp and Dohme. All other authors declare no conflict of interest.

Received 25 February 2022
Returned for modification 29 March 2022
Accepted 15 April 2022
Published 9 May 2022

typing methods. Such is the case of *emm*89 clade 3, which emerged during the 2000s and quickly outcompeted other *emm*89 lineages (7–9). Isolates from this lineage lack the genes encoding hyaluronic acid capsule biosynthesis and carry a high-expression promoter in the operon encoding streptolysin O and NAD-glycohydrolase (P*nga*-3) (8, 10). More recently, an *emm*1 lineage ($M1_{UK}$), differing from the contemporary globally disseminated *emm*1 lineage ($M1_{global}$) by 27 single nucleotide polymorphisms (SNPs), was identified in the United Kingdom (11) and subsequently reported in The Netherlands, the United States, and Canada (12–14).

Whole-genome sequencing has been decisive in clarifying the molecular and evolutionary mechanisms underlying the success of long-term-circulating lineages (15, 16) and has proven useful in the identification of outbreak-related cases (17, 18). Genomic data have the additional potential benefit of providing information on the variability of candidate vaccine antigens and genes involved in antimicrobial resistance (19, 20), further supporting the use of high-throughput sequencing (HTS) in GAS surveillance.

Most genome-wide analyses performed on GAS have been based on the comparison of SNPs between isolates. This usually involves mapping short-read sequence data or aligning *de novo*-assembled sequences to a selected reference genome (8–11, 15–19). However, the choice of an appropriate reference is challenging when simultaneously comparing diverse lineages (21, 22), such as in population-based studies of *S. pyogenes* infection isolates. SNP-based phylogenetic analysis also requires the removal of regions of recombination, which are an important source of diversity in GAS (19, 21–23). These limitations can be largely overcome by the use of gene-by-gene approaches like whole-genome MLST (wgMLST) or core-genome MLST (cgMLST) (24), which do not require comparison to a reference genome and which intrinsically dampen the effect of recombination (21, 22, 25). Minimum-spanning-tree (MST)-like downstream analyses further facilitate the use of wg/cgMLST. Additionally, wg/cgMLST schemas can be curated and maintained in centralized databases, providing a standardized nomenclature and ensuring reproducibility and comparison of results across laboratories (21, 24, 26, 27). Indeed, similarly to SNP-based approaches, cgMLST schemas have been successfully used for both outbreak identification and population-based surveillance of multiple pathogens (22, 25–30). However, it is important to remember that wg/cgMLST is not designed to interrogate noncoding regions of the genome and therefore would have been unable to detect the polymorphisms in the *nga-ifs-slo* promoter present now in the $M1_{global}$ lineage (11).

The aims of this study were to define a publicly available annotated wgMLST schema for *S. pyogenes* and evaluate its suitability for high-resolution typing and documenting the variability of loci encoding proteins of biological relevance.

MATERIALS AND METHODS

Bacterial strains and data sets. A collection of 265 nonduplicate GAS strains isolated from pharyngitis, skin and soft tissue infections, and normally sterile sites in Portugal between 2001 and 2009 was selected for HTS and comparison of cgMLST with other typing methods (see supplemental Data Set 1 in reference 31). These isolates were previously characterized regarding *emm* type, T type, PFGE profile, SAg gene profile, and antimicrobial resistance (5, 32–35) and represent four *emm* types: *emm*1, *emm*3, *emm*4 (including erythromycin-resistant and -susceptible isolates), and *emm*89 (including isolates carrying P*nga*-1, P*nga*-2, and P*nga*-3) (7).

In order to evaluate the performance of the proposed wgMLST schema in more diverse collections, outbreak recognition, and the identification of recently emerged intra-*emm* lineages of interest, publicly available data sets from three previous publications were also included (11, 18, 19). Data Set 2 comprises 2,006 assemblies from a collection of isolates previously selected to represent the genetic, geographic, temporal, and clinical diversity of GAS (19). Data Set 3 consists of 119 isolates associated with 21 outbreaks recorded in England from 2010 to 2015 and 170 contemporaneous sporadic isolates with the same *emm* types (18). Data Set 4 comprises 135 assemblies from noninvasive *emm*1 isolates recovered in the United Kingdom from 2009 to 2016 (11) and the MGAS5005 complete genome that was used as a reference. The United Kingdom assemblies include 123 isolates carrying 27 SNPs characteristic of the recently emerged $M1_{UK}$ lineage and 5 intermediate isolates carrying 13 or 23 of those SNPs (11). Data Set 5 includes all the *emm*89 assemblies included in Data Sets 1 to 4 ($n = 194$) and the 7 complete genomes of *emm*89 that were used to create the schema.

The majority of the strains included in Data Sets 2 to 4 (31) were retrieved from collections of publicly available genome assemblies (36, 37). For strains for which it was not possible to retrieve a public genome assembly, the raw sequencing data were downloaded from the European Nucleotide Archive (ENA) and subsequently assembled. All assemblies were filtered according to assembly quality, *emm* type, and multilocus sequence type (ST) criteria, as detailed below.

High-throughput sequencing. Genomic DNA was extracted from cultures of GAS grown overnight in Todd-Hewitt broth (Oxoid, Basingstoke, UK) using the PureLink genomic DNA minikit (Invitrogen, Carlsbad, CA, USA). The initial bacterial lysis step was carried out in the presence of 45 U of mutanolysin (Sigma-Aldrich, St. Louis, MO, USA) and 86 μg of hyaluronidase (Sigma-Aldrich, St. Louis, MO, USA). Whole-genome sequencing libraries were generated using the Nextera DNA library preparation kit (Illumina, San Diego, CA, USA). The libraries were sequenced in an Illumina MiSeq or NextSeq instrument.

Sequencing data analysis. Raw sequence reads were assembled with INNUca v4.2.2 (38), with the following parameters: -s *Streptococcus pyogenes*, -g 2, –estimatedMinimumCoverage 10, –trueCoverageProceed, and –fastQCproceed. Samples that failed any of the quality control steps related to sequence quality or assembly coverage were excluded from the data sets. Assemblies are available as supplemental material (31).

In silico ST prediction was performed using MLST v2.19.0 (39) with default parameters and the PubMLST database updated on 11 March 2021. Genome assemblies with partial matches to any of the MLST genes or for which it was not possible to identify at least one of the MLST genes were excluded, except for ST293, ST403, ST404, and ST688, which lack the *yqiL* gene, and ST1087, which lacks the *xpt* gene. Strains with a predicted ST that was inconsistent with the classification reported in the original study were also excluded.

The *emm* type was determined using emmTyper v0.2.0 (40) with verbose mode and the CDC M-type-specific sequence databases updated on 11 March 2021. Genome assemblies without an identified *emm* type, with matches only for alleles flagged in the CDC database as possible *emm*-like genes, or with a predicted *emm* type that was inconsistent with the classification reported in the original study were excluded from the data sets. Assemblies classified with multiple *emm* types were also excluded (multiple subtypes of the same *emm* type were accepted), except for *emm*34/*emm*230 (*emm*34 corresponds to the *enn* gene, and the *emm* type is 230), *emm*13L/*emm*13 (these two types correspond to the same sequence), and other cases that were inspected in Geneious v8.1.9 to validate matches to the *emm* gene.

Variant calling to determine the set of SNPs in each assembled genome from Data Set 4 (31) was performed with Snippy v4.6.0 (41) with default parameters and the complete genome of strain MGAS5005 (RefSeq accession no. GCF_000011765.3) as the reference strain.

The P*nga* variant was determined with SeqTyper v2.3 (42) with default parameters. A Fasta file with the sequences for all variants was given as the input to the blast module, followed by variant calling with the assembly module.

Schema creation, annotation, and curation. The complete genomes available in the NCBI RefSeq database as of 20 July 2020 were downloaded to select a set of 208 genome assemblies (see Table S1 in the supplemental material) (31) for schema creation with chewBBACA v2.7.0 (43). Assemblies with a status of suppressed in the NCBI database were excluded, except for accession no. GCF_001535505.1, GCF_001547815.1, GCF_000013525.1, and GCF_900636425.1, whose status was changed to suppressed after the schema creation and allele-calling processes. Loci originating from these four genomes were inspected to ensure their validity. This initial schema seed, composed of 3,318 distinct loci, was populated through the inclusion of allelic variants from all assemblies included in the data sets and sourced from public databases (36, 37). For schema annotation, the chewBBACA UniprotFinder process and custom scripts (44) were used to create a file with locus coordinates and annotation terms selected from the UniProt database, prioritizing the selection of terms from Swiss-Prot over terms from TrEMBL, and from matches against the translated coding sequences in the GenBank files of the genomes used for schema creation. Some product and gene names were further complemented with relevant literature references. The annotated schema was thoroughly curated to identify and remove spurious loci such as gene fusions, truncated genes, and paralogous loci. These loci were identified based on the retrieved annotations, the inspection of the genomic context, and the list of paralogous loci reported by the chewBBACA AlleleCall process and a custom script evaluating interlocus similarity (44). Due to the minimum sequence length parameter enforced during schema creation, the *sagA* gene, present in the streptolysin S-encoding operon, was not in the initial schema. Given the importance of this gene for GAS pathogenesis and the potential interest in its variability, a locus was added representing the *sagA* gene. The full list of changes applied to the schema is available in Table S2 (31).

