JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Cumulative inaugural

DISSERTATION

for the degree of

Doctor rerum naturalium

(Dr. rer. nat.)

To Strepto and beyond!

Large-scale bacterial genomics with a focus on pathogenic streptococci

---

by

Linda Fenske

Faculty of Biology and Chemistry
Department for Bioinformatics and Systems Biology
Justus Liebig University, Giessen

March, 2026

1st reviewer: Prof. Dr. Alexander Goesmann,
Department for Bioinformatics and Systems Biology,
Justus Liebig University, Giessen Germany

2nd reviewer: Prof. Dr. Tobias Eisenberg,
Department of Veterinary Medicine,
Hessian State Laboratory, Giessen Germany

"If at first you don’t succeed; call it version 1.0."

Contents

List of Figures

List of Tables

Abbreviations

α	alpha
β	beta
γ	gamma
AA	amino acid
AI	artificial intelligence
AMR	antimicrobial resistance
BakRep	Bakterien Repository
bp	base pairs
CC	clonal complex
CLI	command line interface
CNS	central nervous system
COBS	Compact Bit-sliced Signature
cps	capsular polysaccharide
CPU	Central Processing Unit
CRISPR	clustered regularly interspaced short palindromic repeats
de.NBI	German Network for Bioinformatics Infrastructure
DFI	diabetic foot infection
DNA	deoxyribonucleic acid
EMBL-EBI	European Bioinformatics Institute
ENA	European Nucleotide Archive
FAIR	findability, accessibility, interoperability, reusability
FASTA	text-based format for nucleotide or amino acid sequences
GB	gigabyte
GBS	group B streptococci
GTDB	Genome Taxonomy Database
GWAS	genome-wide association study
HGT	horizontal gene transfer
HPC	high-performance computing
HTML	HyperText Markup Language
IAP	intrapartum antibiotic prophylaxis
INSCD	International Nucleotide Sequence Database Collaboration
JSON	JavaScript Object Notation
LPSN	List of Prokaryotic names with standing in Nomenclatur
Mbp	mega base pairs
MGE	mobile genetic element

MLSB	macrolide-lincosamide-streptogramin B
MLST	multilocus sequence typing
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
ONT	Oxford Nanopore Technologies
PBP	penicillin-binding protein
RAM	random-access memory
RKI	Robert Koch Institute
RNA	ribonucleic acid
S.	*Streptococcus*
S3	Simple Storage Service
ST	sequence type
STSS	streptococcal toxic shock syndrome
TB	terabyte
TSV	tab-separated value
UK	United Kingdom
WGS	whole-genome sequencing
WHO	World Health Organization

Abstract

Bacteria are central research organisms across many disciplines and capturing their diversity is essential for understanding their genomic features, which in turn illuminate virulence, transmission, and the rapid evolution and spread of antimicrobial resistance. Advances in whole-genome sequencing has driven an explosion of publicly available data, with thousands of new datasets deposited daily. However, despite broad data availability, major hurdles remain in terms of accessibility and comparability due to heterogeneous processing pipelines, inconsistent quality control, and incomplete or unstructured metadata.

Streptococcus agalactiae is an opportunistic multi-host pathogen with major relevance in both human and veterinary medicine. Beyond humans and cattle, *S. agalactiae* has also been reported in endangered species such as elephants, but elephant-associated strains lack detailed genomic characterization to date. Although extensively studied, existing genomic work is often restricted to specific regions or single outbreaks, offering only fragmented views of its true diversity.

This thesis addresses these challenges through three linked projects, motivated by an interest in the diversity of *S. agalactiae*. Initial attempts to perform a large-scale comparative analysis of *S. agalactiae* were hampered by the lack of suitable reference datasets. To enable robust comparative genomics from uniformly processed public data, BakRep was developed: A large-scale, searchable web repository built on the assemblies from the *AllTheBacteria* project. BakRep connects consistent genome-based characterizations, like taxonomic information, subtypings and annotations, with descriptive metadata and provides an integrated search interface complemented by interactive visualizations of genomic features. As a use case, BakRep was used to conduct a population-scale comparative analysis of all *S. agalactiae* genomes in the repository, confirming dominant stable lineages while exposing substantial metadata gaps that limited biological interpretation and highlighted the need for curated, structured metadata. Additionally, isolates from elephant-derived *S. agalactiae* were analyzed, to address the present date deficit.

These isolates were phylogenetically distinct from strains found in other hosts, and several lineage-specific genes suggested potential niche adaptation.

Together, this work demonstrates how consistent, scalable resources support large-scale comparative studies across thousands of bacterial species to identify shared patterns of adaptation, virulence, as well as resistance, to guide focused follow-up research in lesser-studied hosts and ecological niches. While comparative genomics is crucial for understanding genetic variation and host adaptation in multi-host pathogens like *S. agalactiae*, its impact is undermined when sequencing data are generated and shared without well-curated metadata. Such metadata gaps present significant obstacles to biological interpretation and clinical translation, thereby necessitating rigorous data and metadata quality control. Furthermore, this work identified opportunities to enhance methods devised here, highlighted new research questions, and set a foundation for future investigations.

Zusammenfassung

Bakterien sind zentrale Organismen in verschiedensten Forschungsdisziplinen, und die Erfassung ihrer Vielfalt ist entscheidend, um ihre genomischen Merkmale zu verstehen. Diese Erkenntnisse wiederum beleuchten Virulenz, Übertragungswege sowie die rasche Entwicklung und Verbreitung von Antibiotikaresistenzen. Fortschritte in der Gesamtgenomsequenzierung haben zu einer explosionsartigen Zunahme öffentlich zugänglicher Daten geführt, wobei täglich Tausende neuer Datensätze hinzukommen. Trotz der weitreichenden Datenverfügbarkeit bestehen jedoch oft noch große Hürden hinsichtlich der Zugänglichkeit und Vergleichbarkeit, die auf heterogene Verarbeitungsprozesse, uneinheitliche Qualitätskontrollen und unvollständige oder unstrukturierte Metadaten zurückzuführen sind.

Streptococcus agalactiae ist ein opportunistischer Erreger mit einem breiten Wirtsspektrum, der sowohl in der Human- als auch in der Veterinärmedizin von großer Bedeutung ist. Neben Menschen und Rindern wurde *S. agalactiae* auch bei gefährdeten Arten wie Elefanten nachgewiesen, doch für mit Elefanten assoziierte Stämme liegen bisher keine detaillierten genomischen Charakterisierungen vor. Obwohl umfangreiche Forschung durchgeführt wird, beschränken sich bestehende genomische Arbeiten häufig auf bestimmte Regionen oder einzelne Ausbrüche und bieten nur einen bruchstückhaften Überblick über die tatsächliche Vielfalt des Erregers.

Diese Arbeit befasst sich mit diesen Herausforderungen anhand von drei miteinander verbundenen Projekten, die durch das Interesse an der Diversität von *S. agalactiae* motiviert sind. Um eine robuste vergleichende Genomanalyse aus einheitlich verarbeiteten öffentlichen Daten zu ermöglichen, wurde BakRep entwickelt, ein umfangreiches, durchsuchbares Repository, das auf den Assemblies des *AllTheBacteria* Projekt basiert. BakRep verbindet konsistent genomische Charakterisierungen, wie taxonomische Informationen, Subtypisierungen und Annotationen, mit deskriptiven Metadaten und bietet eine integrierte Suchschnittstelle ergänzt durch interaktive Visualisierungen genomischer Merkmale. Als Anwendungsbeispiel wurde BakRep genutzt, um eine populationsweite vergleichende Analyse aller, im Datensatz enthaltenen, *S. agalactiae* Genome durchzuführen, wodurch dominante und stabile Linien bestätigt

wurden, während gleichzeitig erhebliche Metadatenlücken aufgedeckt wurden, die die biologische Interpretation einschränkten und die Notwendigkeit kuratierter, strukturierter Metadaten hervorhoben. Darüber hinaus wurden Genome von aus Elefanten stammenden *S. agalactiae* Isolaten untersucht. Diese Isolate unterschieden sich phylogenetisch von Stämmen die in anderen Wirten identifiziert wurden und mehrere stammspezifische Gene deuteten auf eine mögliche Nischenanpassung hin.

Zusammengenommen zeigt diese Arbeit, wie konsistente, skalierbare Ressourcen wie BakRep große Vergleichsstudien über Tausende von Bakterienarten hinweg fördern können, um gemeinsame Muster der Anpassung, Virulenz und Resistenz zu identifizieren und gezielte Folgestudien in weniger untersuchten Wirten und ökologischen Nischen anzuregen. Während die vergleichende Genomik für das Verständnis der genetischen Variation und der Wirtsanpassung bei Pathogenen mit breitem Wirtsspektrum wie *S. agalactiae* von entscheidender Bedeutung ist, wird ihre Bedeutung untergraben, wenn Sequenzierungsdaten ohne gut gepflegte Metadaten generiert und geteilt werden. Solche Metadatenlücken sind ein großes Hindernis für die biologische Interpretation und die klinische Anwendung, weshalb eine strenge Daten- und Metadatenkontrolle unerlässlich ist. Weiterhin ergaben sich aus dieser Arbeit konkrete Möglichkeiten zur Verbesserung der entwickelten Ansätze, wodurch neue Forschungsfragen aufgezeigt und eine Grundlage für zukünftige Studien geschaffen wurden.

1. Motivation and outline

"If you don't like bacteria,
you are on the wrong planet."

(Stewart Brand - 2010)

Prokaryotes, including bacteria, are considered the oldest life forms on Earth and constitute the most prevalent cellular organisms, playing a significant role in maintaining the balance and functionality of every biome [1, 2]. Currently, prokaryotic organisms are estimated to account for approximately 60 % of Earth's biomass, displaying markedly greater genetic, metabolic, and physiological diversity than plants and animals [3]. Serving as the main drivers of global biogeochemical processes, prokaryotes are found in almost all ecosystems and represent a largely untapped reservoir of genetic and metabolic diversity. Exploring this resource holds great promise for society, and may deliver benefits in areas such as improved food production, medicine, waste bioremediation, and agriculture [1, 4]. Despite this dominance, an estimated 42 % of bacterial diversity currently has no genomic representation and only about 10 % of bacterial taxa can be successfully cultured in diagnostic laboratories [2]. This is because they may grow too slowly to be detected, require tightly defined growth conditions that are difficult to reproduce, or have their proliferation suppressed by competing microorganisms or by inhibitory substances produced by other bacteria [5, 6]. Furthermore, existing molecular tests often do not capture emerging genetic traits in rapidly evolving pathogens [7]. But still, molecular biology has improved considerably through studies on prokaryotes, especially bacteria, which have provided important perspectives on the fundamental components of life: DNA, RNA, and proteins, as well as key biological processes such as gene expression, cell division and horizontal gene transfer (HGT) [8].

For researchers in clinical microbiology and public health, examining bacterial genome diversity and the dynamics of their functional elements is essential [9]. Although most bacteria are harmless, or even beneficial, and support human health, a small fraction are pathogenic, accounting for many serious diseases. In fact, only about 7 % of all validly described bacterial species are recognized as human pathogens [10]. Whether a bacterium causes disease, however, is influenced by many traits, including virulence factors (e.g., toxins and surface proteins) that enable host invasion, as well as host characteristics such as genetics, lifestyle, and immune status [11]. Moreover, it is also assumed that pathogenicity determinants are specific features that enable microorganisms to cause disease, while not being unconditionally essential for their basic survival [12].

The introduction of next-generation sequencing (NGS) ushered in an era of high-throughput genomics, enabling the study of bacterial genomes at resolutions previously unattainable. Nowadays, advances in sequencing technologies, coupled with their growing accessibility, have led to a rapid expansion in the scale and depth of bacterial genome data generation [13, 14]. DNA sequencing ranges from high-resolution approaches that sequence large genomic regions to low-resolution methods targeting a few conserved genes. For example, one widely used low-resolution approach is multilocus sequence typing (MLST), which types bacterial isolates by sequencing DNA fragments (~450 – 500 bp) of typically five to seven housekeeping genes. Each unique sequence is assigned an allele number, and the combination of alleles defines the sequence type (ST). This method allows for billions of possible genotypes and facilitates global comparisons of isolates using centralized online databases [15–17]. Similarly important is monitoring the rapid emergence of antimicrobial resistance (AMR). Bacterial genome sequencing enhances surveillance, improves our understanding of resistance mechanisms, and provides insight into the dissemination of resistance genes, therefore supporting a more effective use of current drugs [18].

Nowadays, bacterial genome data from around the world are more accessible, providing an invaluable resource for microbiologists [19]. However, a noticeable bias favors clinically or industrially relevant bacterial species, while many other species and ecological niches remain underrepresented [20, 21]. Half of all publications these days focus on just a few species, while almost 74 % of all known species have never been

discussed in a scientific study [22]. This skewed representation means that the true diversity of bacterial populations remains overlooked, creating significant gaps in our understanding of their evolution [23]. Consequently, obtaining suitable comparative data for some scientific analyses proves challenging, particularly for isolates from less common host species and ecological niches.

The research domain of microbial genomics applies modern sequencing technologies to study the complete genomes of microorganisms. Additionally, microbial genomics supports studies on microbial diversity in many different clinical and industrial research areas [24]. These approaches enable detection, characterization, and functional analysis of microbes, promoting our understanding of diseases, AMR, and subsequent drug development. During my Master's thesis, which involved a comparative genomic analysis of *Streptococcus (S.) uberis*, a major cause of bovine mastitis, I was first introduced to the field of microbial genomics. This work sparked my interest in the analysis of streptococcal pathogens and their complex ecology. While *S. uberis* is typically framed as an opportunistic pathogen associated with environmental reservoirs, streptococci of the species *S. agalactiae* merit particular attention as important agents more closely tied to mucosal colonization. Although *S. uberis* also inhabits mucosal sites, *S. agalactiae* is more prominently associated with these niches and manifests remarkable host adaptability, with an even greater zoonotic potential. While *S. uberis* is mainly known as a pathogen in dairy cattle, *S. agalactiae* exemplifies how streptococcal species can cross the host barrier, affecting both humans and animals. This switch from a mainly host-restricted to a multi-host pathogen raised my interest in the mechanisms that drive such zoonotic potential, and motivated me to extend my research focus toward *S. agalactiae*. Therefore, the initial idea for my PhD thesis was to perform a comprehensive analysis of *S. agalactiae* isolates from different host species, with a particular focus on its zoonotic potential.

While searching public databases for appropriate comparative datasets, I repeatedly encountered genomes generated with different pipelines, inconsistent quality control, and vast amounts of unprocessed raw data accompanied by poorly organized metadata. At one point, I came across the publication of Blackwell et al., which provided an extensive collection of uniformly processed bacterial genome assemblies [19]. This work motivated the systematic characterization of the assemblies to create a uniformly

processed, consistently annotated resource that supports robust comparative analyses across several bacterial taxa, thereby moving even beyond streptococci. What started as a side project soon evolved into an independent project and ultimately became a central component of this thesis. At the same time, a parallel project was initiated to analyze *S. agalactiae* isolates from elephants, a previously rather unexplored host species lacking any comparative data so far. After establishing the standardized genome dataset, it was applied to a focused use case within streptococci: an in-depth characterization of *S. agalactiae*, intending to deepen the understanding of its genomic diversity, virulence, and epidemiology. This species served as an ideal example to demonstrate how the resource can be leveraged to systematically explore genomic diversity, virulence-associated traits, and epidemiological patterns at scale. It also illustrates an analytical approach that can be applied to many other bacterial species. This thesis, *"To Strepto and beyond!"*, reflects how the project evolved from an initial focus on streptococci into a much more expansive effort.

The following background Chapter 2.1 offers a brief introduction to the genus *Streptococcus*, a diverse group of bacteria able to colonize and infect a wide range of hosts and tissues. Followed by a brief summary of *S. uberis* as preliminary work for this thesis in Chapter 2.1.1 and more detailed insights into *S. agalactiae*, as the pathogen of main interest in the work presented here, in Chapter 2.1.2. Chapter 2.2 addresses the availability of bacterial genomes and the challenges posed by continuous sequence data growth. Having introduced the key background of this thesis, Chapter 3 outlines the resulting research questions and the objectives addressed in this work. Based on this, Chapter 4 presents the first-autorships that are included in this thesis, followed by a presentation and discussion of the resulting outcomes in Chapter 5. It also provides a brief outlook on future research directions, proposing strategies to address remaining challenges and improve current approaches. Finally, Chapter 6, summarizes the thesis with a concluding statement. In Chapter 7 the scientific contributions achieved throughout this work are listed. The appendix (A) enumerates the tool calls and the parameters used in the the bioinformatic analyses.

2. Background

> "The enemy was the microbial world, and over the centuries, it has killed more people than all of man's wars combined."
>
> ***(Tess Gerritsen – 1999)***

2.1. The genus *Streptococcus*

Several streptococcal species, like *S. pneumoniae*, *S. pyogenes*, or *S. agalactiae*, contribute to millions of deaths worldwide, driven by their high virulence and severe impact on human and animal health [25]. Recent reports of streptococcal infections highlight the ongoing relevance of these bacteria for disease-control. In December 2022, the World Health Organization (WHO) reported a rise in severe and sometimes fatal invasive human *S. pyogenes* infections across at least five European countries [26]. In 2024, the Robert Koch Institute (RKI) documented the highest absolute value of *S. pneumoniae* infections in Germany in the past decades [27]. In Japan, cases of life-threatening human *S. pyogenes* infections have recently reached record levels, where in just the first two months of 2024, 378 cases of streptococcal toxic shock syndrome (STSS) were described, compared to 941 cases in all of 2023 [28]. Furthermore, *S. uberis* and *S. agalactiae* are among the leading pathogens causing bovine mastitis, accounting for an estimated 25–50 % of global cases, which remains the most widespread and costly disease affecting the dairy industry [29, 30].

Streptococci are members of the lactic acid bacteria and are commonly found in various warm-blooded animals, including humans. They represent the dominant species in the oral cavity and upper respiratory tract but can also be widespread throughout the whole human body [31–33]. The LPSN (List of Prokaryotic names with standing in Nomenclatur) currently lists 139 validly published streptococcal

species, many of which are capable of causing invasive disease [34]. Based on their disease manifestations, pathogenic streptococcal species can be broadly divided into three groups: those that commonly cause infections, commensal species, and zoonotic species, with bidirectional transmission between animals and humans, that can cause infections under certain conditions [33]. However, streptococcal infections are not considered classical zoonoses. Although disease symptoms appear similar across host species, streptococci are highly host-adapted, meaning infections typically rely less on crossing host barriers and more on host-specific virulence mechanisms. Moreover, distinct genotypes and pathotypes can exist within a single species [35]. Nevertheless, while most species display host preference, sporadic animal-to-human transmission can occur [36, 37]. Several comparative studies showed that genotypic and phenotypic analyses often identify distinctions between strains responsible for human infections and those from animals [38–40]. Most species are considered commensals, typically residing on mucosal surfaces, but they can cause both localized and systemic infections under appropriate conditions [31].

In clinical laboratories, streptococci are traditionally classified using a phenotypic approach, based on their properties to lyse red blood cells when cultured on blood agar: Producing greenish discoloration from partial red blood cell lysis (α-hemolysis), producing clear zones from complete lysis (β-hemolysis), or showing no hemolysis (γ-hemolysis) [31]. Many α-hemolytic species exhibit low virulence, with *S. pneumoniae* and *S. suis* being notable exceptions, whereas β-hemolysis is typically associated with more virulent strains [31, 41]. The most β-hemolytic species can be further subdivided using the Lancefield serogrouping scheme, a system introduced by Rebecca Lancefield in 1933 [42]. This approach assigns isolates to groups (A-W) based on the presence of a specific C-polysaccharide in their cell wall (Figure 2.1). In contrast, many α-hemolytic streptococci lack a typeable group antigen or do not express one at detectable levels, and therefore cannot be assigned to a Lancefield group in routine diagnostics [43, 44]. While serogrouping and hemolysis patterns remain helpful for rapid preliminary identification, the reliability of these classifications for precise taxonomic resolution has diminished with the growing number of described streptococcal species.

The advent of DNA sequencing has enabled more precise species-level identifications, leading to the current classification of streptococci into eight phylogenetic groups based on gene gain and loss analyses: mitis, sanguinis, anginosus, salivarius, downei, mutans, pyogenic, and bovis [32, 45]. The primary pathogenic species include the pyogenic group (e.g., *S. pyogenes*, *S. agalactiae*, *S. uberis*), along with *S. pneumoniae* from the mitis-group [31].

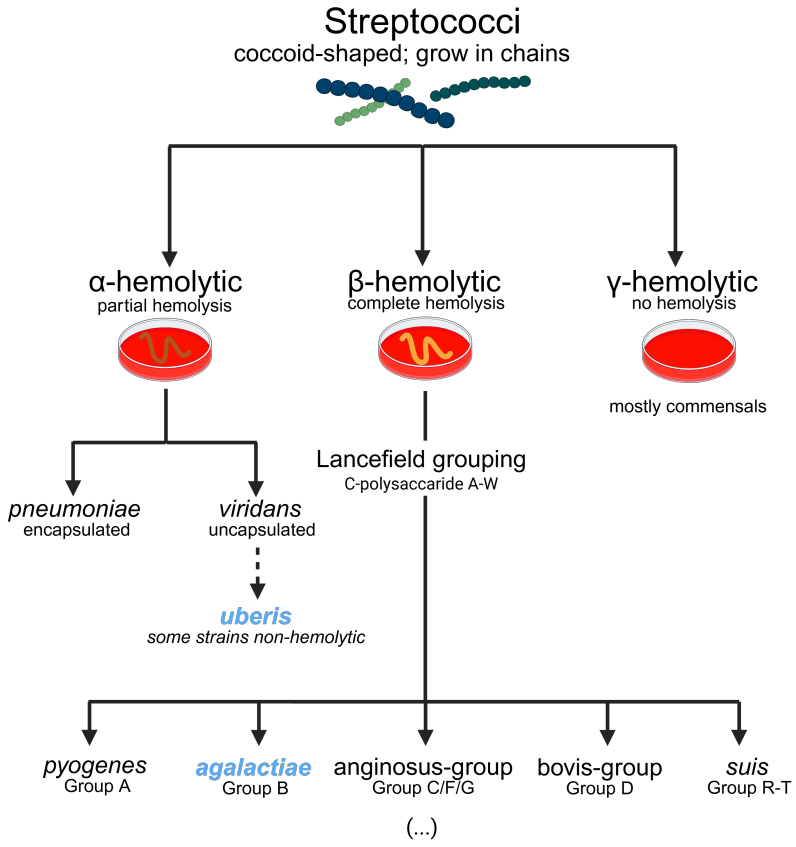


Figure 2.1.: Classification of streptococci by their hemolysis patterns and surface antigens. Initial classification is based on hemolysis on blood agar. Following this, streptococci are further classified by capsule production or surface proteins. This classification is not absolute, as some species deviate from these patterns. Species of main interest for this work are marked in blue. (*Figure created with BioRender.com*)

2.1.1. Preceding research: *Streptococcus uberis*

S. uberis is a major environmentally associated pathogen responsible for bovine mastitis, a disease that causes major economic losses in the dairy industry [46]. Mastitis caused by *S. uberis* can manifest in both subclinical and clinical forms, with the latter frequently leading to mild-to-severe visible inflammation [47]. Building on this background, before starting my PhD, I conducted a comparative genomic analysis of 24 *S. uberis* isolates obtained from three German dairy farms. Two of the farms experienced sporadic cases with predominantly mild infections, whereas the third farm experienced numerous infections within a short time period. The genomic comparison included virulence genes, AMR genes, and prophage regions, but no specific features were identified that could explain the severe course observed on one of the farms. However, nearly all isolates from this farm shared the same previously undescribed MLST profile (ST1373), suggesting a possible clonal outbreak. In summary, the results indicated that pathogenicity in *S. uberis* cannot be explained solely by gene content but is apparently influenced by host conditions and environmental factors. This work introduced me to the diversity and adaptability of streptococcal pathogens. Understanding how such bacteria evolve and adapt across hosts became a central motivation for my subsequent research.

2.1.2. *Streptococcus agalactiae* (GBS)

S. agalactiae is a β-hemolytic *Streptococcus* and the only member of the Lancefield group B, which is why it has earned the common name group B streptococci (GBS). Like other streptococci GBS strains are non-motile, coccoid-shaped, and typically form chains when growing in liquid media [31] (Figure 2.2). GBS was initially recognized as the causative agent of bovine mastitis in the late 19th century, leading to the name "*agalactiae*", which literally translates as "*without milk*" [48, 49]. It is a common agent of mastitis in dairy herds but rarely leads to systemic disease, unlike other mastitis-causing pathogens. In most cases, GBS infections are subclinical and difficult to detect, as they neither causes visible changes at the udder, nor changes the appearance of the milk [50].

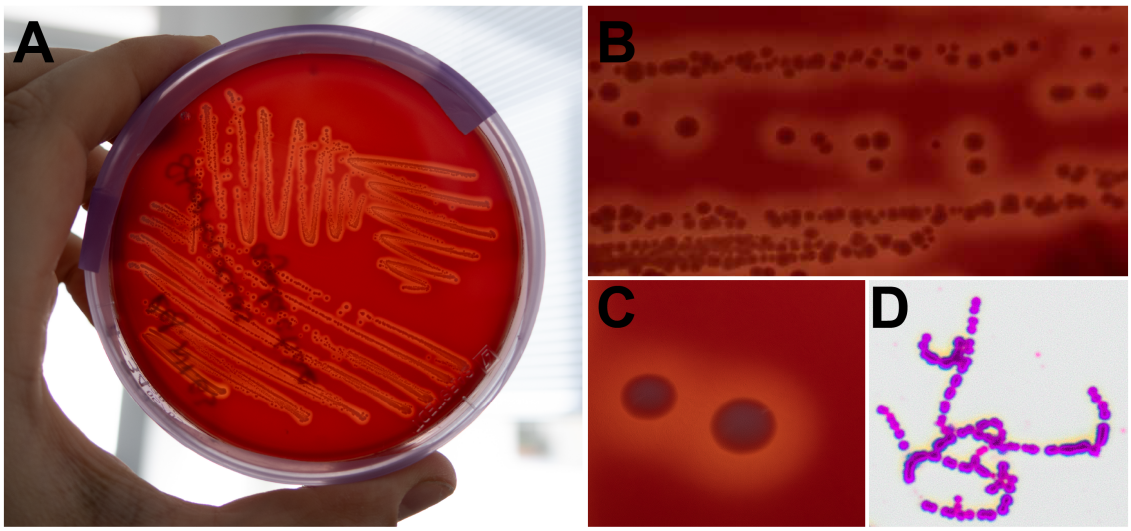


Figure 2.2.: GBS appearance on sheep blood agar **(A)**, colonies in clusters **(B)**, individual round-shaped colonies **(C)**, and chains in liquid medium **(D)**. Transmitted light on agar shows β-hemolysis. (*Photos: Linda Fenske/Hessian State Laboratory*)

Since the 1970s, GBS has also been identified as a major pathogen responsible for invasive neonatal infections [37]. An estimated 20–30 % of pregnant women carry GBS asymptomatically in their vaginal tract, posing a risk of transmission to the newborn during childbirth, where it may turn into a deadly pathogen [51]. Approximately 50 % of newborns become colonized, with around 1 % developing invasive disease, which may result in pneumonia, meningitis, or septicemia [52]. Despite the implementation of several screening programs in many countries, GBS continues to be the leading cause of neonatal invasive infections in high-income nations [50]. While GBS research has largely centered on invasive disease in newborns, adults account for more than half of GBS-attributable deaths in the United States [53]. GBS is often carried asymptomatically in the urogenital or intestinal tract. However, in adults, particularly in the elderly or immunocompremised, it can also cause infections such as sepsis, meningitis, pneumonia, or endocarditis [54, 55]. Additionally, GBS is among the leading pathogens isolated from diabetic foot infections (DFIs) [56, 57]. A study published in 2021 even proposes the hypothesis that GBS may be associated with Alzheimer's disease [58].

While the majority of GBS research focuses on human and bovine infections, it has also been linked to pyogenic diseases in other dairy species, like sheep [59], goats [60] and camels [50], as well as a multitude of additional, even exotic and wildlife species like rats [61], crocodiles [62], aquatic species such as frogs, seals, and dolphins [63, 64], and even elephants [65]. The ability of GBS to infect these and even a wider range of host species is ascribed to its high genomic plasticity, which promotes the integration of various accessory genes into its chromosome. This ability allows the organism to establish itself in a wide range of ecological niches, conferring an evolutionary benefit [66, 67]. Although a high number of streptococci are strictly host specific, cross-species transmission of GBS is periodically documented, but the full zoonotic potential has not yet been completely elucidated. A well-known case of dissemination occurred in Singapore in 2017, when healthy adults developed infections after the consumption of raw fish [68]. This was also the first case, describing a major foodborne outbreak linked to GBS. However, several studies also highlight strong host adaptation, defined by specific STs, which can further be clustered into clonal complexes (CCs) of related isolates that likely share a common ancestor [69]. Isolates belonging to CC17, for instance, are recognized as a hypervirulent lineage increasingly linked to neonatal infections [70, 71], while CC61 is mainly associated with bovine infections [72]. But, studies have likewise shown the presence of identical STs across different host species. CC1, including its namesake ST1, for instance, has been identified in multiple studies in both human and cattle isolates [73, 74], while CC283 was found in humans and fish [68]. CC23 has likewise been found across multiple host species, including cows [66], crocodiles [62], and seals [63].

Such host versatility is likely facilitated by an array of virulence-associated traits, the most prominent of which are summarized in Table 2.1. Virulence genes can be categorized according to their function into four categories: Regulation, Adherence, Invasion, and Immune evasion [67]. One key immune evasion factor and potential vaccine target of GBS is the sialic acid-rich capsule, which has a vital role in the bacterium's ability to persist and survive within the host [75]. This capsule is encoded by the capsular polysaccharide (*cps*) locus, consisting of 16 to 18 genes [76, 77]. Variations within these gene sequences serve as the basis for GBS serotyping. To date, ten different serotypes are recognized (Ia, Ib, II-IX), with serotype III additionally being subdivided into four subtypes (III-1-III-4) [78].