The schema was uploaded to chewie-NS (45), where a more detailed description of schema creation, annotation, and curation can be found (https://chewbbaca.online/species/1/schemas/1).

cgMLST analysis. Allelic profiles of the core loci (shared by 100% of the isolates under analysis [cgMLST-100]) were used to create MSTs with the goeBURST algorithm in the desktop or online version of PHYLOViZ (46, 47). Groups of isolates linked by up to *n* different loci in the MST were determined using the desktop version. The genes present in 95% (cgMLST-95), 99% (cgMLST-99), and 100% (cgMLST-100) of the isolates were identified with the chewBBACA ExtractCgMLST process for Data Set 2. The lists of genes for each gene presence threshold are available as supplemental material (31).

Intracluster and intercluster pairwise distances were determined using custom scripts (44).

PFGE cluster definition. Previously generated SmaI/Cfr9I macrorestriction PFGE patterns (5, 32, 34, 35) were used to create a UPGMA (unweighted pair group method with arithmetic means) dendrogram with BioNumerics software (Applied Maths, Sint-Martens-Latem, Belgium). The Dice similarity coefficient was used, with optimization and position tolerance settings of 1.0 and 1.5, respectively. PFGE clusters were defined based on ≥80% relatedness on the dendrogram (4).

Statistical analysis. The results of cgMLST-100 and other typing methods were compared using Simpson's index of diversity (SID), the adjusted Wallace (AW) coefficient, and the adjusted Rand (AR) coefficient (4, 48), calculated with an online tool (http://www.comparingpartitions.info/). For comparison with other typing methods, groups were defined by cutting MSTs at a suitable allelic difference to have an SID similar to that of the method to which it was being compared.

Data availability. The annotated wgMLST schema and a detailed description of its development are publicly available in chewie-NS (45) at https://chewbbaca.online/species/1/schemas/1. The genome assemblies

TABLE 1 Simpson's index of diversity and 95% confidence intervals for the typing methods used to characterize 265 *S. pyogenes* isolates recovered in Portugal

Typing method	No. of partitions	SID ($CI_{95\%}$)[c]
emm type	4	0.742 (0.727–0.756)
ST	15	0.826 (0.800–0.852)
T type[a]	6	0.744 (0.720–0.768)
SAg profile	19	0.835 (0.813–0.857)
PFGE	16	0.792 (0.766–0.817)
MST_{1000}[b]	5	0.743 (0.728–0.758)
MST_{45}[b]	15	0.807 (0.779–0.835)
cgMLST-100	245	0.999 (0.998–1.000)

[a]The SID for T type was calculated for the subset of 248 isolates with a defined T type (17 isolates were nontypeable).
[b]Groups of isolates linked by up to *n* different loci in the MST (MST_n).
[c]SID, Simpson's index of diversity; $CI_{95\%}$, 95% confidence interval.

and allele-calling results for each data set, a static version of the wgMLST schema, the list of loci in each subschema, the pairwise distances computed for Data Sets 2 and 3, and Tables S1, S2, S5, and S9 can be found in the supplemental material (31). Raw sequencing data and sample metadata for the 265 isolates included in Data Set 1 have been deposited in the European Nucleotide Archive (ENA) under project accession number PRJEB49967. The custom scripts used for schema annotation, curation, and result analyses are part of the Schema Refinery repository (44).

RESULTS

Development of the wgMLST schema for *S. pyogenes*. The final annotated wgMLST schema comprises 3,044 loci with 371,549 alleles. Out of these, 1,096 (36%) loci presented low variability, presenting 1 to 19 DNA alleles (see Fig. S1 in the supplemental material). These correspond essentially to genes that were identified in a minority of assemblies, mostly associated with prophages and other mobile genetic elements. The exception is *sagA*, encoding the streptolysin S precursor peptide, which presented only 13 alleles despite being ubiquitous among *S. pyogenes* isolates. The short length of this locus (162 bp) may be partly responsible for the limited number of alleles.

On the opposite extreme, among the 10 most variable loci (>750 alleles) are genes encoding well-known surface-exposed virulence factors but also transcriptional regulators known to play a major role in GAS pathogenesis and virulence, namely, CovS, RopB, and Mga. For these loci, the diversity of DNA alleles also results in a very large number of protein variants (range, 535 to 1,695) (Fig. S2).

Specific analyses can be performed through the identification and creation of subschemas for smaller sets of biologically relevant loci, such as genes encoding virulence factors and transcriptional regulators, for which subschemas are provided as supplemental material (31).

Comparison of cgMLST with other typing methods. To compare cgMLST analysis with conventional typing methods, a collection of 265 infection isolates with previous information on *emm* type, ST, T type, PFGE profile, PCR profile of 11 SAg genes, and antimicrobial resistance was used (see Data Set 1 in reference 31). This data set includes isolates of *emm* types 1, 3, 4, and 89; 15 distinct STs; 6 T types (17 isolates were nontypeable); 19 SAg profiles; and 16 PFGE clusters (Table 1).

Allele calling using the wgMLST schema followed by cgMLST-100 analysis generated 245 different profiles representing 1,230 loci. The resulting MST separated the isolates according to *emm* type, except for one *emm*4 isolate that did not cluster with the others (Fig. 1). The minimum distance between clusters of different *emm* types varied between 1,084 and 1,105 allelic differences, while those among isolates of the same *emm* type were ≤28 for *emm*1, ≤100 for *emm*3, ≤157 for *emm*4 (excluding the distantly related isolate), and ≤225 for *emm*89. Clustering of isolates at a cutoff of 1,000 differences created four groups separating the four *emm* types and one singleton, corresponding to the *emm*4 isolate, resulting in high concordance between the MST groups linked by up to 1,000 different loci and *emm* types (Table 1; see also Tables S3 and S4 in the supplemental material).

A lower congruence was obtained between the distributions of isolates in the MST and the remaining typing methods (ST, PFGE, SAg profiling, and T type) (Fig. S3 to S6).

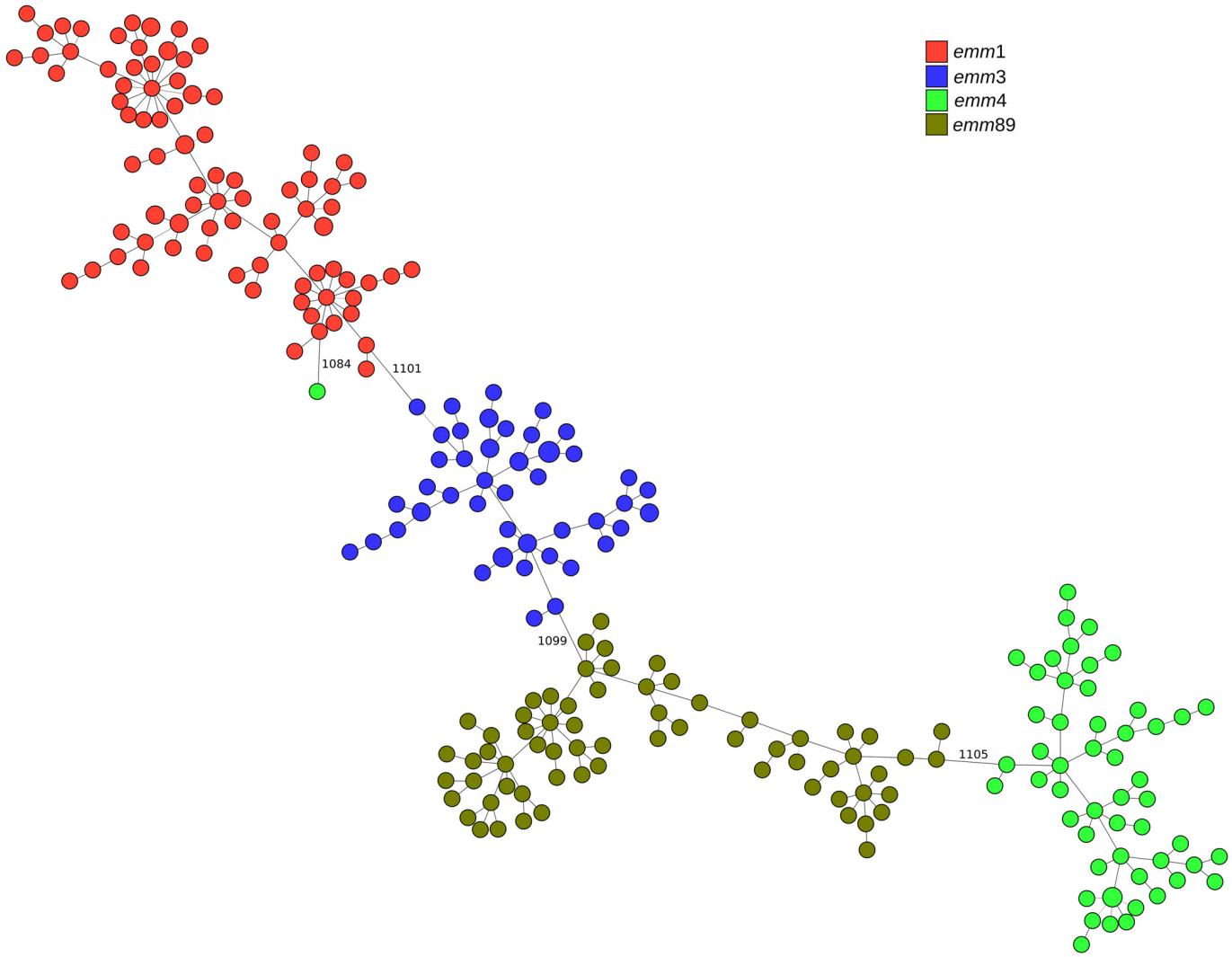


FIG 1 Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST-100 profiles of 265 *S. pyogenes* isolates recovered in Portugal (see Data Set 1 in reference 31). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to *emm* type. Link distances of ≥1,000 allelic differences are labeled (from a total of 1,230 compared loci).

The AW and AR values between T types and MST groups linked by up to 1,000 different loci were only slightly lower than those of *emm* types (Tables S3 and S4), but T type had a lower typeability since 17 isolates were nontypeable. Although MST groups linked by up to 45 different loci resulted in a number of partitions and SIDs comparable to those of ST, SAg profiling, and PFGE (Table 1), the AW coefficient between MST groups linked by up to 45 different loci and these typing methods was lower than that between MST groups linked by up to 1,000 different loci and *emm* type (Table S4). This means that MST groups linked by up to 45 different loci could not confidently predict the ST, PFGE cluster, or SAg profile, or the converse, which was also reflected in lower AR values (<0.900) (Table S3).

The use of a wgMLST schema instead of a universally defined cgMLST-100 set of loci allows scalable analysis in which higher resolution can be obtained by including larger numbers of common loci when analyzing closely related isolates. As an example, the cgMLST-100 obtained exclusively for the *emm*4 isolates grouped into the same MST group linked by up to 1,000 different loci (n = 54) comprises 52 profiles of 1,382 cgMLST-100 loci. The *emm*4 isolates presenting the M phenotype of macrolide resistance (erythromycin resistant and clindamycin susceptible) shared ST39 and an SAg profile with most susceptible isolates (see Data Set 1 in reference 31), rendering these two methods unable to differentiate macrolide-resistant isolates. One PFGE cluster was associated with macrolide resistance (49), although it also included two susceptible isolates. Similarly, one of the MST groups

linked by up to 33 different loci comprised exclusively all but two of the macrolide-resistant isolates (Fig. S7). Not surprisingly, the set of 46 loci that were present universally and exclusively in the subset of erythromycin-resistant isolates (list available in the supplemental material in reference 31) represents mostly phage-related genes, including *mef*(A) and *msr*(D), the genes most commonly associated with the M phenotype in GAS (50, 51).

Performance of the wgMLST schema on a large and genetically diverse data set. The genetic structure of the GAS population is known to vary temporally and geographically, with an associated impact on the disease spectrum and incidence (52, 53). To evaluate the performance of the proposed wgMLST schema on the analysis of genetically diverse data sets, we used a large collection of isolates previously selected to represent the genetic, geographic, temporal, and clinical diversity of GAS (19). A total of 2,006 assemblies were included in the data set, comprising 140 *emm* types and 443 STs and organized into 292 phylogroups defined by PopPUNK (19, 54) (see Data Set 2 in reference 31).

We defined 1,321-locus cgMLST-95, 1,204-locus cgMLST-99, and 763-locus cgMLST-100 schemas (available in the supplemental material in reference 31).

Allele call results identified 1,700 cgMLST-100 profiles. The resulting MST indicates that many *emm* types include diverse genetic lineages, with 12 of the 19 most prevalent *emm* types ($>$30 isolates) comprising isolates distributed in multiple tree regions (Fig. 2). Accordingly, 50 of the 67 *emm* types comprising $\geq$10 isolates included assemblies that differed in $>$50% of the 763 cgMLST-100 loci (up to 708 differences [93%] in *emm*4) (Fig. 3A). In 31 of these *emm* types, the mean intra-*emm* allelic difference was larger than the smaller difference from another *emm* type (Table S5). This is possibly due to the diversity of geographic and temporal origins of the isolates in this data set and is in line with a previous report of genetic heterogeneity within *emm* types (19). It is also reflected in a low congruence between *emm* types and MST groups linked by up to 450 different loci despite similar SID values (Tables S6 to S8).

The overall congruence between STs and MST groups linked by up to 50 different loci was poor although slightly higher than that observed for the less diverse Data Set 1 (AR coefficients of 0.810 and 0.709, respectively) (Table S7). In contrast, there was good congruence, with high AW and AR values, between PopPUNK phylogroups and MST groups linked by up to 200 different loci (Tables S7 and S8). Still, PopPUNK phylogroups can be rather diverse, including multiple STs and isolates differing in up to 61% of the core 763 loci (phylogroup 27) (Fig. 3B; Table S9), highlighting the advantage of using multiple methods for analyzing the evolution of GAS lineages.

Performance of the wgMLST schema in an outbreak context. To evaluate the potential contribution of the proposed wgMLST schema for outbreak recognition, we used a previously published data set comprising isolates from 21 outbreaks in England and contemporaneous nonrelated isolates with the same *emm* types (18). A total of 119 outbreak isolates and 170 sporadic isolates were included (see Data Set 3 in reference 31). Allele calling for the 119 outbreak isolates identified 58 profiles of 1,263 cgMLST-100 loci. In agreement with the SNP-based clustering presented previously (18), the MST clustered the isolates according to *emm* type, with a minimum distance of 1,079 allelic differences between isolates of different *emm* types, while isolates of different subtypes or outbreaks of the same *emm* type were more closely related (Fig. 4).

Individual MSTs were created for *emm* types 1, 5, 11, 28, 75, 89, and 94, including outbreak and sporadic isolates (Fig. S8 to S14). Since these MSTs included only isolates sharing the same *emm* type, they comprised larger sets of cgMLST-100 loci (1,384 to 1,547 loci), potentially allowing higher resolution in the discrimination of outbreak isolates. Ten isolates with epidemiological links could be excluded from the respective outbreaks because they did not cluster with isolates of the same outbreak or differed by too many loci (Table S10 and Fig. S8 and S11 to S13). These isolates also matched the outbreak exclusion criteria based on SNP analysis (18).

Except for these 10 excluded isolates, outbreak isolates linked in the MSTs shared $>$99.5% of their core genome (maximum link distance of 6 allelic differences), and the mean distance within a given outbreak was much lower than the mean distance among sporadic