The prevalence and distribution of serotypes differ depending on geographic region, host species, and clinical presentation, with certain serotypes also linked to varying levels of virulence [79]. For example, serotype III seems to be responsible for most cases of meningitis in neonates [80, 81], whereas serotypes Ia and V are more commonly associated with invasive infections in non-pregnant adults [23, 55]. In cases of bovine infections, the serotypes differ significantly by geographic region. Studies conducted in Argentina, New York State, Brazil and Iran identified serotype III as the most prevalent [82–85], while studies from China found serotype Ia in most bovine mastitis cases [86, 87]. A study from Denmark reported serotype V predominating, which was previously only found in human isolates [88]. Some GBS isolates remain phenotypically non-typeable because they do not react with available anti-capsular sera, which may result from low or absent capsule expression or from polysaccharide variants unrecognized by current antisera [89]. The growing diversity of GBS serotypes, combined with the potential for capsular switching, poses a major challenge for vaccine development, since *cps*-based vaccines may exert selective pressure that enables virulent genotypes to escape coverage. Different studies already demonstrated such switching from serotype III to IV within the hypervirulent CC17 lineage, underscoring that even highly conserved clones can alter one of their main vaccine targets [90, 91]. The likelihood of this occurring may increase if horizontal transfer of capsular operons between strains, or even across different host species, becomes more frequent. Furthermore, certain serotypes are often associated with specific CCs, illustrating a relationship between capsular type and genetic lineage. For example, ST1/V is likely accountable for a high percentage of bloodstream infections in nonpregnant adults [92, 93], while CC17/III represents the most virulent lineage, being responsible for the majority of neonatal meningitis cases [71]. A recent study from Brazil, examining bovine mastitis cases, reported an association of CC103 with serotype Ia and CC91 with serotype III [94].

Moreover, some lineages are also linked to elevated levels of AMR, which is an increasing concern [94, 95]. More than 80 % of GBS strains exhibit resistance to tetracyclines [96]. This is mainly caused by the acquisition of resistance genes such as *tetM* or *tetO* by certain GBS clones, due to the extensive use of tetracyclines since 1948, which has contributed to the clonal expansion [97]. As the majority of clinical GBS isolates have been classified as susceptible to β-lactam antibiotics,

these agents remain the current first-line choice for both treatment and prevention of GBS infections, while clindamycin and erythromycin are commonly prescribed for patients with penicillin allergies. However, intrapartum antibiotic prophylaxis (IAP) is given to expectant mothers based on established risk factors or screening protocols, pointing to the need for ongoing surveillance of resistance trends [96]. For mastitis cases, the extensive use of antimicrobials, including broad-spectrum and critically important antimicrobials, drives resistance and limits therapeutic options [98]. In recent years, resistance to clindamycin and erythromycin has risen [99], and reduced susceptibility to penicillin has also been observed, mainly linked to several amino acid (AA) substitutions in the penicillin-binding proteins (PBPs), which impair antibiotic binding [96, 100, 101]. Continuing efforts to develop vaccines against GBS aim not only to prevent disease but also to reduce the preventive use of antibiotics. But, achieving this goal necessitates a thorough comprehension of the clinical and microbiological characteristics of GBS [102].

Multiple virulence and resistance genes are further linked to known or suspected mobile genetic elements (MGEs). In many bacterial pathogens, shifts into new hosts or ecological niches have been strongly connected to the uptake of a broad range of MGEs [50]. Genetic exchange between coexisting streptococcal species is common and significantly shapes their pathogenic potential. MGEs transferred through HGT drive species diversity, host adaptation, and the emergence of virulence traits [66]. Moreover, certain AMR determinants are often carried on conjugative transposons and can spread swiftly across GBS lineages, amplifying the risk of AMR dissemination [103].

Table 2.1.: Overview of the key virulence factors of GBS, outlining their role in pathogenicity and their biological function.

Category	Gene	Function	Role in pathogenicity
Regulation	*covR/covS*	two-component system CovS/CovR	Global virulence regulator controlling hemolysis, epithelial adherence, invasiveness, and colonization.
Adherence	*scpB, pavA*	C5a peptidase	Reduces chemotaxis and neutrophil recruitment, promoting colonization and immune evasion.
	fbsA, fbsB	fibrinogen-binding proteins	Binds fibronectin, enhancing epithelial adherence and colonization.
	lmb	laminin-binding protein	Binds laminin, promoting invasion into microvascular endothelial cells.
	hvgA	HvgA adhesin	Mediates intestinal colonization and barrier crossing; linked to meningitis.
	PI-1, PI-2a, PI-2b	Pili/Pilus islands	Filamentous adhesins supporting biofilms and persistent infection.
Invasion	*cyl*	β-hemolysin/cytolysin	Pore-forming cytotoxin causing tissue damage and promoting bloodstream/CNS invasion in neonatal meningitis.
	cfb	CAMP factor	Forms host-cell pores, driving tissue injury and invasion.
	hylB	Hyaluronate lyase	Degrades extracellular matrix, promoting tissue infiltration and systemic dissemination.
Evasion	*cps*	Polysaccharide capsule	Inhibits complement opsonization, evading phagocytosis and promoting bloodstream survival.
	rib	Surface proteins	Resistance to proteases; enabling mucosal colonization and infection.

2.2. The availability of bacterial genomes

Over the past decades, bacterial whole-genome sequencing (WGS) has greatly expanded our knowledge of microbial life by revealing the immense diversity within populations as well as the elaborate complexity of ecosystems [7, 104]. The growing availability of genome data has given rise to a new discipline that merges informatics with biological research: bioinformatics [105]. Today, DNA-based bioinformatics has become indispensable for the field of microbial genomics, and access to cost-effective sequencing platforms have accelerated bacterial WGS to a new level [7]. Sequence-based analyses have become an essential tool for public health tasks, such as detecting bacterial outbreaks, tracking AMR, or simply monitoring bacterial pathogens [106]. Today, high-throughput sequencing has produced vast amounts of bacterial genomic data from around the world, much of which is openly deposited in public repositories, creating a uniquely valuable resource for research [19]. Moreover, the submission of genomic data to public repositories has become a central and often mandatory step, as most journals now require accession numbers as a prerequisite for publication [107]. Due to this, public databases comprise tens of thousands of bacterial genomes, and there is no end in sight to this data growth. For example: In 2024, the International Nucleotide Sequence Database Collaboration (INSCD) published, that it currently contains 32 million NGS runs and 4.76 billion assembled sequences [108]. And at the end of 2025, the GenBank release 269.0 (published in 12/2025), maintained by the National Center for Biotechnology Information (NCBI), stated that an average of 12 952 sequence datasets are added or updated per day, with the repository containing 4.5 billion WGS records with more than 42 trillion base pairs [109] (Figure 2.3A+B).

Although most sequence data is publicly available, certain difficulties with data accessibility persist. The accelerating generation of sequence data, has shifted the trend away from submitting complete or draft genomes to predominantly depositing raw sequence reads. These typically require extensive preprocessing before any further analysis [19]. This trend is also evident in GenBank, where contig- and scaffold-level assemblies by now outnumber fully complete genomes (Figure 2.3C).

Consequently, comparing WGS data analysis results is increasingly challenging. On the path from raw sequencing data to tangible biological information, several factors act as critical bottlenecks. Each sequencing run can generate terabytes (TBs) of data, demanding efficient strategies for storage, processing, and analysis. The massive output of NGS is increasingly exceeding the capacity of traditional storage systems, making scalable and accessible cloud-based solutions increasingly essential [110]. Moreover, NGS data analysis involves multiple steps, and manual data processing is not only time-consuming but also compromises comparability, as diverse workflows often employ different software tools, parameter settings, and reference datasets. Ideally, to avoid any bias and to ensure reproducibility, every dataset in a project has to be processed with identical tools, software versions, and parameters. Achieving this typically requires reprocessing large portions of the data to avoid unwanted variability arising from mixed tools, settings, and databases [19, 111]. In addition, many researchers lack access to computational infrastructure needed for large-scale genomic analyses. Although cloud- and high-performance computing (HPC) systems offer powerful resources, their use is often limited by cost, access, or technical expertise [110]. Another major challenge is the need for specialized bioinformatics expertise, including programming skills and the use of varied computational tools to generate precise and meaningful results [112]. Yet, the value of genomic datasets exists not only in the sequences themselves but also in the accuracy and completeness of the accompanying metadata. A sequence uploaded in a database that gives no additional information, like isolation time or source, is often practically worthless for comparative examinations and reliable interpretations. To improve the availability and usability of (meta)-data in public repositories, the FAIR (**f**indability, **a**ccessibility, **i**nteroperability, **r**eusability) principles were introduced in 2016 and have since been widely adopted. Accordingly, many databases request and collect metadata associated with the uploaded data. Still, filtering and searching this metadata to generate specific bacterial cohorts for downstream comparative analyses is a tedious task.

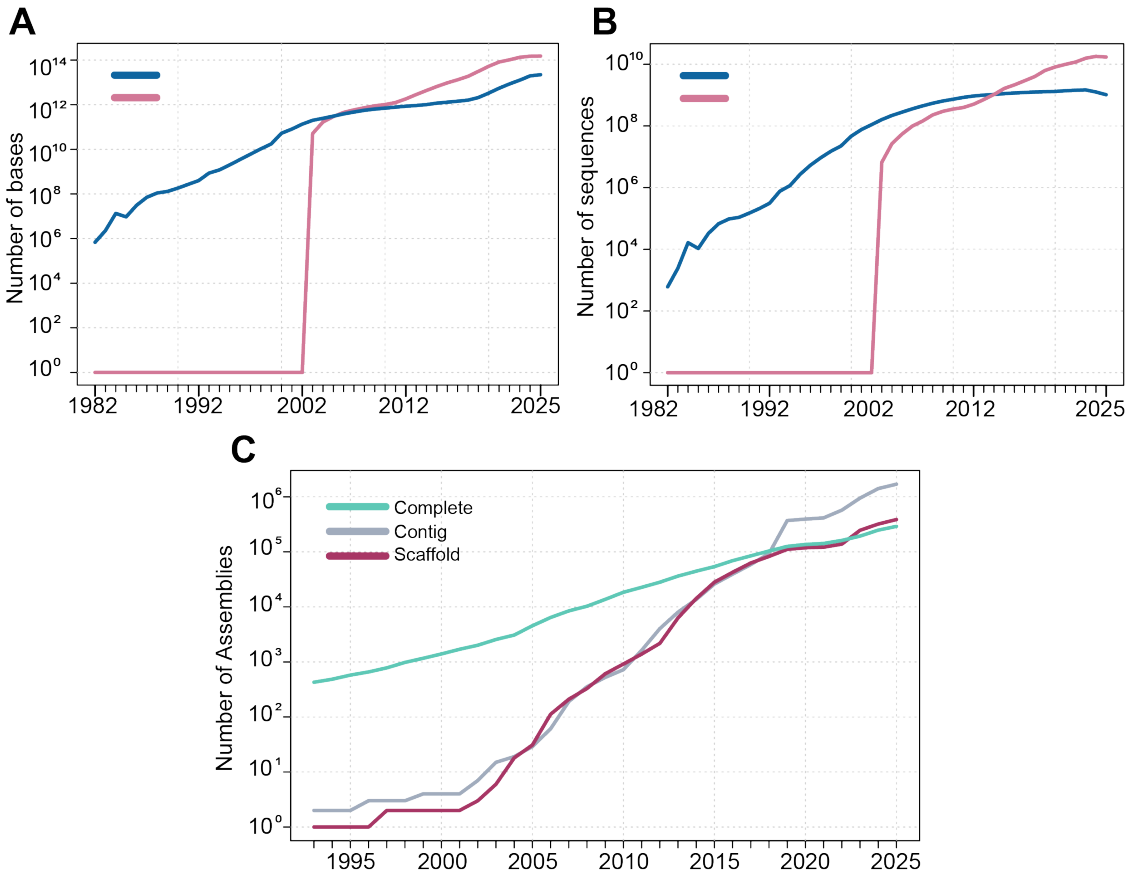


Figure 2.3.: Sequence data growth in the NCBI. The number of bases **(A)** and the number of sequences **(B)** in each GenBank (blue) and WGS (pink) record listed in the NCBI. **(C)** Number of genome assemblies at different assembly levels in the GenBank section of the NCBI [113, 114].

To address some of the beforementioned challenges, the team of Zamin Iqbal at the European Bioinformatics Institute (EMBL-EBI) in London began taking steps toward a more uniform processing of public bacterial genome data. In 2021, they reported the systematic assembly of all bacterial genomes available in the European Nucleotide Archive (ENA) as of November 2018. This initial effort resulted in a collection of 661 405 standardized assemblies, including 311 006 that had never been assembled before [19]. To improve accessibility and usability, they provided several indices along with the assemblies: the Compact Bit-sliced Signature (COBS) index [115], which enables users to search for sequences by breaking them into constituent k-mers, the MinHash index [116] allowing for the identification of similar genomes, and an index generated using the library sketching function of PopPUNK [117], which provides

precomputed distances between the assemblies. Nevertheless, users without advanced bioinformatics expertise, command line interface (CLI) skills, or access to larger compute infrastructure still face serious challenges in locating and extracting genomes or genes of interest from those assemblies. All genome FASTA files were bundled into a single archive, and even though the assemblies are pre-indexed, there remains a barrier to findability and accessibility. Furthermore, the authors themselves stated that they had not included additional valuable information beyond the assemblies, such as gene annotations or MLST. Genomic sequences alone have limited value for research if we cannot identify which genes are present. When the *661k* dataset was published, there was already great anticipation that this vast resource would be widely utilized. However, it soon became evident that analyzing and processing such an enormous volume of data is a major task. So, it became clear that this large-scale project would benefit from contributions from the wider scientific community.

3. Aim and objectives of this thesis

This thesis is based on three interconnected projects, all driven by an interest in the zoonotic potential of GBS as a multi-host pathogen. Suitable comparative data for the initially planned GBS analysis were either lacking or inconsistent. In addition, such data were often processed in a non-reproducible way. Therefore, the *661k* dataset generated by Blackwell et al. served as the basis for creating a consistently characterized dataset meant to increase the consistency of sequenced bacterial genomes stored in public data repositories. Additionally, GBS isolates from elephants were analyzed, a rarely studied host species with no prior public comparative data. The first-mentioned dataset allowed for a renewed analysis of GBS. This led to another objective focused on a comprehensive comparison of all GBS genomes contained in this dataset. The following sections provide a brief overview of the requirements for each task, sorted by publication date:

1.) Generation and provision of a large-scale, consistently characterized bacterial genome dataset.

2.) Genomic characterization of GBS in elephants.

3.) Detailed profiling of GBS using a comprehensive, uniformly processed dataset.

3.1. Uniform bacterial genome data for comprehensive, comparative analyses

As previously mentioned, databases are overflowing with sequencing data, forming an ever-growing reservoir of genetic information. For the planned comparative analyses of GBS, this abundance is both an advantage and a challenge. Despite the public availability of most genome data, they are often processed with heterogeneous pipelines or variable quality control, all of which limit scientific reusability. To obtain a robust set of GBS genomes, it was therefore necessary to work from a consistently processed resource. The uniformly assembled genomes from the *661k* project offer a wealth of such data with substantial untapped potential. Given the scale of the dataset, extracting and analyzing the GBS subset would require the development of an automated and reproducible workflow. Therefore, the goals of this objective were:

- Developing of a reproducible workflow for large-scale GBS analysis.
- Identifying all GBS assemblies in the *661k* dataset.
- Characterizing all GBS assemblies included in the *661k* dataset.