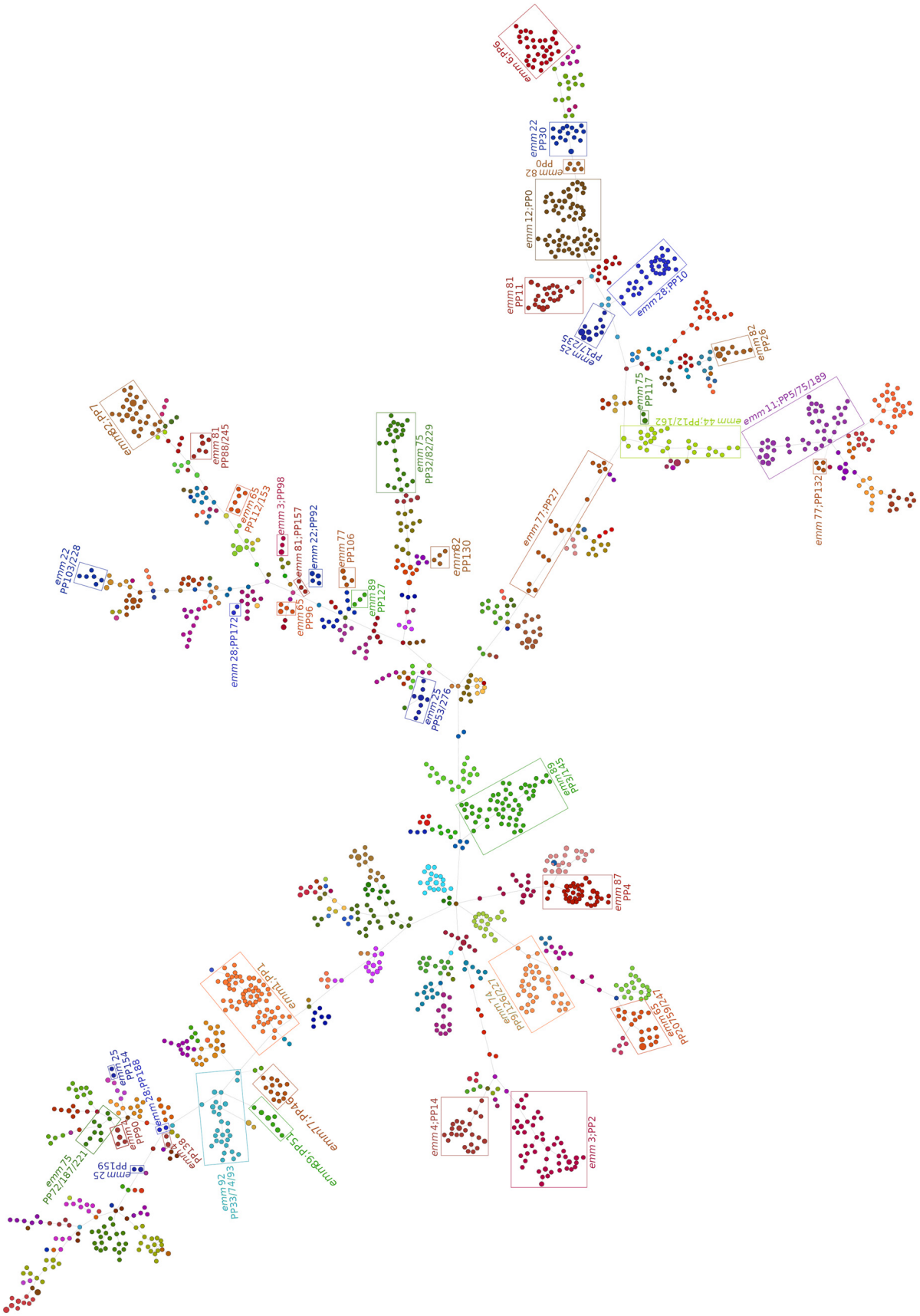


FIG 2 Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST profiles of 2,006 genetically diverse *S. pyogenes* isolates recovered worldwide (19) (see Data Set 2 in reference 31). The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. Nodes are colored according to *emm* type. Groups of clustered *emm* types represented by >30 isolates are highlighted inside rectangles and labeled with the respective *emm* types and PopPUNK (PP) phylogroup numbers (for simplicity, isolated nodes of *emm* types 4, 22, 44, 65, 75, 77, 81, and 92 are not highlighted). A total of 763 core loci were compared.

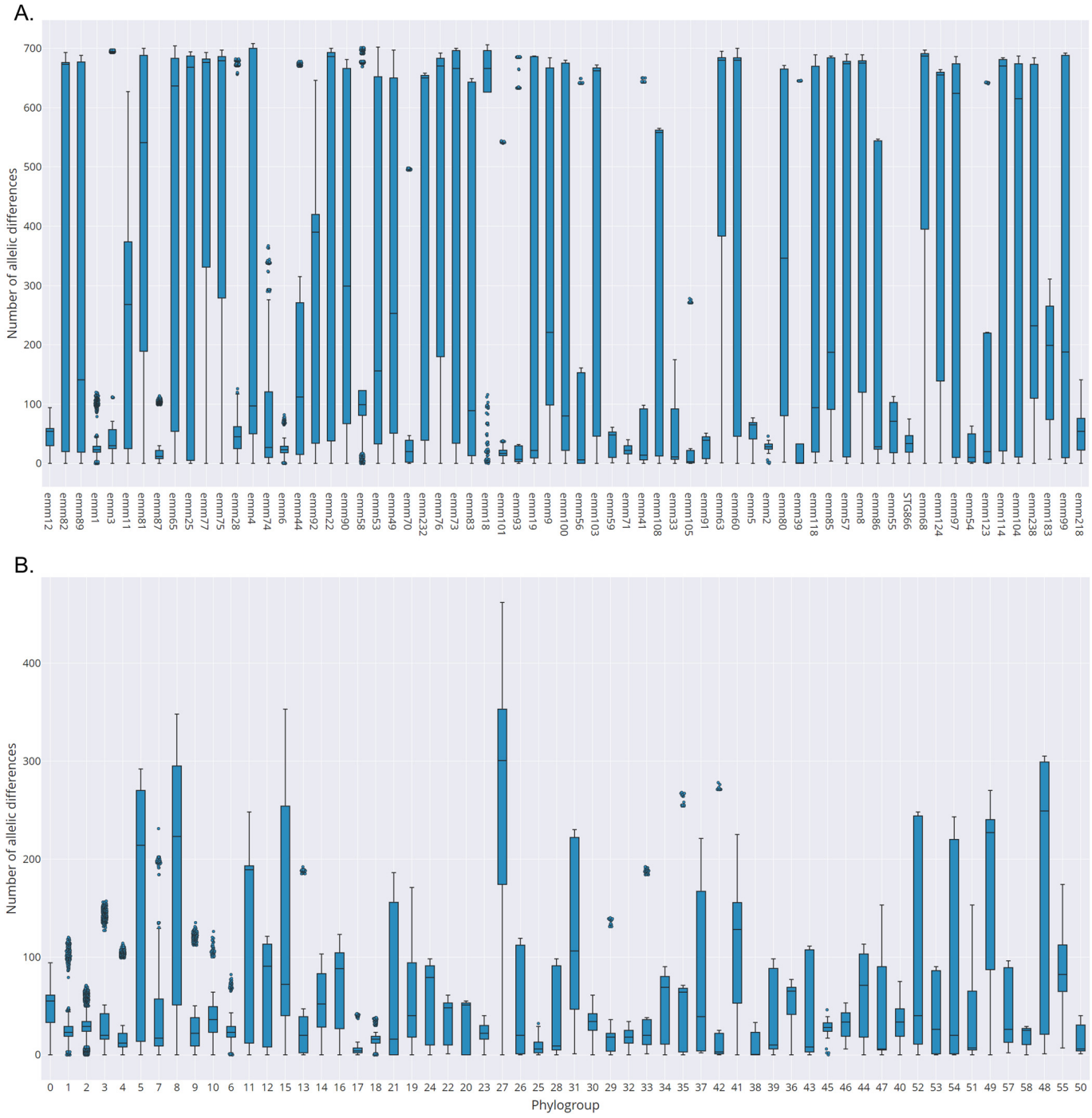


FIG 3 Box-and-whisker plots for the pairwise distances of the assemblies from Data Set 2 (19, 31) included in each *emm* type with ≥10 isolates (A) or in each PopPUNK phylogroup with ≥10 isolates (B). The distances were calculated based on the allele call results for the 763 cgMLST-100 loci of the 2,006 assemblies (interactive versions of these plots are available as supplemental material in reference 31).

isolates of the same *emm* type (Table 2). However, in *emm* types 1 and 5, there were sporadic isolates with cgMLST profiles very similar to those of outbreak 1 (OB1) and OB19, respectively (0 to 2 allelic differences), indicating that these outbreak strains were also present in the community (Table 2; Fig. S8 and S9).