However, identifying and extracting the GBS assemblies from the full dataset proved difficult. Although some species were stored within individual ENA projects, it was not fully clear whether all GBS entries had been taxonomically classified correctly. Indeed, as the authors noted, the database used for taxonomic assignment had some limitations. Because this was not sufficiently reliable for my purposes, reclassifying the entire dataset would have been necessary in any case. All assemblies would need to be processed through a consistent workflow. After identifying the GBS subset, a comparable workflow was also required to characterize it. Therefore, it was a natural step to extend the whole pipeline beyond GBS. The same pipeline planned for the GBS analysis can be applied to the remaining assemblies with minimal modifications. It can also be run at scale with my available compute resources and throughput

capacity. At the same time, broadening the characterization to the full dataset would substantially increase its value for the wider community. Many researchers, as I did, are actively searching for well-described and characterized genomes. Instead of stopping when my analysis requirements are met, systematically process the remaining assemblies makes the collection more reusable, greatly increasing accessibility and downstream utility for little extra cost. Furthermore, making the resulting processed data publicly available would additionally support the FAIR principles to ensure findability, accessibility, and reusability. Hence, further goals were:

- Processing all data contained in the *661k* dataset.
- Providing a large number of uniformly characterized bacterial genomes.
- Integrating information like assembly statistics, taxonomic classification, MLST, annotations, and metadata in one place.
- Creating a curated resource that supports robust comparative studies.
- Providing comprehensive statistics to allow insights into bacterial diversity.

3.2. Uncovering GBS in elephants

Most existing GBS research has focused on human isolates, while studies involving animal-derived isolates remain limited. Nonetheless, GBS is a multi-host pathogen. Even though it is best known for human and bovine infections, GBS has also been isolated from endangered species such as elephants [65, 118, 119]. However, a detailed genomic characterization of GBS in elephants is lacking. Despite scant reporting, pododermatitis is widespread in captive elephants [120]. Parallels between the role of GBS in human DFIs and similar elephant lesions point to a potential common mechanism [65]. An aspect that should not be underestimated is that isolates in less-studied hosts may serve as reservoirs where resistance and virulence genes can evolve undetected. A thorough understanding of bacterial species, particularly pathogenic ones, requires closer examination of exactly these kinds of ecological niches. Investigating elephant-derived GBS in detail aims to contribute to a better understanding of GBS diversity, virulence, and zoonotic potential, pointing out the importance of including isolates from diverse geographic and host backgrounds. Therefore, the goals of this objective were to:

- Characterize the genotype of elephant-derived GBS isolates.
- Assess the prevalence and etiology of GBS in elephants.
- Identify AMR genes in a less-studied ecological niche of GBS.
- Provide insights insights into GBS population evolution specific to elephants.

3.3. Epidemiological profiling of GBS using a large-scale cohort

GBS is a widely adaptable pathogen capable of infecting multiple host species. This host range is facilitated by substantial genome flexibility and the uptake of accessory genetic elements that support niche and host adaptation. Worldwide, GBS populations are structured into numerous capsular serotypes and CCs, which vary in their likelihood of causing invasive disease [121]. Although GBS has been studied extensively, the available literature is often dominated by region-specific surveys or single-outbreak investigations, yielding an incomplete and uneven view of global diversity. The comprehensive genomic and metadata resource introduced in the first objective offers a way to overcome these constraints. By providing a large set of consistently processed genomes together with contextual information, it enables consistent comparisons across studies, locations, and host settings. Therefore, the goals of this objective were to:

- Conduct a comparative analysis of a large-scale global GBS cohort.
- Characterize the microbiological features of a comprehensive set of GBS genomes.
- Analyze the trends in the current research landscape of GBS.
- Identify understudied areas in publicly available GBS datasets.

4. Thesis contributions

This thesis encompasses two peer-reviewed publications and one unrefereed preprint, which are outlined and summarized in the subsequent sections.

BakRep - a searchable large-scale web repository for bacterial genomes, characterizations and metadata
Linda Fenske, Lukas Jelonek, Alexander Goesmann, Oliver Schwengers.
Microbial Genomics (2024), DOI: 10.1099/mgen.0.001305

Evidence of a novel sublineage of *Streptococcus agalactiae* in elephants from zoo populations in Germany
Linda Fenske, Elita Jauneikaite, Maria Getino, Yu Wan, Alexander Goesmann, Tobias Eisenberg.
Microbial Genomics (2025), DOI: 0.1099/mgen.0.001489

Towards a holistic epidemiology of *Streptococcus agalactiae* using the BakRep repository
Linda Fenske, Oliver Schwengers, Alexander Goesmann.
bioRχiv (2026), DOI: 10.64898/2026.03.02.709001 (Preprint)

4.1. BakRep

BakRep:
A searchable large-scale web repository for bacterial genomes, characterizations and metadata

Linda Fenske, Lukas Jelonek, Alexander Goesmann, Oliver Schwengers (2024).

Microbial Genomics
DOI: 10.1099/mgen.0.001305

This publication describes BakRep, an innovative web-based repository developed to improve the findability and accessibility of publicly available bacterial genome data. The widespread availability of affordable sequencing technologies has dramatically increased bacterial WGS output, but this rapid growth has also brought serious challenges in data access, computational capacity, and inconsistent analytical pipelines, which together constrain scientific utility. In response to this obstacles, a previous study published a uniformly processed collection of 661 405 bacterial genome assemblies from the ENA as of November 2018. Building upon that foundation, BakRep serves as a searchable, large-scale web repository for bacterial genomes, augmented with consistent genome characterizations and associated metadata. All genomes included in the repository were consistently quality controlled, taxonomically classified, multilocus sequence typed, and annotated. Furthermore, the repository provides a flexible search engine combining taxonomic, genomic, and metadata information, as well as interactive elements to visualize genomic features. Results are presented in standard bioinformatic file formats and can be downloaded for offline analyses via an accompanying command-line tool. By providing a uniformly processed dataset, BakRep addresses key challenges in microbiological research, including data normalization, integration, and practical reusability. With its user-friendly interface and wide-ranging dataset, BakRep offers a valuable resource for comparative and clinical studies, fostering new insights and discoveries in microbial genomics. The platform can be accessed via `bakrep.computational.bio`. It has been demonstrated that specific cohorts of bacteria can easily be filtered out, downloaded, and used for further analyses. One project demonstrating the function and benefits of BakRep is described in Chapter 5.3.

4.2. GBS in elephants

Evidence of a novel sublineage of *Streptococcus agalactiae* in elephants from zoo populations in Germany

Linda Fenske, Elita Jauneikaite, Maria Getino, Yu Wan, Alexander Goesmann, Tobias Eisenberg (2025).

Microbial Genomics
DOI: 10.1099/mgen.0.001489

This publication presents the first comprehensive whole-genome characterization of GBS isolates from elephants. GBS studies have mainly focused on human disease and on bovine mastitis. However, GBS is known to affect a broader host range and can cause infections in many animal species. Although GBS has been repeatedly isolated from elephants, few studies have reported infections in wild or captive populations. To address this gap, a comparative genomic analysis of 24 elephant-derived isolates from three German zoos was conducted. Their phylogenetic placement within the broader GBS population was evaluated, and the genomes were screened for features associated with host adaptation. The elephant-derived isolates showed pronounced phylogenetic divergence from GBS originating from other host species and formed well-defined clusters that aligned with the zoo of origin and MLST profiles. Notably, capsular serotypes could not be predicted for most isolates (20/24), pointing to atypical or highly divergent capsular loci and raising questions about immune evasion and virulence mechanisms in this host context. In addition, several accessory genes were found to be associated with the elephant-derived isolates, potentially contributing to increased fitness, persistence, or pathogenicity in the elephant host and its husbandry environment. This study broadens the current view of GBS diversity and provides a genomic foundation for future research into disease ecology in exotic animals. The results demonstrate the need for further research across additional zoos, countries, and wild populations. Such efforts will be essential to better understand GBS evolution, host adaptation, and virulence, and to more accurately evaluate potential zoonotic risks.

4.3. GBS in BakRep

Towards a holistic epidemiology of *Streptococcus agalactiae* using the BakRep repository

Linda Fenske, Oliver Schwengers, Alexander Goesmann (2026).

bioRχiv (Preprint)
DOI: 10.64898/2026.03.02.709001

This study conducted a large-scale comparative analysis of all GBS genomes contained in BakRep. With its extensive genomic content and accompanying metadata, the BakRep repository serves as a valuable resource for gaining a broad overview of publicly available sequence data, enabling analyses that extend beyond separate studies. In total, 37 970 GBS genomes were analyzed and characterized by MLST, capsular serotypes, AMR genes, and putative lineage-specific genes. These genomic features were further integrated with metadata on geographic origin, host species, and disease association. The results largely recapitulated the currently reported population structure, with serotypes III, Ia, and V being most prevalent and stable serotype/clonal complex associations, alongside evidence of increasing serotype diversity. Lineages showed distinct accessory gene repertoires: III-2/CC17 was enriched for virulence and adhesion factors, while other lineages displayed greater MGE content or signatures matching niche specialization. AMR determinants were common, including very high tetracycline resistance, frequent macrolide-lincosamide-streptogramin B (MLSB) resistance genes, and the emergence of aminoglycoside resistance in a subset of genomes. However, the analysis also revealed substantial gaps in associated metadata. Missing or incomplete contextual information constrains biological interpretation and reinforces the need for well-curated, structured metadata to fully realize the value of large-scale sequencing efforts.

5. Results and discussion

"While a lot of modern computation,
may look like magic, it isn't.
It's algorithms all the way down."

(Helena Rasche – 2025)

5.1. Consistent characterization of over 2 million bacterial assemblies with BakRep

Motivated by the increasing expansion of bacterial WGS data and the consequent loss of standardization and comparability across public datasets, BakRep was designed: A searchable, large-scale web repository for bacterial genomes. Through consistent data processing, it reduces barriers to data integration, normalization, and reusability, supporting comparative and targeted analyses. Building on the uniformly processed set of 661 405 bacterial genome assemblies from the *661k* project [19], BakRep adds consistent per-genome characterizations such as standard assembly metrics, robust taxonomic classifications, MLST subtypings, genome annotations, and original metadata. All genomes in the repository are quality checked with CheckM2, which uses machine learning approaches to predict completeness and contamination levels of genomic bins [122]. Basic assembly statistics, such as the number of contigs or the L50 and N50 metrics, were calculated with assembly-scan to enhance the amount of information derived from the genomes [123]. A species-specific MLST subtyping was performed as an established standard to detect lineages below the subspecies level [124]. To become aware of potential errors in the species information of the original metadata, a robust taxonomic classification based on the Genome Taxonomy Database (GTDB) was conducted [125]. By normalizing taxonomic ranks, using relative evolutionary divergence, the GTDB removes polyphyletic groups (a grouping

of organisms by superficial similarities despite lacking a single common ancestor) and enables reproducible, genome-based species assignment [126]. For the genome annotation, the widely accepted tool Bakta was used, due to its combination of comprehensive functional annotations and efficient runtimes, which is essential given the large number of genomes analyzed [127]. Shortly after I finished processing the *661k* dataset, the *AllTheBacteria* project was introduced, tackling some of the limitations of the first approach [9]. Since the initial dataset only included data up to the end of 2018 and, as previously mentioned, public data continues to grow daily, an update was necessary. In March 2024, version 0.1 of *AllTheBacteria* was launched, integrating 1 271 428 new assemblies and covering all ENA data up to May 2023. A subsequent update in August 2024 introduced an additional 507 566 assemblies, which made the entire dataset grow to 2 440 377 assemblies [128]. I also processed these assemblies with the established workflow to further expand the dataset for the subsequent objectives.

The BakRep workflow was built with a modular architecture, allowing for straightforward integration of additional processing steps. For the initial *661k* dataset, it was run utilizing 2 128 CPU cores across 86 server nodes, each equipped with up to 128 GB of RAM. The workflow design and cloud-based infrastructure, using the compute cloud of the German Network for Bioinformatics Infrastructure (de.NBI), ensure scalability and enable rapid adaptation to expanded analyses and growing datasets. So, to handle the additional 1.7 million assemblies, provided by the *AllTheBacteria* consortium, the cloud project was expanded to 320 nodes, totalling 3 560 CPU cores and with up to 256 GB of RAM. Output files from all analysis tools were parsed in JSON format, to enhance technical accessibility. Additionally, annotation results are provided in GenBank format, along with nucleotide and AA FASTA files for all annotated coding sequences. In total, 39 TB of genomic result files were generated and stored in a Simple Storage Service (S3) (Figure 5.1).

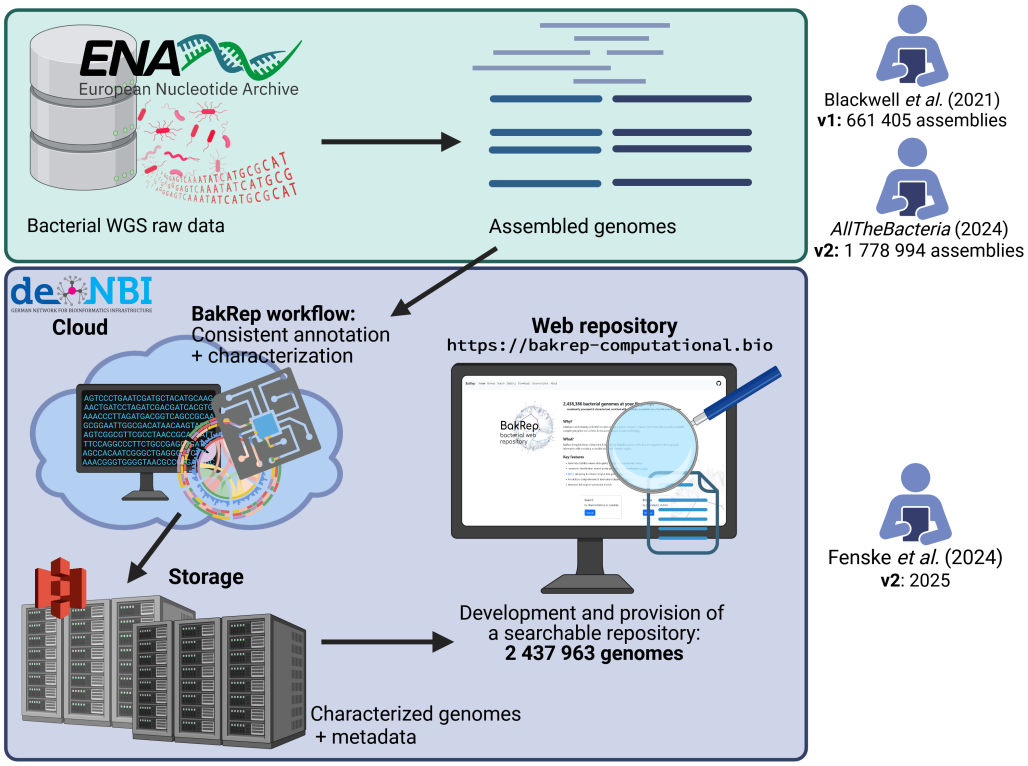


Figure 5.1.: BakRep is a searchable large-scale web repository for bacterial genomes, characterizations and metadata. Original data were retrieved from the ENA and uniformly assembled by Blackwell et al. and the *AllTheBacteria* consortium. Assemblies were consistently processed using the infrastructure of the de.NBI cloud and stored together with the original metadata in a S3. Results are available via a searchable web-interface. (*Figure created with BioRender.com*)

A substantial amount of research would be limited or even impossible without accessible data. Whilst maintaining a high degree of standardization is essential, ensuring findability and accessibility is equally important. For instance, in outbreak investigations where detecting specific AMR genes is critical, researchers must be able to rapidly identify genomes of a given species that exhibit defined characteristics such as particular MLST types or virulence factors. To advance accessibility, a comprehensive web interface was created to provide all results as user-friendly and easily accessible as possible. Furthermore, interactive reports and several filter options are available via the webpage, found at `bakrep.computational.bio`. For a general summary of the repositories' content, all available genomes can be browsed based on GC content, number of contigs, genome size, estimated completeness and contamination level, or the GTDB species. For more in-depth investigations, BakRep provides a powerful

search engine that supports flexible queries by integrating various information from the analyses as well as the descriptive metadata (Figure 5.2 and Figure 5.3). In total, the repository supports searches for 37 distinct query tags (Table 5.1). Search results are presented as a result table and can be exported in TSV format. Every single genome is further presented in a human-readable way, including a summary page summing up all results, a feature table, a genome browser, and a summary of the most relevant metadata (Figures 5.4). Datasets can be downloaded for offline analyses either directly via the website or via an accompanying CLI tool, available at `github.com/ag-computational-bio/bakrep-cli`, to enable large-scale analyses and the retrieval of extensive genome datasets.