Performance of the wgMLST schema in the identification of recently emerged lineages. We tested if the proposed wgMLST schema has enough discriminatory power to identify two recently emerged intra-*emm* lineages that were originally identified by whole-genome SNP analysis, namely, $M1_{UK}$ and *emm*89 clade 3 (8, 9, 11). Allele calling was performed

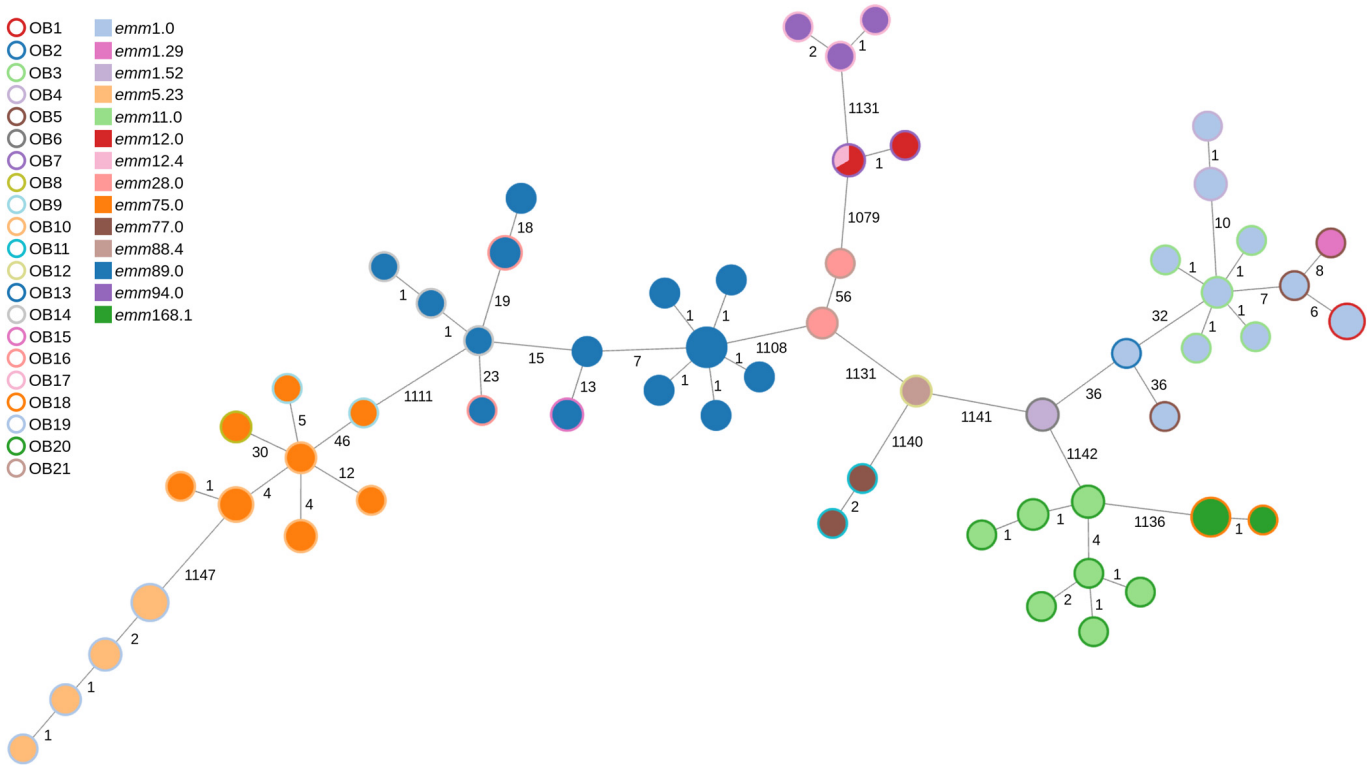


FIG 4 Minimum-spanning tree generated with the goeBURST algorithm for the cgMLST-100 profiles of 119 outbreak *S. pyogenes* isolates recovered in the United Kingdom (18) (see Data Set 3 in reference 31). The size of each node is proportional to the number of isolates with that particular cgMLST profile on a logarithmic scale. The nodes are colored according to the *emm* type, and the outer ring is colored according to the outbreak number. Link distances are labeled as the number of allelic differences between nodes (from a total of 1,263 compared loci).

for the 135 assemblies from noninvasive *emm*1 isolates (11) together with the complete genome of strain MGAS5005, a reference representative of the $M1_{global}$ lineage (see Data Set 4 in reference 31). The graph in Fig. 5 represents the resulting MST with all links of up to 19 differences depicted. All $M1_{UK}$ isolates were tightly clustered, together with an intermediate isolate ($M1_{inter}$) carrying 23 of the 27 SNPs characteristic of the $M1_{UK}$ lineage (11). The MST links within this cluster ranged between 0 and 13 differences, while the closest links to the $M1_{inter}$ cluster (13 SNPs) and an $M1_{global}$ isolate were 20 and 31 differences, respectively. $M1_{global}$ isolates presented higher genomic diversity, with MST links of up to 49 differences.

Allele calling was performed for all *emm*89 assemblies included in the four data sets described above and all the complete *emm*89 genomes used to create the schema (n = 201) (see Data Set 5 in reference 31). In addition, the P*nga* variant was determined for all isolates. The absence of the *hasA* gene of the capsule locus was confirmed in all P*nga*-3 isolates, while all other isolates carried this gene, except for two ST568 isolates that have an internal nonsense codon in *hasA*. The graph depicting all links of up to 55 differences (Fig. 6) showed limited diversity in the isolates carrying P*nga*-3, which clustered closely, with MST links with 0 to 27 differences, while the shortest link to a P*nga*-2 isolate was 57 differences. The P*nga*-2 and, especially, P*nga*-1 isolates were more diverse, presenting fewer links with up to 55 differences and comprising multiple sublineages associated with different STs (Fig. 6; Fig. S15). Both the wider geographic range and collection time span may contribute to this higher diversity. As previously reported (7), MLST was not suitable for discriminating P*nga*-3 isolates from those carrying P*nga*-2 since ST101 was prevalent among both lineages (Fig. S15). Analysis of the *emm*89 isolates from Data Set 1 showed that P*nga*-3 isolates and most P*nga*-2 isolates were also grouped into the same PFGE cluster, and some of them shared the same SAg profile, while the T serotype B3264 was ubiquitous, except for the single P*nga*-1 isolate (T11) and one P*nga*-3 isolate that was nontypeable (Fig. S16 to S18). PopPUNK clustering

TABLE 2 Distances (numbers of allelic differences) among outbreak isolates and between each outbreak and sporadic isolates of the same *emm* type determined by cgMLST-100 analysis for each *emm* type, using a collection of isolates from the United Kingdom[a]

emm type	No. of loci in cgMLST	Subset (no. of isolates)	Mean distance within subset (range)	Mean distance to sporadic isolates (range)
1	1,488	OB1 (6)	0.6 (0–1)	17.9 (2–59)
		OB3 (6)	1.7 (1–2)	18.4 (8–53)
		OB4 (4)	0.5 (0–1)	24.8 (15–60)
		OB6 (3)	0.7 (0–1)	50.1 (46–55)
		Sporadic (30)	25.7 (3–63)	NA
5	1,485	OB19 (14)	1.6 (0–4)	124.1 (0–174)
		Sporadic (27)	112.0 (0–175)	NA
11	1,384	OB20 (10)	3.8 (0–8)	89.8 (35–592)
		Sporadic (26)	118.6 (1–597)	NA
28	1,510	OB21 (2)	0	44 (12–67)
		Sporadic (11)	51.0 (0–74)	NA
75	1,547	OB8 (2)	0	20.1 (14–65)
		OB10 (11)	5.4 (0–11)	19.6 (12–70)
		Sporadic (39)	19.5 (0–76)	NA
89	1,392	OB13 (17)	0.93 (0–2)	28.4 (11–42)
		OB14 (3)	1.3 (1–2)	29.0 (16–41)
		OB15 (3)	0 (0–0)	32.5 (16–47)
		OB16 (4)	0.5 (0–1)	33.1 (14–45)
		Sporadic (31)	31.7 (0–50)	NA
94	1,506	OB17 (3)	2.7 (1–4)	29.6 (10–48)
		Sporadic (6)	31.4 (2–50)	NA

[a]See reference 18 and Data Set 3 in reference 31. Ten outbreak isolates were excluded according to the results of both cgMLST-100 and SNP analyses (18). NA, not applicable.

also could not discriminate P*nga*-3 isolates, which were clustered with isolates carrying P*nga*-1 and P*nga*-2 in phylogroup 3 (see Data Set 5 in reference 31).

DISCUSSION

The reduced costs of HTS have facilitated a wider application of whole-genome data to the epidemiological surveillance of multiple pathogens. This leads to a requirement for standardized analysis pipelines producing reproducible and portable results that can be easily compared across laboratories and with those of previously used typing methods (55). Here, we propose a wgMLST schema for *S. pyogenes*, consisting of 3,044 loci. Hard-defined cgMLST schemas comprising the subsets of loci present in 95% (1,321 loci), 99% (1,204 loci), and 100% (763 loci) of the assemblies of a collection representing the genetic diversity of *S. pyogenes* (19) are also presented. However, the use of a wgMLST schema from which the cgMLST loci are selected according to the specific data set under analysis has the advantage of allowing the inclusion of larger subsets of loci and, hence, increased resolution when comparing closely related isolates (21). This can be particularly important to track the emergence of intra-*emm*-type sublineages or identify outbreak-related isolates.