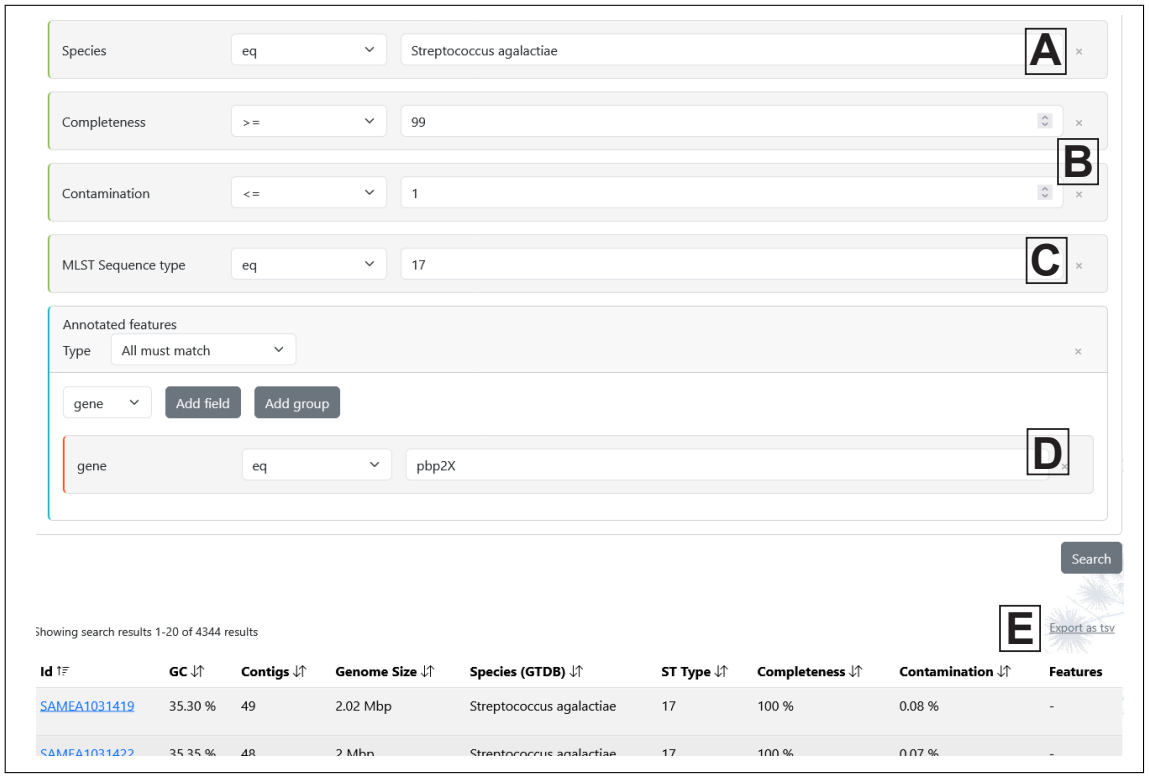


Figure 5.2.: Overview of the search engine in the BakRep web repository. Users can perform advanced queries targeting genomes with specific attributes such as **(A)** species, **(B)** genome quality, **(C)** MLST, or **(D)** annotated features. For example, to identify all *Streptococcus agalactiae* genomes with specific quality criteria, belonging to ST17 and containing the gene *pbp2X*. Search results are displayed in a compact summary table **(E)**, with detailed information accessible on individual dataset pages.

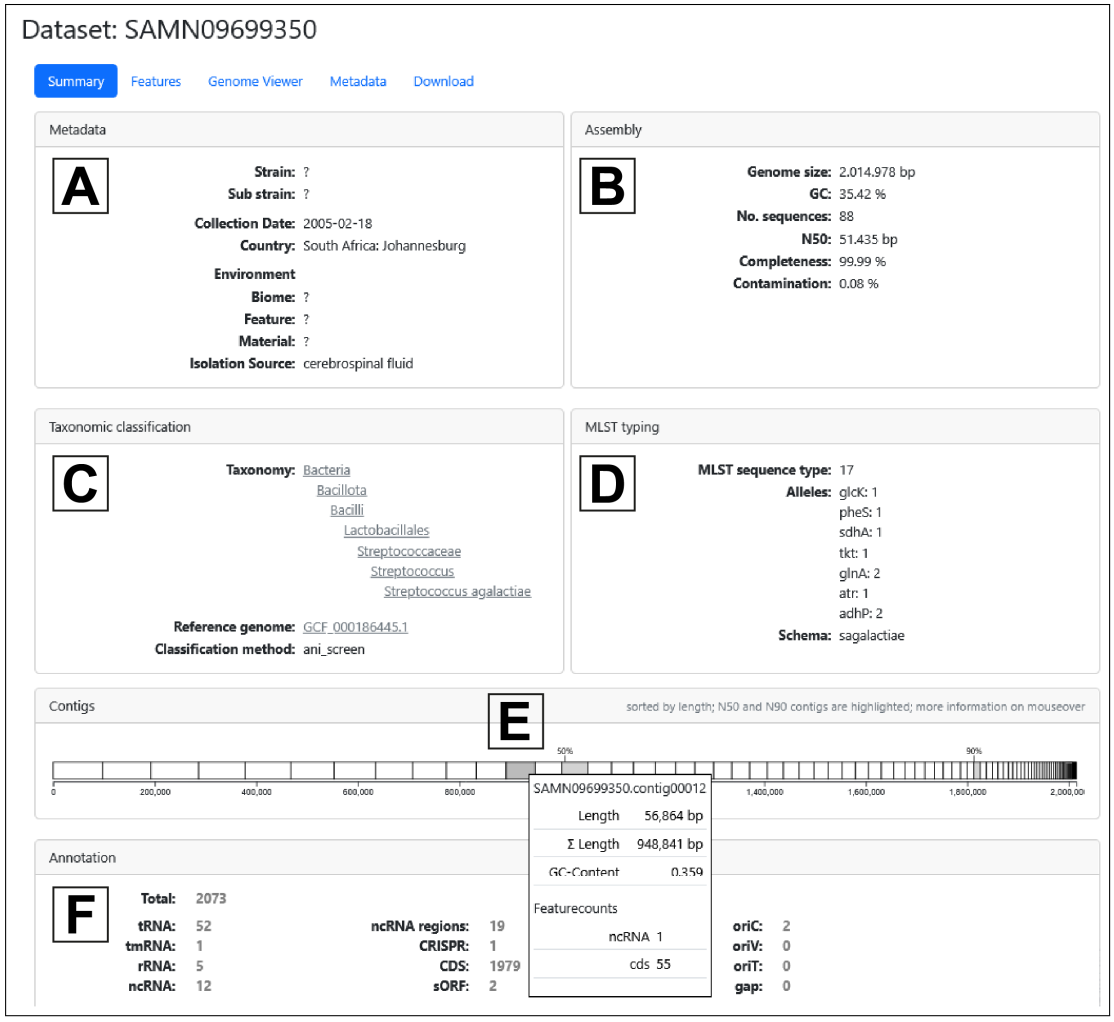


Figure 5.3.: Overview of the detailed result page of the BakRep web repository. Presented are: **(A)** A short summary of the metadata. **(B)** The key assembly statistics including estimated completeness and contamination. **(C)** A summary of the GTDB taxonomy. **(D)** A summary of the MLST subtyping. **(E)** An interactive contig bar, showing length, GC content and features of each contig. **(F)** A summary of the annotation results.

Table 5.1.: List of all search tags available in the BakRep web interface. 37 query tags derived from four different tools, along with associated metadata fields, can be used to filter the data.

Search tag	Source	Description
`Dataset ID`	metadata	BioSample accession number
`Number of contigs`	Bakta	the number of assembled contigs
`Assembly size`	Bakta	size of all assembled contigs
`GC content`	Bakta	GC content of all assembled contigs
`N ratio`	Bakta	ratio of unknown bps
`N50`	Bakta	shortest contig length that covers 50 % of the genome
`Coding ratio`	Bakta	ratio of annotated bps and genome size
`Annotated features`	Bakta	annotated features
`Completeness`	CheckM2	estimation of genome completeness
`Contamination`	CheckM2	estimation of genome contamination
`Domain`	GTDB	taxonomic domain
`Phylum`	GTDB	taxonomic phylum
`Class`	GTDB	taxonomic class
`Order`	GTDB	taxonomic order
`Genus`	GTDB	taxonomic genus
`Species`	GTDB	taxonomic species
`MLST sequence type`	mlst	species-specific ST
`Strain`	metadata	submitted strain designation
`Study accession`	metadata	ENA internal project accession
`Project name`	metadata	submitted project name
`Isolation source`	metadata	submitted isolation source
`Instrument platform`	metadata	submitted sequencing platform

Search tag	Source	Description
`Host species`	metadata	submitted host species
`Country`	metadata	submitted isolation country
`Collection date`	metadata	submitted date of collection
`Accession`	metadata	BioSample accession number
`Host sex`	metadata	submitted sex of the host species
`Host status`	metadata	submitted host health status
`Host taxid`	metadata	numerical identifier for a node in the NCBI taxonomy tree
`Instrument model`	metadata	submitted model of the sequencing platform
`Isolate`	metadata	submitted isolate identifier
`Sample alias`	metadata	ENA internal sample alias
`Secondary sample accession`	metadata	ENA internal accession number
`Secondary study accession`	metadata	ENA internal secondary study accession
`Serotype`	metadata	submitted serotype
`Serovar`	metadata	submitted serovar
`Submission accession`	metadata	ENA internal submission accession

Figure 5.4.: Interactive results and visualizations of the BakRep web repository. **(A)** Dynamic table of annotated genomic features with detailed information and cross-referenced database links. **(B)** Interactive linear genome browser with customizable tracks for genes, non-coding RNAs, CRISPR arrays, and replication-origin markers. **(C)** Summary of associated metadata. **(D)** File-download preview panel.

BakRep is not the only effort aimed at making genome data available in a more uniform manner. Conceptually related initiatives exist, but they emphasize different aspects. The beforementioned *AllTheBacteria* consortium focuses on community-led approaches. Up to now, they have broadened the scope by launching a global collaborative project to generate species-specific annotations tailored to the needs of different research communities [128]. But it is less geared toward interactive filtering of genome and metadata. GTDB offers a strictly curated, phylogenetically consistent reference taxonomy with standardized genome quality criteria and representative genomes, making it ideal for robust taxonomic frameworks, but it is not intended as a metadata-centric exploration platform [129]. Broad repositories such as NCBI RefSeq [130] or Esemble Bacteria [131] provide comprehensive access to genomes and basic metadata, but require substantial filtering, harmonizing, and infrastructure to achieve a BakRep-like level of usability. The latest version of proGenomes [132] compiles a very large, quality-controlled collection of prokaryotic genomes sourced from the NCBI. Its main strengths lie in the breadth and depth of downstream analyses, providing rich functional and ecological annotations and offering species-specific pan genome computations. But, because input comes from NCBI assemblies, the dataset reflects the heterogeneity of that repository. Stringent QC filters are applied, meaning not all available genomes are retained, and the accompanying website is primarily geared toward browsing species-level results rather than filtering by detailed metadata or typing schemes.

In conclusion, BakRep offers a unified characterization of one of the largest bacterial genome collections and serves as a high-quality, open-access resource for large-scale mibrobial genomics. By delivering a consistently processed genome collection, BakRep helps overcome obstacles in microbiological research by harmonizing analyses across studies, bringing genomic data together with contextual information, and making the resulting resource easier to explore and reuse. Its key advantage is the combination of consistent processing with user-orientated, populating-focused queying, whereas comparable resources typically emphasize either broad diversity coverage, a consistent taxonomic framework, or rich functional annotations, and offer less in terms of an interactive search interface designed for straightforward downstream reuse. This enables comparison of a vast number of bacterial genomes and large cohorts, keeping pace with future research demands.

5.1.1. Recent progress and challenges ahead

The volume of data in the INSCD has more than doubled between the release of the initial *661k* dataset and the latest *AllTheBacteria* update in August 2024, with no signs of slowing down [9]. Including those assemblies, BakRep comprises at the time of writing 2 437 963 genomes, and due to its scalable implementation and underlying cloud infrastructure, it allows for further swift adjustments to extend analysis and expand datasets. Even though it is obvious that keeping pace with this never-ending growth is an enormous challenge, continuing to offer such consistent characterization is of great value. Following the *661k* project, the *AllTheBacteria* project was designed as an open community project with the goal of engaging a broader research community, and especially to have experts who focus on particular topics or species.

BakRep has greatly improved the accessibility and, most importantly, the searchability of the data assembled so far, but the challenge rests in ensuring this remains feasible. Not only does the volume of newly generated sequence data grow daily, but tools and underlying databases are continually updated. Due to that, it is necessary to periodically rerun the BakRep workflow on the previously stored data to ensure the repository remains as up-to-date as possible. It is therefore essential to guarantee that the workflow is reproducible and reusable. Ultimately, the long-term objective would be to automate some of the updating processes. Over time, the workflow could also be enhanced by adding additional analyses or populating BakRep with existing results from the *AllTheBacteria* project.

As the data keeps growing, ensuring adequate storage capacity becomes a major challenge. A potential future improvement involves managing data storage with our in-house tool Aruna [133], which would facilitate more efficient data orchestration and flexible data handling in alignment with the FAIR principles.

During the statistical analysis of all data contained in the web repository, I observed a marked bias toward certain phyla, alongside significant gaps in the accompanying metadata. In the absence of corresponding information, the reusability of genome data is significantly constrained. Therefore, another long-term objective is to enhance

metadata availability, potentially by automatically extracting missing information from associated publications or using tools for metadata normalization like MetaMiner [134]. In future updates, it could also be considered to restrict inclusion to genomes with sufficient metadata. This would ensure not only that genuinely reusable data is released but would also improve scalability by reducing the overall data volume.

Artificial intelligence (AI) and machine-learning approaches are increasingly applied to extract insights from large data collections. Thus, BakRep would provide a suitable development platform for such methods, for instance, to address biological questions using large language models. Several AI-powered tools like these are already in use, like CellWhisperer, which enables chat-based questions over large-scale gene-expression data [135], Evo, which learns patterns and functional signals from microbial genomes [136], or platforms such as PanKB and AskMicrobe that provide chatbot assistants to query pan genome resources and microbiology knowledge [137, 138].

5.2. A novel sublineage of GBS in elephants

GBS is a well-known pathogen, most commonly linked to humans and cattle. Prior to this work, no genomic data existed for elephant-derived GBS, limiting the ability to contextualize elephant isolates within the wider GBS population. To attain insights into this less-explored ecological niche of GBS, a total of 24 isolates from elephants obtained from German zoos were analyzed, marking the first whole-genome characterization of GBS in this host species.