The application of the schema proposed here to previously published data sets and analysis of the resulting MSTs showed a performance comparable to that of SNP-based methods in distinguishing recently emerged intra-*emm*-type sublineages as well as in identifying clusters of epidemiologically and genetically related isolates associated with local, short-term outbreaks (8, 11, 18). Analyses based on wg/cgMLST build upon the strengths of gene-by-gene approaches, which do not require a reference genome or the removal of regions of recombination (21, 22, 25). This is particularly important when analyzing collections of genetically diverse lineages, such as in long-term surveillance studies, particularly in organisms

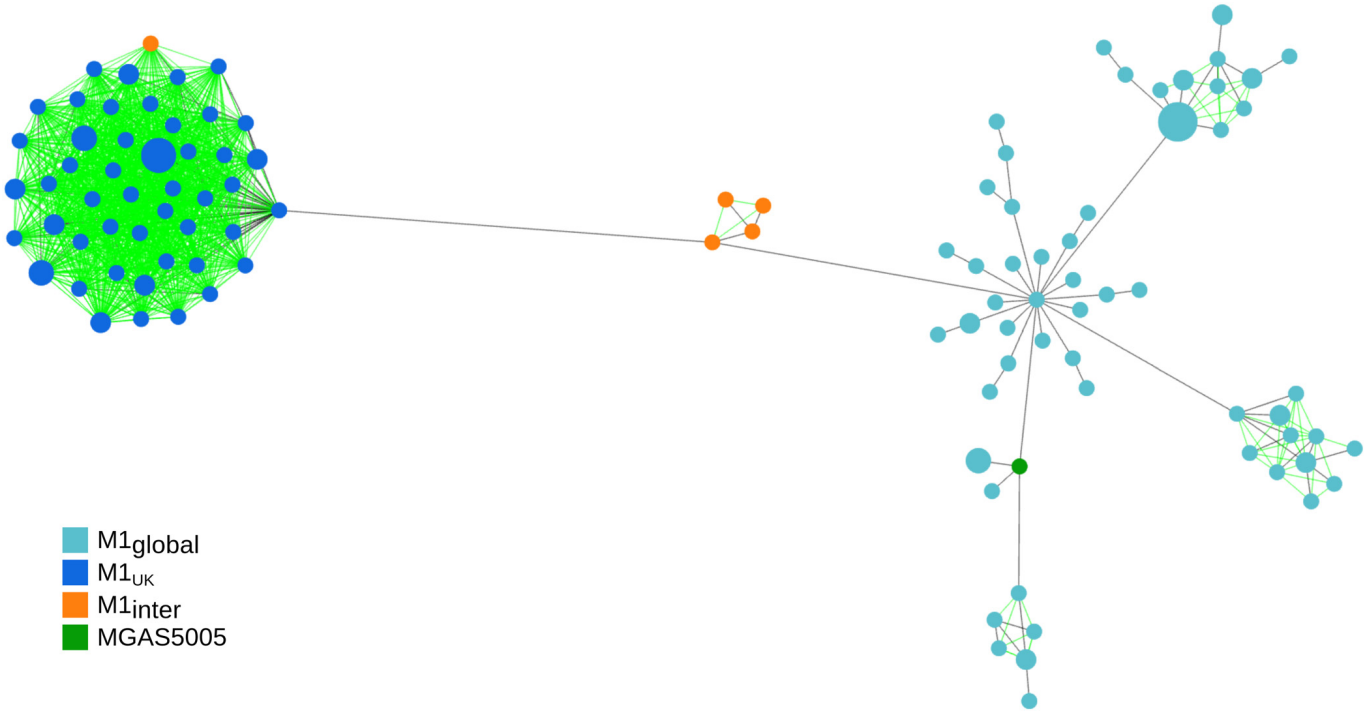


FIG 5 Graph representation of the relationships between the cgMLST-100 profiles of 135 noninvasive *emm*1 isolates recovered in the United Kingdom (11) and reference strain MGAS5005 (see Data Set 4 in reference 31), depicting all links with ≤19 allelic differences (from a total of 1,404 compared loci). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to the M1 lineage, with MGAS5005 (reference genome for the $M1_{global}$ lineage) in green. Links that would not be present in the standard MST are shown in green. Links shown in black represent the MST links and may represent distances with >19 allelic differences.

where mobile genetic elements and recombination play major roles in genomic plasticity and evolution, such as *S. pyogenes* (19, 23). Moreover, gene-by-gene approaches constitute a framework that has been widely used in surveillance, which can facilitate the transition to wg/cgMLST by reference laboratories involved in surveillance activities.

Comparison of cgMLST-based clustering with other typing methods used for *S. pyogenes* revealed poor concordance, although in temporally and geographically restricted data sets, the groups defined by *emm* typing were also supported by cgMLST. By including a much higher number of loci, cgMLST was expected to present a higher discriminatory power than the traditional seven-gene MLST schema and to further discriminate isolates sharing the same ST (22, 27). However, such a simplistic expectation was not universally borne out by the data, which highlights the limitations of the seven-gene MLST schema to correctly identify GAS lineages based on broader genomic information. It is worth noting that from the seven genes included in traditional MLST, two (*gtr* and *yqiL*) were excluded from the wgMLST schema because they shared alleles with paralogous genes, and one (*xpt*) was absent in at least one GAS lineage and therefore was not always included in the cgMLST analysis.

In contrast, a good correlation was found between cgMLST clustering and PopPUNK (19, 54), another whole-genome-based clustering method. However, the flexibility of wg/cgMLST allows increased resolution by lowering the number of allelic differences used to define clusters and a dynamic cgMLST definition, providing further discrimination within PopPUNK clusters.

The proposed wgMLST schema is publicly available on the chewie-NS platform (45), where multiple statistics regarding the whole schema and individual loci can be visualized (https://chewbbaca.online/species/1/schemas/1). The close integration with the chewBBACA suite (43) facilitates its use in surveillance and epidemiological studies and the maintenance of a common nomenclature across different studies. By virtue of the comprehensive annotation, the database can be used to obtain relevant data for basic research, such as the variability of genes of interest (virulence factors,