All isolates were collected during routine bacteriological investigations from three different zoos (A-C) between 2010 and 2023 (Table 5.2; Figure 5.5). Zoo A originally kept four female Asian elephants, but only two remained at the time of sampling, in an 830 m^2 outdoor and a 92 m^2 indoor enclosure. GBS was repeatedly isolated from both elephants, mostly from the feet, especially the nail matrix, and also from abscess material, fistulas, skin lesions, and vaginal discharge. Although treatment led to lesion healing, GBS remained detectable at chronic sites. Zoo B housed three adult females and one male African elephant in a 6 450 m^2 outdoor area and an 880 m^2 indoor space. From 2010 to 2015, multiple elephants developed infections from which GBS was consistently cultured. In 2010, "Aruba" suffered from left forefoot pododermatitis with high GBS loads. "Wankie" died in 2014 after developing multijoint arthritis. Necropsy indicated septicemia following acute arthritis, with GBS predominant in the nasal cavity, spleen, urinary bladder, and heart. "Tamo" developed skin abscesses in 2014, which worsened in 2015 with additional ulcers around the right ear, while "Zimba" developed a suppurative infection of the right temporal gland in 2015 [65]. Histories of cases in Zoo C were not available.

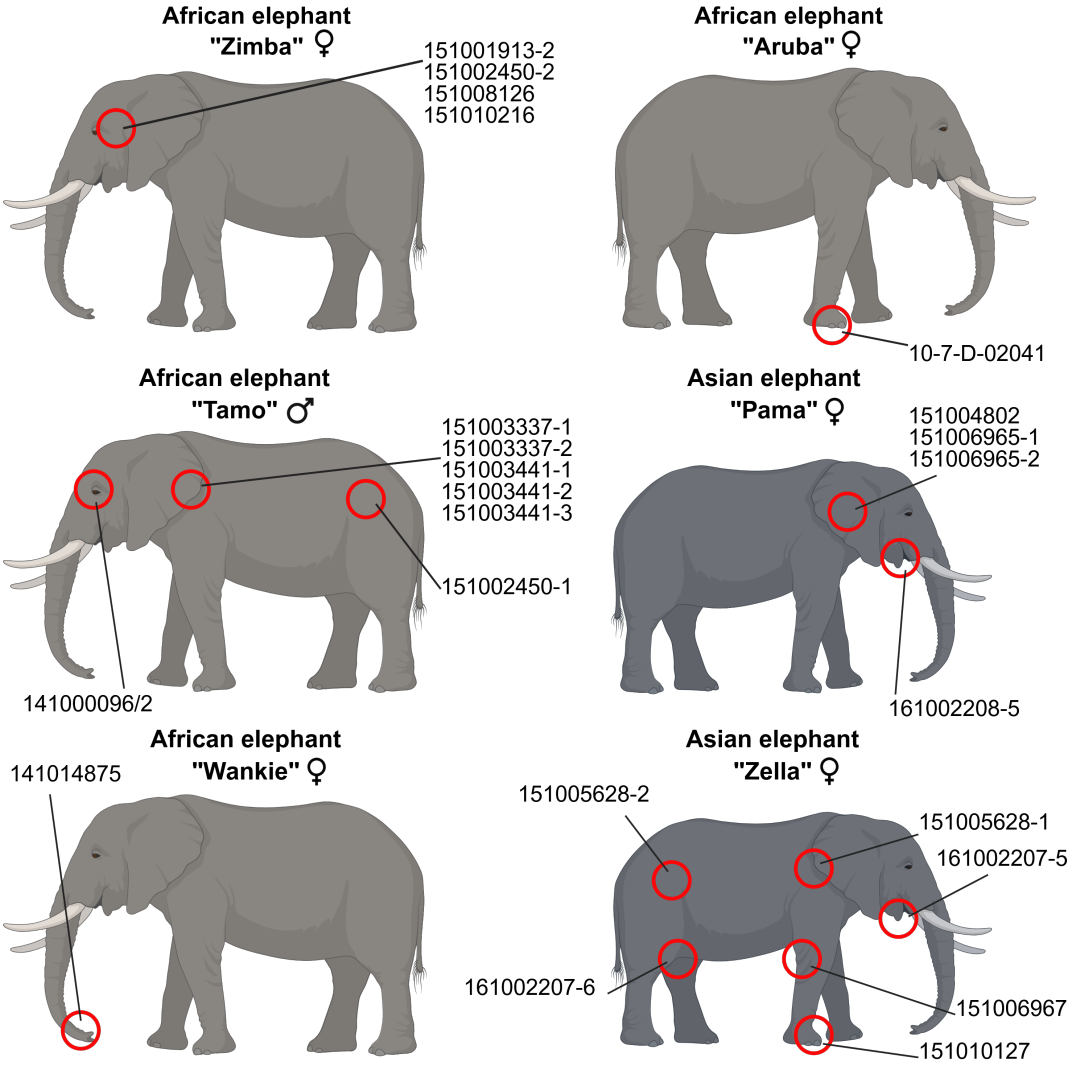


Figure 5.5.: Isolation source of the 24 elephant-derived GBS isolates. A total of six elephants were sampled, four of which were African (*Loxodonta africana*) and two Asian (*Elephas maximus*). The respective isolation site is highlighted with a red circle. (*Figure created with BioRender.com*)

Table 5.2.: Origin of the 24 elephant-derived GBS isolates. All isolates were collected from three zoos in Germany between 2014 and 2023. The latest isolate was provided by Prof. Dr. Christa Ewers from the Institute of Hygiene and Infectious Diseases of Animals, Gießen; all other isolates were provided by Prof. Dr. Tobias Eisenberg from the Hessian State Laboratory, Gießen.

Isolate	Source	Zoo	Year
161002207-5	mouth	A	2016
161002207-6	vagina	A	2016
161002208-5	mouth	A	2016
10-7-D-02041	nail (fistula)	B	2010
141000096/2	abscess (eye)	B	2014
141014875	swab (nose)	B	2014
151001913-2	temporal gland	B	2015
151002450-1	abscess (skin)	B	2015
151002450-2	temporal gland	B	2015
151003337-1	swab (ear)	B	2015
151003337-2	swab (ear)	B	2015
151003441-1	swab (ear)	B	2015
151003441-2	swab (ear)	B	2015
151003441-3	swab (ear)	B	2015
151004802	swab (ear)	A	2015
151005628-1	swab (ear)	A	2015
151005628-2	abscess	A	2015
151006965-1	swab (ear)	A	2015
151006965-2	swab (ear)	A	2015
151006967	swab (elbow)	A	2015
151008126	temporal gland	B	2015
151010127	foot	A	2015
151010216	temporal gland	B	2015
IHIT53690	placenta (abort)	C	2023

All isolates were sequenced with short-read technology (Illumina HiSeq X Ten; 150-cycle paired-end), and for two of the isolates (161002207-5, 161002207-6), long-read sequencing (MinION Mk1B) was additionally used. Variations in the core genome, which refers to the set of genes shared by all genomes of a dataset, can shape a lineage's host range. These genes primarily encode essential metabolic functions or universally shared features, making them promising targets for vaccine or drug development [139]. Genes present in one group but not in all genomes in a dataset are classified as accessory genes, which can confer niche adaptation or host-specific virulence factors [67]. The core genome of the elephant-derived isolates was compared to a subset of isolates from other host species. The observed differences supported the idea of niche-specific adaptation through the acquisition of certain accessory genes.

A gene variant of the hyaluronate lyase (*hylB*) was found in all elephant-derived isolates. This gene is thought to play an important role in bacterial invasion. However, hyaluronidase activity is not required for GBS infections in humans [140]. Another notable gene is the CAMP factor (*cfb*). This toxin creates pores in host cells and is present in nearly all virulent isolates, except for fish isolates belonging to CC552 [67, 141]. In the course of this work, *cfb* was detected in all elephant-derived isolates and in all other GBS genomes analyzed, including two CC552 fish isolates. This suggests that *cfb* is typically present, but often remains either non-functional or undetectable through biochemical CAMP reaction testing.

Certain virulence genes, specifically *scpB* (C5a peptidase) and *lmb* (laminin-binding protein), show a strong correlation with human disease [142]. While *scpB* was present in all elephant-derived isolates, *lmb* was absent from all of them. In humans, *scpB* and *lmb* are often co-located on the same transposon [67, 143, 144], but at the same time, most bovine GBS lack *lmb* [143]. The absence in elephant isolates supports the idea that these loci have a key role in human invasion and infection, but in elephants laminin-binding proteins may not contribute to pathogenesis. Nevertheless, the elephant-derived isolates possess a range of genes that jointly contribute to enhanced adaptability across various conditions.

A pan-genome-wide association study (GWAS) using scoary2 identified multiple regulatory systems involved in environmental adaptation, alongside genes linked to

metabolism and energy balance, and a range of MGEs (Table 5.3). Elephants likely impose distinct selective pressures, including a different microbiome, diet composition, and immune environment, which may require GBS to adapt its metabolism accordingly. They also carried a potentially novel variant of the fibrinogen-binding protein. All these genetic features likely support GBS in coping with diverse physiological and ecological challenges, facilitating its adaptation to a more exotic host like elephants and promoting its survival in distinct environments. GBS likely consists of populations that range from broadly adapted generalists to more host-restricted specialists, allowing it to persist throughout various ecological niches. This diversity makes GBS a useful model for studying host adaptation and how it shapes genetic exchange among populations persisting in different environments [40, 94].

Table 5.3.: Scoary2 results are reported for the most significant elephant-associated genes (specificity and sensitivity greater than 90 %). **Sensitivity:** Sensitivity of using the presence/absence of this gene as a diagnostic test to determine trait-positivity; **Specificity:** Specificity of using the absence/precence of this gene as a diagnostic test to determine trait-negativity; ***p*-value:** The *p*-value of the post-hoc permutation test for the best gene [145].

Annotation	**Sensitivity**	**Specificity**	***p*-value**
Maf family protein	100.00	100.00	0.2076
marR	100.00	100.00	0.2076
acyl-CoA thioester hydrolase	100.00	100.00	0.2076
Lin0465 protein	95.83	97.22	0.8184
histidine kinase	100.00	84.72	0.9441
prrA	100.00	84.72	0.9441
yjjG	100.00	84.72	0.9441
XRE family transcriptional regulator	79.17	94.44	0.0279
AAA family ATPase	83.33	90.28	0.2514
NADH-quinone oxidoreductase subunit J	100.00	72.22	0.0419
ATP-binding protein	100.00	70.83	0.0798
Phosphoglycerate mutase family protein	83.33	86.11	0.3573
Peptidase C51 domain-containing protein	75.00	88.89	0.0978
Phage protein	83.33	81.95	0.0040
TM2 domain-containing protein	83.33	76.39	0.0080
Fibrinogen-binding protein variant 3	79.17	77.78	0.8822
Sporulation protein Cse60	83.33	70.84	0.0060

The capsular serotypes are a major focus of research, especially in the context of vaccine development. However, this attention is primarily directed toward strains isolated from humans [146]. Defining the serotype of the elephant-derived GBS isolates proved challenging, despite different tools and methods used, indicating the predominant emphasis in research on human-specific strains. Only four isolates could be assigned to serotype Ia, which was also reported in humans, cattle, fish, seals, dogs, and camels [147]. Most of the isolates were non-typeable due to a deletion in the *cps* locus already previously described in human isolates [148]. Since the elephants

seem to represent a more specialized ecological niche, capsule expression may not be essential for colonization and persistence in this host species. However, compared to other streptococcal species, GBS shows comparatively few serotype variants. For example, *S. suis* comprises 35 known serotypes [149], whereas *S. pneumoniae* includes more than 100 [150]. This limited diversity suggests the possibility that additional, currently uncharacterized, variants may exist in GBS. Supporting this notion, a study investigating isolates from camels identified a previously unrecognized variant of serotype V, highlighting that undiscovered serotype sequences likely persist in underexplored host species [146].

Another indication supporting the existence of a distinct sublineage is that none of the elephant isolates could be assigned to a known CC, and 11 out of the 24 isolates displayed previously unknown MLST profiles (ST2304, ST2305). The remaining 13 isolates belonged to ST2019, for which the PubMLST database contains only one additional record, which was also non-typeable by capsular serotyping and sourced from an elephant from the United Kingdom (UK) (personal communication). All genotyping results are summarized in Table 5.4. Taken together with the core genome analysis, the elephant isolates likely form a host-specialist GBS lineage absent from other host species. But, while ST and CC provide broadly comparable classification schemes, they do not allow fine-scale genetic differentiation, since MLST captures only a tiny fraction of the core genome. For instance, phylogenetic analyses based solely on MLST data once incorrectly suggested that the hypervirulent clone ST17 originated from the bovine lineage CC61/67 [142, 151].

Table 5.4.: Sequencing, serotyping and MLST results of the elephant-derived GBS isolates (**nt**: non-typeable).

Isolate	Genome size [bp]	GC [%]	Serotype	ST
161002207-5	1 929 621	35.3	nt	2304
161002207-6	1 914 764	35.2	Ia	2304
161002208-5	1 914 593	35.2	Ia	2304
10-7-D-02041	1 985 838	35.4	nt	2019
141000096/2	1 986 022	35.4	nt	2019
141014875	1 986 390	35.4	nt	2019
151001913-2	1 863 397	35.3	nt	2019
151002450-1	2 026 806	35.4	nt	2019
151002450-2	1 863 022	35.3	nt	2019
151003337-1	1 986 202	35.4	nt	2019
151003337-2	1 985 853	35.4	nt	2019
151003441-1	1 986 206	35.4	nt	2019
151003441-2	1 986 069	35.4	nt	2019
151003441-3	1 986 613	35.4	nt	2019
151004802	1 930 241	35.3	nt	2304
151005628-1	1 928 632	35.3	nt	2304
151005628-2	1 930 371	35.3	nt	2304
151006965-1	1 929 610	35.3	nt	2304
151006965-2	1 914 570	35.2	Ia	2304
151006967	1 930 032	35.3	nt	2304
151008126	1 863 139	35.3	nt	2019
151010127	1 914 960	35.2	Ia	2304
151010216	1 863 537	35.3	nt	2304
IHIT53690	2 041 196	35.3	nt	2305

Although an increasing number of GBS isolates exhibit AMR, including multidrug resistance, no such genes were identified in the elephant-derived isolates, except for a single gene conferring resistance to cationic antimicrobial peptides (*mprF*). This gene may enhance the tolerance to the aminoglycosides gentamicin and daptomycin by altering the bacterial cell membrane [152]. Several mutations in the PBPs, which are critical enzymes for peptidoglycan synthesis in the cell walls of Gram-positive bacteria, are linked to reduced susceptibility to β-lactam antibiotics. Although several AA substitutions were identified in the elephant-derived isolates, none were associated with reduced penicillin susceptibility. Even so, ongoing surveillance is needed for the recurrent appearance of these substitutions, because resistance may infiltrate the population through spread and further mutations. Since penicillin remains first-line therapy in many cases, monitoring is especially critical.

5.2.1. Limitations and future research directions

All elephant-derived GBS isolates examined in this study were collected from captive elephants in Germany. To better determine whether elephants in general constitute a distinct GBS sublineage, future research needs to incorporate isolates from wildlife elephants into phylogenetic and virulence analyses to better understand the pathogenic characteristics and potential zoonotic implications of this pathogen. Distinguishing between captive and wild animals would be particularly necessary for making definitive conclusions and for excluding humans as potential transmitters. Only coordinated, parallel sampling of elephants and their zookeepers can determine the scope of cross-species transmission, similar to approaches used for dairy cattle and farmworkers [153, 154].

One other limitation of the study was the potential linking between host species and geographic location, as each zoo housed only one elephant species. This makes it impossible to differentiate whether the observed STs may be driven by elephant species or by environmental factors associated with each zoo. Additionally, the dataset spans 13 years of irregular sampling intervals, complicating the interpretation of phylogenetic relationships and strain persistence. Continued monitoring of GBS prevalence and

population structure in elephants is important to enable early detection of virulent and zoonotic strains, which can spread in zoos and even beyond. Moreover, surveying isolates from even lesser-studied hosts could reveal uncharacterized strains, and such hosts may constitute long-term zoonotic risks or reservoirs of AMR. Continued, interdisciplinary sequencing of GBS from diverse hosts is essential to elucidate host-adaptation dynamics and the organism's evolution.

5.3. A large-scale analysis of GBS using BakRep

GBS is a multi-host pathogen recognized in many species, including humans, cattle, camelids, fish [67], and as mentioned before even elephants. Most studies, however, focus on specific regions, outbreaks, STs, or serotypes, and large-scale comparative analyses are less common, which may limit our global understanding of GBS diversity. BakRep's extensive genomic data and metadata, as described in Chapter 5.1, enabled broader research trend analysis and more comprehensive population-level assessments of GBS. The BakRep dataset comprised 37 970 GBS genomes at the time of writing. The cohort is largely human-derived, and a considerable fraction of records lacks information on the host species. Animal-derived genomes represent only a small fraction of the collection, primarily from bovine and fish sources, with a minimal representation from other species such as dogs, cats, llamas, and crocodiles. Disease status information was notably sparse across the dataset, with the vast majority of genomes lacking documentation on whether they originated from diseased individuals, healthy carriers, or asymptomatic cases. The temporal distribution revealed a gradual expansion in genome availability over time. Collection efforts remained minimal from the early 1970s through the mid-2000s, followed by a marked increase beginning in 2006. The peak collection period occurred during the mid-2010s through the early 2020s, with a subsequent decline in recent years. A significant proportion of genomes lacked temporal metadata altogether.

Large-scale genomic collections derive their true utility from two components: the genetic sequences and the contextual information describing their origin. This documentation allows scientists to repurpose existing data for novel investigations and hypothesis testing [155]. Especially in comparative studies, comparing multiple genomes, with consistent and detailed metadata, is essential for drawing valid conclusions [134]. Some organizations provide guidelines for submitting metadata, but often they are recorded using unvalidated free-text entries, which introduce ambiguities, inaccuracies, and omissions that undermine data trustworthiness [156].

Since more than half of the genomes are human-derived, the dominance of serotypes III, Ia, and V is unsurprising. Yet, it remains unclear whether this distribution

is mirrored across non-human hosts worldwide. However, the frequency of rarer serotypes has increased in recent years, e.g., serotypes Ib and IX. Taken together, expanding serotype diversity and putative capsular switching complicate vaccine development, as polysaccharide-based vaccines may impose selective pressure that enables immune escape by virulent lineages. Several strong, also previously indentified, serotype/sequence type associations were found, e.g., CC17 with serotype III-2 [157, 158], ST19 with serotype III-1 [81, 157], CC23 with serotype Ia [91, 159], CC1 with serotype V [160], or ST459 with serotype IV [161]. Though disease status cannot be inferred due to missing metadata, these specificities suggest stable clonal lineages with conserved capsules, aiding targeted surveillance.

Serotype and CC distributions also varied across continents. North America was dominated by serotypes Ia, II, and III, with CC1 and CC23 being the most common. In Africa and Europe serotype III occurred more frequently (especially III-1/III-2), with CC17 predominant. While serotype III is often reported as overrepresented in Africa, I found no corresponding subtype data [75, 162–164]. Asia was marked by a high prevalence of serotype III-4 and CC283 alongside CC1. Notably, III-4 and CC283 were almost exclusive to Asia. This combination seems to be linked to fish disease in Asia [63, 165].

Across all genomes, more than 80 % carried at least one AMR gene (132 determinants and 24 drug classes), indicating a considerable genomic reservoir for resistance that warrants ongoing surveillance and phenotypic confirmation. Tetracycline resistance genes dominated, consistent with historical tetracycline use driving clonal expansion and human adaptation. MLSB resistance determinants were the second most common, chiefly *erm(A)*, *erm(B)*, and the *mef(A)-mrs(D)* efflux pair. These trends call into question clindamycin and erythromycin as dependable alternatives for penicillin-allergic patients and justify shifting to cephalosporins or vancomycin [166]. Nevertheless, β-lactam resistance genes were rare, but since their main mechanism of resistance arises from AA substitutions in the PBPs, surveillance is still essential. Tetracycline and erythromycin resistance determinants are often co-localized on MGEs [25]. Nearly 30 % of genomes harbored both macrolide and tetracycline resistance genes, suggesting that MGEs enable the swift dissemination of multidrug resistance. Resistance patterns were additionally serotype- and lineage-linked, and differed by

geographic region, indicating the value of targeted surveillance strategies. AMR genes were rarely detected in genomes collected before the 1980s, with only occasional tetracycline and macrolide resistance determinants identified in the late 1980s. Over the last 20 years, AMR gene frequency and diversity increased overall. *Tet(M)* remained the most common gene, *tet(O)* rose markedly after 2008, and *erm(A)/erm(B)* became increasingly frequent from 2000 onward. By contrast, aminoglycoside-resistance genes displayed no consistent upward trend, appearing instead in occasional, individual peaks. This expansion of AMR highlights, moreover, the need for continued surveillance, yet the presence of resistance genes does not automatically translate into phenotypic resistance. Genomic results should be interpreted cautiously and, where possible, validated with phenotypic testing. Standardized, detailed metadata are also essential to accurately assess trends and compare findings across regions.

Gene-content analysis revealed a large pan genome (274 610 genes) with a comparably small core genome (1 398 genes). However, most genes were rare, with over 99 % occurring in fewer than 15 % of genomes. Using a GWAS, genes associated with the five most common serotype/CC associations (Ia/23, II/22, III-2/17, IV/459, V/1) were identified. Ia/23 was enriched for MGEs, phage elements, as well as metal homeostasis and resistance genes. II/22 showed enrichment for metal uptake, whereas III-2/17 accumulated virulence and adhesion determinants, notably the SecA2/SecY2 secretion system, linked to ST17 hypervirulence [167]. IV/459 and V/1 were dominated by phage-associated genes. Overall, these lineages carry characteristic accessory gene repertoires, spanning from high genomic plasticity to niche-adaptive specialization, which likely underpins lineage-specific disease patterns.

5.3.1. Limitations and future research directions

When investigating population-wide genomic patterns, it is essential to sample a diverse set of genomes. Not only in terms of genetic features such as MLST and serotype, but also regarding geographic origin, host species, or year of isolation. However, it is equally important to avoid an overrepresentation of lineages that are disproportionately studied. Current GBS sequencing is heavily skewed toward invasive isolates from humans, especially those causing neonatal disease. Consequently, public repositories are dominated by these genomes, even though they may not represent the organism's actual diversity and prevalence in natural settings. Moreover, large-scale genomic studies do not always apply stringent sampling strategies and may select isolates largely based on availability rather than defined criteria. Chiara Crestani also raised the above concern in her work about the host-specificity across GBS lineages and noted that public GBS genome collections are affected by strong sampling biases [67].

I observed the same bias in BakRep, where certain lineages and isolation contexts are markedly overrepresented. To maximize diversity, Chiara Crestani curated her dataset by removing duplicates and selecting unique combinations. However, she also acknowledged that this strategy can distort frequencies, and so rare isolation events became overly represented. In contrast, I intentionally retained potential duplicates in order to preserve and make visible the underlying sampling bias and to highlight underrepresented sources. For future analyses, it may be useful to cluster genomes from the same lineages into groups to reduce such artifacts. Rather than constructing one single pan genome across all genomes, it could be better to build pan genomes one per duplicate cluster and then compare these genes in more detail. Nevertheless, broad comparative-genomics analyses, especially those examining lineage-specific genetic differences and factors that may promote niche adaptation remain crucial for elucidating multi-host pathogens such as GBS. In subsequent projects, profiling MGEs, such as transposons and plasmids, could help to elucidate additional resistance mechanisms. Moreover, a detailed examination of the PBPs and possible AA variants would help clarify pathways to penicillin resistance.

Taking advantage of such an expansive global dataset offers rich opportunities, especially for pathogens of clinical relevance. However, the greatest prospects may lie in lesser-explored organisms. BakRep currently contains 14 071 distinct species, enabling a wide range of further comparative projects. In particular, rarely studied species can open up valuable avenues for discovery, as they may more likely capture unexplored evolutionary and ecological diversity. Furthermore, comparing repeatedly occurring features, like shared determinants of adaptation, virulence, or AMR genes, across species boundaries may help address new biological questions and can also highlight areas where targeted sampling and follow-up research are still needed.

6. Conclusion

> "For the microbial ecologist, what can be cultured is the basis of his conception of what exists. This is exactly like learning about animals from visiting zoos."
>
> ***(Carl R. Woese – 1994)***

Prokaryotes constitute an ancient, highly diverse domain of life and are among the most abundant organisms on earth. Attempts to characterize and classify this diversity produce extensive volumes of sequencing data, with thousands of sequences added to public repositories every day. Sequencing prokaryotic genomes has transformed our understanding of the diverse roles of microorganisms, and sequence-based analyses are indispensable for detecting and comparing genes of interest in various genomes. However, identifying appropriate comparative datasets is often challenging when data have not been processed in a consistent, uniform way.

This thesis provides a repository of consistently and comprehensively characterized bacterial genomes usable for the targeted analysis of specific cohorts. In contrast to the results collected by the *AllTheBacteria* consortium so far, BakRep enables substantially easier retrieval of specific datasets. Researchers can query the repository through a flexible search engine that integrates genomic information and associated metadata. This thesis also demonstrates the utility of this resource by analyzing all GBS genomes in BakRep, which enabled inferences to be made of serotype and CC distribution, as well as AMR, and lineage-specific genes and helped identify the study bias in current GBS research.

A majority of WGS data is strongly biased toward a few phyla and especially specific model species. It is also worth noting that WGS studies are largely limited to organisms that can be grown under laboratory conditions. Consequently, most microbial species, which have not been successfully cultured to date, remain underrepresented despite the fact that these organisms are likewise important to Earth's ecology [3]. The microbiologist Jeffrey Gralnick once remarked that: "*E. coli* is a great model organism - for *E. coli*." [22]. This underscores that repeatedly analyzing the same strains does not necessarily improve our understanding of the full breadth of bacterial diversity. In GBS, accessory genes are known to be exchanged among different streptococcal species and even with other Gram-positive bacteria [67]. Such shared adaptations are called ecotypes: bacterial groups that occupy the same niche and exploit similar resources, using common genetic traits to outcompete less-adapted strains [168]. However, investigating a single species within that niche does not inevitably broaden our understanding of the other species that inhabit it. Although the elephant-derived isolates carry certain host-specific genes, it cannot be excluded that they could still cause invasive infections in humans or other host species. Nevertheless, my results point to a potentially distinct sublineage, potentially shaped by adaptations to their host and environment. Despite the study's limited geographic scope, comparative genomics revealed that elephant-derived GBS differs in several ways from strains found in other hosts. This highlights the need for further research on GBS evolution and zoonotic potential, especially in rarely studied ecological niches.

It is crucial to further explore understudied niches, e.g., by characterizing species that are rarely detected in metagenomic investigations. A 2025 study on literature bias in bacteriology notes: "Our suggestion is to focus on the production of knowledge, not the collection of data." [22]. Estimates suggest that Earth may harbor around one trillion microbial species [169], and we only know a fraction of them. Under a niche-based view, bacterial specification occurs when a lineage adapts to a new resource or habitat and becomes ecologically stable against displacement by its ancestral population. Such niche shifts can either be driven by mutations in existing genes or by acquiring genes from other species [168]. Rarely studied microbes can have outsized effects on biogeochemical cycles, ecosystems, or host-associated functions, potentially acting as overlooked keystone species in terrestrial and aquatic environments [170]. Incorporating such microbes and niches into future studies will

deepen our understanding of how microbial communities function and help clarify how ecosystems develop. Despite the pronounced species bias, BakRep enables comparable genome characterizations across very large and diverse genome collections. With several thousand species included, it supports broad comparative studies, even for taxa with limited representation, and can therefore support identifying shared traits of adaptation, virulence, or AMR across species and help to pinpoint targets for focused follow-up sampling.

7. Scientific contributions

7.1. First authorships in scientific publications

7.1.1. A dominant clonal lineage of *Streptococcus uberis* in cattle in Germany

Linda Fenske[1,2], Irene Noll[3], Jochen Blom[2], Christa Ewers[4], Torsten Semmler[5], Ahmad Fawzy[1,6], Tobias Eisenberg[1]

[1]Hessian State Laboratory, Department of Veterinary Medicine, Giessen, Germany

[2]Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

[3]Regional Council of Gießen, Wetzlar, Germany

[4]Institute of Hygiene and Infectious Diseases of Animals, Justus Liebig University, Giessen, Germany

[5]NG1 Microbial Genomics, Robert Koch Institute, Berlin, Germany

[6]Department of Medicine and Infectious Diseases, Faculty of Veterinary Medicine, Cairo University, Giza, Egypt

Antonie van Leeuwenhoek (2022), DOI: 10.1007/s10482-022-01740-w

The following contributions are attributed to the thesis author:

Involved in planning and coordination of the study. Conducted the data analyses and interpreted the data. Drafted the manuscript and finalized it with input from the co-authors. Corresponded with the journal.

This publication is not part of the main set of publications for this cumulative thesis, but is included for completeness/background.

Antonie van Leeuwenhoek (2022) 115:857–870
https://doi.org/10.1007/s10482-022-01740-w

ORIGINAL PAPER

A dominant clonal lineage of *Streptococcus uberis* in cattle in Germany

Linda Fenske · Irene Noll · Jochen Blom · Christa Ewers · Torsten Semmler · Ahmad Fawzy · Tobias Eisenberg

Received: 12 December 2021 / Accepted: 9 April 2022 / Published online: 30 April 2022

Abstract Bovine mastitis causes enormous economic losses in the dairy industry with *Streptococcus uberis* as one of the most common bacterial pathogens causing clinical and subclinical variations. In most cases mastitis can be cured by intramammary administration of antimicrobial agents. However, the severity of the clinical manifestations can vary greatly from mild to severe symtoms. In this study, a comparative genomic analysis of 24 *S. uberis* isolates from three dairy farms in Germany, affected by different courses of infection was conducted. While there were sporadic mild infections in farm A and B, a large number of infections were observed within a very short period of time in farm C. The comparison of virulence genes, antimicrobial resistance genes and prophage regions revealed no features that might be responsible for this severe course. However, almost all isolates from farm C showed the same, novel MLST profile (ST1373), thus a clonal outbreak cannot be excluded, whereby the actual reason for the particular virulence remains unknown. This study demonstrates the importance of extensive metagenomic studies, including the host genomes and the environment, to gain further evidence on the pathogenicity of *S. uberis*.

Keywords Bovine mastitis · Comparative genomics · Multilocus sequence typing · Resistance · *Streptococcus uberis* · Prophage regions · Virulence

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10482-022-01740-w.