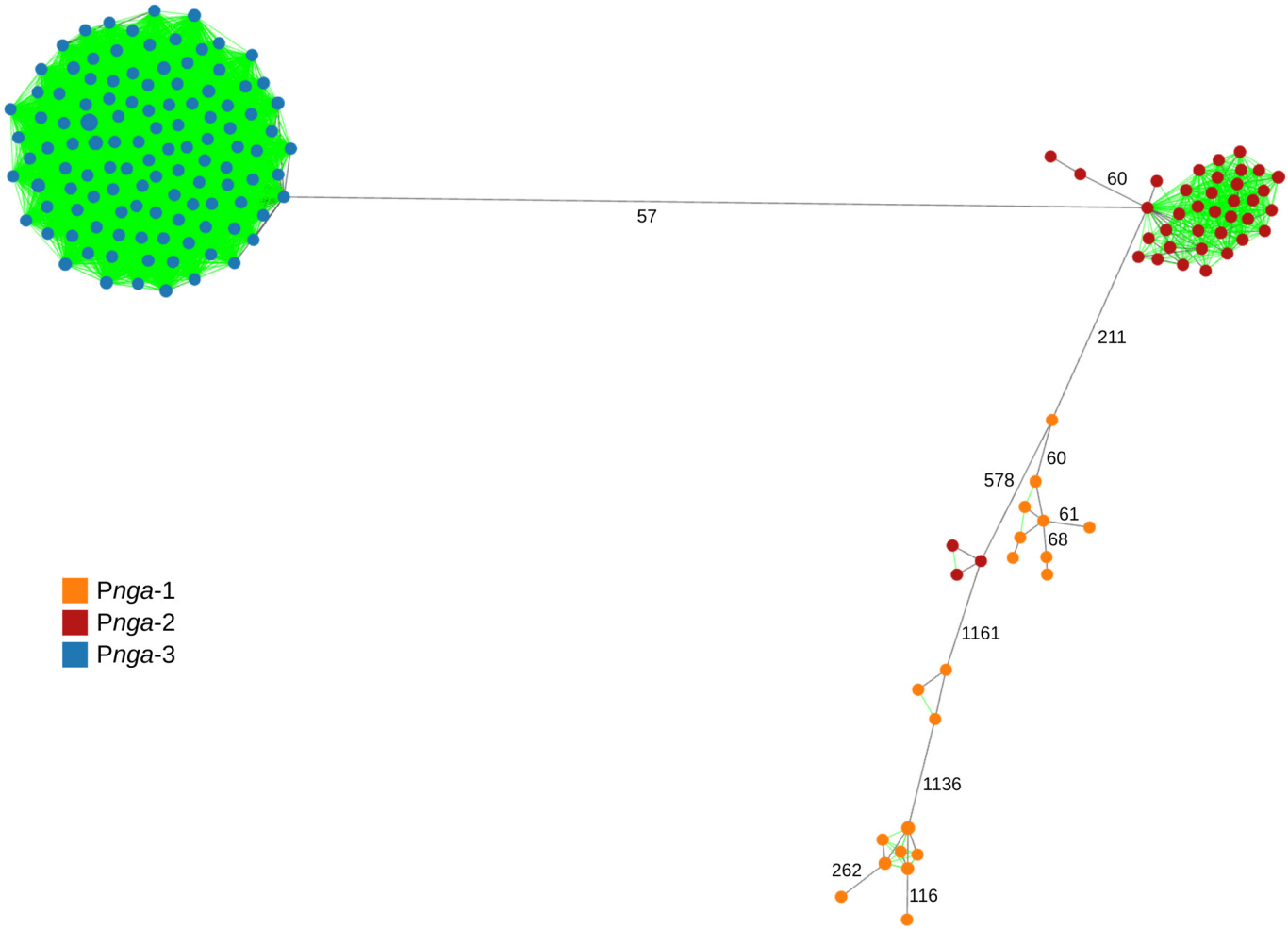


FIG 6 Graph representation of the relationships between the cgMLST-100 profiles of 201 *emm*89 isolates (see Data Set 5 in reference 31) depicting all links with ≤55 allelic differences (from a total of 1,279 compared loci). The size of each node is proportional to the number of isolates with that particular cgMLST-100 profile on a logarithmic scale. Nodes are colored according to the variant of the *nga* promoter (P*nga*). Links that would not be present in the standard MST are shown in green. Links shown in black represent the MST links and may represent distances with >55 allelic differences (labeled links).

antimicrobial resistance genes, candidate vaccine antigens, and transcriptional regulators, etc.) (19, 20) in addition to its use for typing purposes.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.2 MB.

ACKNOWLEDGMENTS

R.M. was supported by the Fundação para a Ciência e Tecnologia (FCT) (grant 2020.08493.BD). Partial support was received from the ONEIDA project (LISBOA-01-0145-FEDER-016417), cofunded by the FEEI (Fundos Europeus Estruturais e de Investimento) from the Programa Operacional Regional de Lisboa, Portugal, 2020 (POR Lisboa 2020), and by national funds from the FCT and the LISBOA-01-0145-FEDER-007391 project, cofunded by FEDER through POR Lisboa 2020 and the Fundação para a Ciência e a Tecnologia.

REFERENCES

1. Carapetis JR, Steer AC, Mulholland EK, Weber M. 2005. The global burden of group A streptococcal diseases. Lancet Infect Dis 5:685–694. https://doi.org/10.1016/S1473-3099(05)70267-X.
2. Vekemans J, Gouvea-Reis F, Kim JH, Excler J-L, Smeesters PR, O'Brien KL, Van Beneden CA, Steer AC, Carapetis JR, Kaslow DC. 2019. The path to group A *Streptococcus* vaccines: World Health Organization research and

development technology roadmap and preferred product characteristics. Clin Infect Dis 69:877–883. https://doi.org/10.1093/cid/ciy1143.
3. Beall B, Facklam R, Thompson T. 1996. Sequencing *emm*-specific PCR products for routine and accurate typing of group A streptococci. J Clin Microbiol 34:953–958. https://doi.org/10.1128/jcm.34.4.953-958.1996.
4. Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. 2006. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clin Microbiol 44:2524–2532. https://doi.org/10.1128/JCM.02536-05.
5. Friães A, Pinto FR, Silva-Costa C, Ramirez M, Melo-Cristino J. 2013. Superantigen gene complement of *Streptococcus pyogenes*—relationship with other typing methods and short-term stability. Eur J Clin Microbiol Infect Dis 32:115–125. https://doi.org/10.1007/s10096-012-1726-3.
6. Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE. 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between *emm* type and clone. Infect Immun 69:2416–2427. https://doi.org/10.1128/IAI.69.4.2416-2427.2001.
7. Friães A, Machado MP, Pato C, Carriço J, Melo-Cristino J, Ramirez M. 2015. Emergence of the same successful clade among distinct populations of *emm*89 *Streptococcus pyogenes* in multiple geographic regions. mBio 6: e01780-15. https://doi.org/10.1128/mBio.01780-15.
8. Turner CE, Abbott J, Lamagni T, Holden MTG, David S, Jones MD, Game L, Efstratiou A, Sriskandan S. 2015. Emergence of a new highly successful acapsular group A *Streptococcus* clade of genotype *emm*89 in the United Kingdom. mBio 6:e00622-15. https://doi.org/10.1128/mBio.00622-15.
9. Zhu L, Olsen RJ, Nasser W, Beres SB, Vuopio J, Kristinsson KG, Gottfredsson M, Porter AR, DeLeo FR, Musser JM. 2015. A molecular trigger for intercontinental epidemics of group A *Streptococcus*. J Clin Invest 125:3545–3559. https://doi.org/10.1172/JCI82478.
10. Zhu L, Olsen RJ, Nasser W, de la Riva Morales I, Musser JM. 2015. Trading capsule for increased cytotoxin production: contribution to virulence of a newly emerged clade of *emm*89 *Streptococcus pyogenes*. mBio 6:e01378-15. https://doi.org/10.1128/mBio.01378-15.
11. Lynskey NN, Jauneikaite E, Li HK, Zhi X, Turner CE, Mosavie M, Pearson M, Asai M, Lobkowicz L, Chow JY, Parkhill J, Lamagni T, Chalker VJ, Sriskandan S. 2019. Emergence of dominant toxigenic M1T1 *Streptococcus pyogenes* clone during increased scarlet fever activity in England: a population-based molecular epidemiological study. Lancet Infect Dis 19:1209–1218. https://doi.org/10.1016/S1473-3099(19)30446-3.
12. Rümke LW, de Gier B, Vestjens SMT, van der Ende A, van Sorge NM, Vlaminckx BJM, Witteveen S, van Santen M, Schouls LM, Kuijper EJ. 2020. Dominance of M1UK clade among Dutch M1 *Streptococcus pyogenes*. Lancet Infect Dis 20: 539–540. https://doi.org/10.1016/S1473-3099(20)30278-4.
13. Li Y, Nanduri SA, Van Beneden CA, Beall BW. 2020. M1UK lineage in invasive group A *Streptococcus* isolates from the USA. Lancet Infect Dis 20: 538–539. https://doi.org/10.1016/S1473-3099(20)30279-6.
14. Demczuk W, Martin I, Domingo FR, MacDonald D, Mulvey MR. 2019. Identification of *Streptococcus pyogenes* M1UK clone in Canada. Lancet Infect Dis 19:1284–1285. https://doi.org/10.1016/S1473-3099(19)30622-X.
15. Nasser W, Beres SB, Olsen RJ, Dean MA, Rice KA, Long SW, Kristinsson KG, Gottfredsson M, Vuopio J, Raisanen K, Caugant DA, Steinbakk M, Low DE, McGeer A, Darenberg J, Henriques-Normark B, Van Beneden CA, Hoffmann S, Musser JM. 2014. Evolutionary pathway to increased virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. Proc Natl Acad Sci U S A 111:E1768–E1776. https://doi.org/10.1073/pnas.1403138111.
16. Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ, Goldman TD, Feldgarden M, Birren BW, Fofanov Y, Boos J, Wheaton WD, Honisch C, Musser JM. 2010. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. Proc Natl Acad Sci U S A 107:4371–4376. https://doi.org/10.1073/pnas.0911295107.
17. Turner CE, Bedford L, Brown NM, Judge K, Török ME, Parkhill J, Peacock SJ. 2017. Community outbreaks of group A *Streptococcus* revealed by genome sequencing. Sci Rep 7:8554. https://doi.org/10.1038/s41598-017-08914-x.
18. Coelho JM, Kapatai G, Jironkin A, Al-Shahib A, Daniel R, Dhami C, Laranjeira AM, Chambers T, Phillips S, Tewolde R, Underwood A, Chalker VJ. 2019. Genomic sequence investigation *Streptococcus pyogenes* clusters in England (2010-2015). Clin Microbiol Infect 25:96–101. https://doi.org/10.1016/j.cmi.2018.04.011.
19. Davies MR, McIntyre L, Mutreja A, Lacey JA, Lees JA, Towers RJ, Duchêne S, Smeesters PR, Frost HR, Price DJ, Holden MTG, David S, Giffard PM, Worthing KA, Seale AC, Berkley JA, Harris SR, Rivera-Hernandez T, Berking O, Cork AJ, Torres RSLA, Lithgow T, Strugnell RA, Bergmann R, Nitsche-Schmitz P, Chhatwal GS, Bentley SD, Fraser JD, Moreland NJ, Carapetis JR, Steer AC, Parkhill J, Saul A, Williamson DA, Currie BJ, Tong SYC, Dougan G, Walker MJ. 2019. Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. Nat Genet 51:1035–1043. https://doi.org/10.1038/s41588-019-0417-8.
20. Beres SB, Zhu L, Pruitt L, Olsen RJ, Faili A, Kayal S, Musser JM. 2022. Integrative reverse genetic analysis identifies polymorphisms contributing to decreased antimicrobial agent susceptibility in *Streptococcus pyogenes*. mBio 13:e03618-21. https://doi.org/10.1128/mbio.03618-21.
21. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. https://doi.org/10.1038/nrmicro3093.
22. Neumann B, Prior K, Bender JK, Harmsen D, Klare I, Fuchs S, Bethe A, Zühlke D, Göhler A, Schwarz S, Schaffer K, Riedel K, Wieler LH, Werner G. 2019. A core genome multilocus sequence typing scheme for *Enterococcus faecalis*. J Clin Microbiol 57:e01686-18. https://doi.org/10.1128/JCM.01686-18.
23. McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, Beall B, Bessen DE. 2004. Multilocus sequence typing of *Streptococcus pyogenes* representing most known *emm* types and distinctions among subpopulation genetic structures. J Bacteriol 186:4285–4294. https://doi.org/10.1128/JB.186.13.4285-4294.2004.
24. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. 2018. A primer on microbial bioinformatics for nonbioinformaticians. Clin Microbiol Infect 24:342–349. https://doi.org/10.1016/j.cmi.2017.12.015.
25. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol 52:2365–2370. https://doi.org/10.1128/JCM.00262-14.
26. Higgins PG, Prior K, Harmsen D, Seifert H. 2017. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of *Acinetobacter baumannii*. PLoS One 12: e0179228. https://doi.org/10.1371/journal.pone.0179228.
27. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. J Clin Microbiol 56:e01987-17. https://doi.org/10.1128/JCM.01987-17.
28. Abdel-Glil MY, Chiaverini A, Garofolo G, Fasanella A, Parisi A, Harmsen D, Jolley KA, Elschner MC, Tomaso H, Linde J, Galante D. 2021. A whole-genome-based gene-by-gene typing system for standardized high-resolution strain typing of *Bacillus anthracis*. J Clin Microbiol 59:e02889-20. https://doi.org/10.1128/JCM.02889-20.
29. Pinto M, González-Díaz A, Machado MP, Duarte S, Vieira L, Carriço JA, Marti S, Bajanca-Lavado MP, Gomes JP. 2019. Insights into the population structure and pan-genome of *Haemophilus influenzae*. Infect Genet Evol 67:126–135. https://doi.org/10.1016/j.meegid.2018.10.025.
30. Bardenstein S, Gibbs RE, Yagel Y, Motro Y, Moran-Gilad J. 2021. Brucellosis outbreak traced to commercially sold camel milk through whole-genome sequencing, Israel. Emerg Infect Dis 27:1728–1731. https://doi.org/10.3201/eid2706.204902.
31. Friães A, Mamede R, Ferreira M, Melo-Cristino J, Ramirez M. 2022. Supplemental material for "An annotated whole-genome multilocus sequence typing schema for scalable high resolution typing of Streptococcus pyogenes". Zenodo. https://doi.org/10.5281/zenodo.5901775.
32. Friães A, Pinto FR, Silva-Costa C, Ramirez M, Melo-Cristino J, Portuguese Group for the Study of Streptococcal Infections. 2012. Group A streptococci clones associated with invasive infections and pharyngitis in Portugal present differences in *emm* types, superantigen gene content and antimicrobial resistance. BMC Microbiol 12:280. https://doi.org/10.1186/1471-2180-12-280.
33. Pato C, Melo-Cristino J, Ramirez M, Friães A, Portuguese Group for the Study of Streptococcal Infections. 2018. *Streptococcus pyogenes* causing skin and soft tissue infections are enriched in the recently emerged *emm*89 clade 3 and are not associated with abrogation of CovRS. Front Microbiol 9:2372. https://doi.org/10.3389/fmicb.2018.02372.
34. Friães A, Lopes JP, Melo-Cristino J, Ramirez M, Portuguese Group for the Study of Streptococcal Infections. 2013. Changes in *Streptococcus pyogenes* causing invasive disease in Portugal: evidence for superantigen gene loss and acquisition. Int J Med Microbiol 303:505–513. https://doi.org/10.1016/j.ijmm.2013.07.004.
35. Pato CTC. 2011. Streptococcus pyogenes como agente de infecção da pele e tecidos moles. MSc thesis. Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal.
36. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, Thomson NR, Iqbal Z. 2021. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol 19:e3001421. https://doi.org/10.1371/journal.pbio.3001421.
37. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao

Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, et al. 2016. Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.

38. Bioinformatics @ Molecular Microbiology and Infection Unit. 2021. INNUca (v4.2.2). GitHub. https://github.com/B-UMMI/INNUca/releases/tag/v4.2.2. Accessed 7 July 2021.
39. Seemann T. 2021. mlst (v2.19.0). GitHub. https://github.com/tseemann/mlst/releases/tag/v2.19.0. Accessed 11 March 2021.
40. Microbiological Diagnostic Unit Public Health Laboratory. 2021. emmtyper—emm automatic isolate labeller (v0.2.0). GitHub. https://github.com/MDU-PHL/emmtyper/releases/tag/v0.2.0. Accessed 11 March 2021.
41. Seemann T. 2021. Snippy (v4.6.0). GitHub. https://github.com/tseemann/snippy/releases/tag/v4.6.0. Accessed 31 July 2021.
42. Bioinformatics @ Molecular Microbiology and Infection Unit. 2021. SeqTyper (v2.3). GitHub. https://github.com/B-UMMI/seq_typing/tree/gbs_types. Accessed 29 July 2021.
43. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. 2018. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microb Genom 4:e000166. https://doi.org/10.1099/mgen.0.000166.
44. Bioinformatics @ Molecular Microbiology and Infection Unit. 2022. Schema Refinery (v0.1.0). GitHub. https://github.com/B-UMMI/Schema_Refinery/releases/tag/v0.1.0. Accessed 1 February 2022.
45. Mamede R, Vila-Cerqueira P, Silva M, Carriço JA, Ramirez M. 2021. Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas. Nucleic Acids Res 49:D660–D666. https://doi.org/10.1093/nar/gkaa889.
46. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. 2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 33:128–129. https://doi.org/10.1093/bioinformatics/btw582.
47. Ribeiro-Gonçalves B, Francisco AP, Vaz C, Ramirez M, Carriço JA. 2016. PHYLOViZ Online: Web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res 44:W246–W251. https://doi.org/10.1093/nar/gkw359.
48. Severiano A, Pinto FR, Ramirez M, Carriço JA. 2011. Adjusted Wallace coefficient as a measure of congruence between typing methods. J Clin Microbiol 49:3997–4000. https://doi.org/10.1128/JCM.00624-11.
49. Silva-Costa C, Friães A, Ramirez M, Melo-Cristino J, Portuguese Group for the Study of Streptococcal Infections. 2012. Differences between macrolide-resistant and -susceptible *Streptococcus pyogenes*: importance of clonal properties in addition to antibiotic consumption. Antimicrob Agents Chemother 56:5661–5666. https://doi.org/10.1128/AAC.01133-12.
50. Silva-Costa C, Friães A, Ramirez M, Melo-Cristino J. 2015. Macrolide-resistant *Streptococcus pyogenes*: prevalence and treatment strategies. Expert Rev Anti Infect Ther 13:615–628. https://doi.org/10.1586/14787210.2015.1023292.
51. Iannelli F, Santoro F, Santagati M, Docquier J-D, Lazzeri E, Pastore G, Cassone M, Oggioni MR, Rossolini GM, Stefani S, Pozzi G. 2018. Type M resistance to macrolides is due to a two-gene efflux transport system of the ATP-binding cassette (ABC) superfamily. Front Microbiol 9:1670. https://doi.org/10.3389/fmicb.2018.01670.
52. Steer AC, Law I, Matatolu L, Beall BW, Carapetis JR. 2009. Global *emm* type distribution of group A streptococci: systematic review and implications for vaccine development. Lancet Infect Dis 9:611–616. https://doi.org/10.1016/S1473-3099(09)70178-1.
53. Barnett TC, Bowen AC, Carapetis JR. 2018. The fall and rise of group A *Streptococcus* diseases. Epidemiol Infect 147:e4. https://doi.org/10.1017/S0950268818002285.
54. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316. https://doi.org/10.1101/gr.241455.118.
55. Sabat AJ, Budimir A, Nashev D, Sa-Leao R, van Dijl JM, Laurent F, Grundmann H, Friedrich AW, ESCMID Study Group of Epidemiological Markers (ESGEM). 2013. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill 18:20380. https://doi.org/10.2807/ese.18.04.20380-en.