L. Fenske (✉) · A. Fawzy · T. Eisenberg
Hessian State Laboratory, Department of Veterinary Medicine, Giessen, Germany
e-mail: linda.fenske@computational.bio.uni-giessen.de

L. Fenske · J. Blom
Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

I. Noll
Regional Council of Gießen, Wetzlar, Germany

C. Ewers · T. Eisenberg
Institute of Hygiene and Infectious Diseases of Animals, Justus Liebig University, Giessen, Germany

T. Semmler
NG1 Microbial Genomics, Robert Koch Institute (RKI), Berlin, Germany

A. Fawzy
Department of Medicine and Infectious Diseases, Faculty of Veterinary Medicine, Cairo University, Giza, Egypt

Introduction

Bovine mastitis causes enormous economic losses in the dairy industry (Hogeveen et al. 2011). It is estimated that the loss amounts to approximately 124 Euros per cow annually, resulting in annual losses of 500 million Euros in Germany and up to 125 billion Euros worldwide (Kabelitz et al. 2021). Bovine

mastitis is defined as an inflammation of the udder, which is characterised by an increase in somatic cell count (SCC) and typically by the presence of udder pathogens. In case of subclinical mastitis, there are no clinical signs beside SCC elevations, whereby clinical mastitis is characterised by a visibly altered milk or udder, which can reach various degrees of severity (GVA 2012). If no udder pathogens are detected and there are subclinical or clinical findings, it is a matter of non-specific mastitis. Mastitis can be caused by a wide variety of pathogens, e.g. bacteria, yeasts or algae (Bradley 2002) and some viral pathogens aswell (Wellenberg et al. 2002). *Streptococcus uberis*, a Gram-positive, catalase-negative member of the family *Streptococcaceae* with a genome size of about 1.8–2.3 Mbp is one of the most important environmentally associated pathogens responsible for bovine udder infections (Whiley and Hardie 2015). Mastitis caused by *S. uberis* in particular can be both subclinical and clinical, the latter often causing mild to severe visible signs of inflammation (Zadoks et al. 2003). It is not yet certain whether these different courses of infection are solely based on the genetic diversity of involved bacterial strains or if other (e.g. environmental, immune) factors also play a role (Günther et al. 2016). In general, *S. uberis*, which is also known to be shed with bovine faeces, normally does not spread from udder to udder (non-contagious). Furthermore, severe infections with this pathogen by a lymphogenic or haematogenic route to extra-mammary tissues are rarely observed (Thomas et al. 1994). Occasionally, *S. uberis* seems to induce a contagious type of mastitis as suggested by recent clinical case series (Davies et al. 2016; Wente et al. 2019). In this regard, this bacterium is able to adapt to different environmental conditions, body sites and also stress situations with different gene expressions (Ward et al. 2009). Likewise, antimicrobial and in particular β-lactam resistant strains of *S. uberis* have been observed more frequently in the recent past, which might have been induced by certain substitutions in the sequences of the penicillin binding proteins (pbp) leading to increased resistance to oxacillin (McDougall et al. 2020). Although several strains have been typed using multilocus sequence typing (MLST), only a few fully sequenced genomes exists to this date. Nevertheless, the number of sequenced draft genomes continues to increase and studies in this field are becoming more and more relevant (Hossain et al. 2015; Vezina et al. 2021). Currently, 69 genomes can be found in the NCBI database. But, despite extensive research, prophylactic vaccination against *S. uberis* is still under debate, because little is known about the interaction between *S. uberis* and its host and generally knowledge about the pathogen's genomic traits is still scarce (Hossain et al. 2015).

In this study, we conducted a comparative genomic analysis using 24 *S. uberis* isolates, obtained from 3 dairy farms in Germany, that were experiencing different courses of infection. The results could help to extend the insight into the molecular epidemiology of this important pathogen.

Material and methods

Cattle herds under study

For this study aseptically drawn milk samples were sent to the laboratory for cyto-bacteriological investigation (CBI). A total of 24 *S. uberis* strains collected from 3 different farms in Germany between 2016 and 2019 were included in the study. All three herds on the farms are located in the same surrounding and were characterised by the same breed, age ranges of sick animals, usage of cows, hygienic conditions, used sanitisers and milking practice (Suppl. Tab 1). *S. uberis* infections represented the predominant mastitis pathogen, however, some mixed infections with coagulase-negative staphylococci were detected, but not considered for sampling. No *Mycoplasmopsis bovis* infections were detected by PCR in pooled samples (data not shown). The spread of *S. uberis* infections differed between the farms. While there were sporadic mild infections in farm A and B (group A and B), a large number of infections were observed within a period of eight weeks in farm C (group C). The severity of the infection in farm C was also much more pronounced compared to farm A and farm B and included mostly severe inflammations of the udder and fever. Various antimicrobial treatment schemes (data not shown) ended up in no or only short-term clinical improvement, although the streptococci detected in the antimicrobial susceptibility test were reported to be sensitive to all the substances used (s. below). Although *S. uberis* is usually an environmental pathogen, in the present case it resembled a contagious course of infection that is

spreading from cow to cow. About 30 lactating cows had to be slaughtered, because of the severity of mastitis despite antibiotic treatment. At the time of investigation on farm C, 106 animals were lactating. Milk samples from 69 cows hav been taken and were submitted to the laboratory for CBI. In 32 of the animals examined, *S. uberis* in connection with an increased somatic cell count could be detected in at least one quarter of the udder. Fourteen of these strains were selected for further investigations in this study.

Microbiological culture and identification of mastitis pathogens including *S. uberis*

Milk samples were directly streaked on inhouse prepared Columbia agar with 5% cattle blood and esculin (ingredients provided by Oxoid, Wesel, Germany) and cultivated using aerobic atmosphere conditions for 48 h at 37 °C. Yeast growth was investigated using a Sabouraud glucose agar with gentamicin and chloramphenicol (Oxoid) at 30 °C in samples containing an SCC of more than 3×10^6 somatic cells/mL. Isolates were further evaluated using Gram's staining and matrix assisted laser desorption/ionisation—time of flight mass spectrometry (MALDI-TOF MS; microflex LT Mass Spectrometer, MALDI Biotyper™; Bruker Daltonik, Bremen, Germany) using the direct smear method in sample preparation and the commercial MALDI-Biotyper database (MBT 8468; Bruker Daltonik).

Antimicrobial susceptibility testing

Antimicrobial susceptibility testing (AST) was carried out using broth microdilution testing. Briefly, a commercially available panel layout for mastitis treatment (Micronaut/Bruker according to guidelines of the research group antimicrobial resistance of the German Veterinary Society DVG) was used. In this layout, 11 different antimicrobials were employed [(ranges given in µg mL^{-1}); penicillin (0.063–4), oxacillin (0.063–2), amoxicillin/clavulanic acid (0.031/0.063–8/16), ampicillin (0.125–8), cefazolin (0.5–16), cefoperazon (0.25–4), cefquinome (1–16), kanamycin/cephalexin (0.031–2), marbofloxacin (0.016–2), erythromycin (0.125–4) and pirlimycine (0.25/4.75–2/38)]. Resulting MIC values were interpreted as sensitive, resistant and intermediate resistant based on clinical breakpoints according to CLSI VET01/VET01S (5th ed.) for broth microdilution testing. *Escherichia coli* ATCC 25,922, *Pseudomonas aeruginosa* ATCC 27,853, *Enterococcus faecalis* ATCC 29,212 and *Staphylococcus aureus* ATCC 29,213 served as quality control strains.

Sequencing and bioinformatic processing

Next-generation-sequencing (NGS) was carried out with 24 representative *S. uberis* isolates from dairy farms in Germany. From randomly chosen control farms, A and B, 10 isolates (each $n = 5$) were selected, whereas the remaining *S. uberis* isolates were included from farm C (group C, $n = 14$). Genomic DNA was extracted from a 72 h bacterial culture and sequencing was employed using Illumina MiSeq and Illumina NextSeq 150 bp paired-end sequencing with an obtained coverage of > 70x (Table 1). Raw reads were used for de novo assembly into contiguous sequences and subsequently into scaffolds using SPAdes (Bankevich et al. 2012). Assemblies were annotated using Bakta (Schwengers et al. 2021). MLST profiles were checked with PubMLST (pubmed.org/suberis). For clonal complex assignment of isolates and classification of sequence types in the overall population, goeBURST was used (Feil et al. 2004). For determination of the pan/core genome, as well as for orthologous genes, EDGAR 3.0 was used (Dieckmann et al. 2021). The PHASTER webtool was used to identify prophage sequences within the draft genomes (Arndt et al. 2016). ABRicate, with an individual gene set of *S. uberis* putative pathogenicity factors described in the literature, was used for virulence and resistance gene analysis (https://github.com/tseemann/abricate). As reference for all analyses the genome sequences of the strains 0140J (accession number AM946015)—which has been isolated from milk obtained from clinical case of bovine mastitis in 1972 and sequenced in 2009 as a typical virulent UK isolate pathogenic to lactating as well as non-lactating bovine mammary glands (Ward et al. 2009)—and EF20 (accession number JANW01) as an nonvirulent strain according to Hossain et al. 2015, were used. To gain insight into possible β-lactam antibiotic resistance, the genome of strain 0140J was first searched for penicillin-binding proteins. EDGAR was then used to search for orthologues of these genes in all the draft genomes, and a subsequent BLAST search

Table 1 Assembly statistics of 24 *Streptococcus uberis* isolates

Isolate	Sequencing platform	No. of contigs	N50	Average coverage range
Su-1	Illumina NextSeq 500	28	159,211	275–2463
Su-2	Illumina MiSeq	16	175,018	184–1665
Su-3	Illumina MiSeq	15	457,442	134–1210
Su-4	Illumina NextSeq 500	31	197,492	83–3848
Su-5	Illumina MiSeq	14	206,517	130–1164
Su-6	Illumina MiSeq	24	229,596	100–932
Su-7	Illumina NextSeq 500	23	313,412	138–3145
Su-8	Illumina MiSeq	37	123,181	202–3331
Su-9	Illumina NextSeq 500	21	255,232	129–3911
Su-10	Illumina NextSeq 500	38	206,273	95–4030
Su-11	Illumina MiSeq	13	436,510	160–1445
Su-12	Illumina MiSeq	12	204,844	166–1565
Su-13	Illumina MiSeq	15	439,515	74–1330
Su-14	Illumina MiSeq	12	1,031,011	92–1363
Su-15	Illumina MiSeq	13	1,030,881	265–3465
Su-16	Illumina MiSeq	13	1,030,845	458–3633
Su-17	Illumina MiSeq	12	439,836	406–4468
Su-18	Illumina MiSeq	49	79,688	56–3424
Su-19	Illumina MiSeq	12	439,511	206–1692
Su-20	Illumina MiSeq	18	320,790	114–2080
Su-21	Illumina MiSeq	21	187,067	83–1150
Su-22	Illumina MiSeq	12	439,515	121–2124
Su-23	Illumina MiSeq	13	436,512	474–4427
Su-24	Illumina MiSeq	16	439,517	352–5571

was done to locate substitutions within the amino acid sequences of these genes (Altschul et al. 1990).

Results

Microbiological culture results

Herds under study showed a similar distribution of mastitis pathogens with *S. uberis* found in several quarters. It was the only relevant microorganism that was mostly found. In farm A, B and C 6/48 (12.5 %), 17/140 (12.1 %) and 48/252 (19.0 %) quarters were positive for this microorganism, respectively. All *S. uberis* isolates were confirmed by MALDI-TOF MS and a subset of 24 isolates was chosen for WGS analysis based on monobacterial infections and a high semi-quantitative count (> 200 colonies per quarter) of *S. uberis* together with high SCC above 1×10^6 cells/mL. Samples with mixed infections or possible contamination were not taken into account.

Antimicrobial susceptibility testing

Based on AST as conducted by broth microdilution testing, all strains from this study were susceptible to all antimicrobials tested (Table 2).

Sequencing

The combined lengths of the assembled contigs ranging from 1,805,858 to 1,983,343 bp and the GC-content was in the range of 36.3–36.5 %, which is in line with the literature values (Table 3) (Whiley and Hardie 2015). In comparison, the GC content of all *S. uberis* genomes currently available in the NCBI database ranges between 36.1 and 36.8 % (data not shown). Isolates from farm A had a mean genome

Table 2 Representative *Streptococcus uberis* AST phenotype

Parameter	Interpretation result	MIC values
Amoxicillin/clavulanic acid	S	< = 4/2
Ampicillin	S	< = 4
Cefquinome	S	< = 1
Cefazolin	S	< = 4
Cefoperazon	S	< = 2
Erythromycin	S	< = 0.125
Kanamycin/cephalexin	S	< = 4/0.4
Marbofloxacin	S	= 0.5
Oxacillin	S	< = 1
Penicilling	S	< = 0.125
Pirlimycin	S	< = 1

size of 1,835,641 bp (SB ± 31,282.29), whereas farm B isolates had a mean size of 1,949,274 bp (SD ± 25,325.31) and farm C isolates displayed a mean genome size of 1,886,202 bp (SD ± 28,483.27). With respect to genome size here was a high statistically significant difference ($p < 0.01$) between groups A and B and between groups B and C, whereas groups A and C differed significantly ($p < 0.05$). All genomes of *S. uberis* currently available in the NCBI database had a mean genome size of 1,917,489 bp (SD ± 53,593.74), which is high significant above the genomes examined in our study with a combined mean genome size of 1,888,809 (SD ± 46,408.21; $p < 0.01$; Fig. 1).

Pan and core genome analysis

Examination of the gene content of all 24 isolates (referred to as Su-01—Su-24) resulted in a pan genome of 2508 genes, with a core genome of 1611 genes. With the addition of strain 0140J, the pan genome increased to 2551 genes, and with strain EF20, the pan genome increased to 2642 genes. To address the presumption of a clonal outbreak and to determine the exact relationships between the isolates, a phylogenetic tree based on the core genome was constructed using EDGAR 3.0 (Fig. 2). The topology of this tree was initially divided into three monophyletic groups. One group contained the isolates Su-06 and Su-08 from group B, the second clade only included the strain EF20, and the last branch contained all remaining isolates and the strain 0140J. Group A also formed a separate monophyletic group within the latter group. The isolates Su-11 and Su-23 from farm C form a group together with strain 0140J. All other isolates from group C clustered in the same clade. However, Su-07, Su-09 and Su-10 from group B were intermixed in this group. When looking at the development of the core and pan genome over the number of genomes, within group C, there is neither a significant increase in the pan genome nor a decrease in the core genome (Fig. 3). The increasing of the pan genome as well as the decreasing of the core genome showed when isolates from groups A and B are included.

To get a first impression of the intra-species similarity between isolates, an average nucleotide identity (ANI) matrix was created (Fig. 4). The ANI between all isolates was within 98.37 % and 100 %. Isolates Su-01 and Su-02 gave a 100 % match to each other, as did isolates Su-11 and Su-23. In addition, a group of isolates that also completely matched in their ANI could be identified within group C (Su-14 to Su-17, Su-19 to Su-22, and Su-24).

Multilocus sequence typing

Four of the isolates, as well as the reference isolates, could be assigned to previously known sequence types (ST) namely ST931 (n = 2), ST964 (n = 2), ST55 (n = 1) and ST1 (n = 1). All other isolates, contained allele variants that have not yet been described. In detail, seventeen of the isolates contained new allele variations of the *tdk* gene locus, which were registered as allele numbers 122, 123 and 124 in the *S. uberis* MLST database (pubmlst.org/suberis). With registration of the new locus variants, all remaining isolates could be assigned to new sequence types, namely ST1373 (n = 12), ST1374 (n = 3), ST1375 (n = 2), ST1377 (n = 1), ST1378 (n = 1) and ST1379 (= 1). Overall, we detected four different STs among isolates from farm A (ST1375, ST1377, ST1378, ST1379), two STs in farm B (ST931 and ST1374), and two STs in farm C (ST964 and ST1373) (Table 4). Thus, 12 of the 14 isolates (85.7%) from group C showed the same MLST type. The novel STs ST1373 and ST1374 belonged to clonal complex CC5, with possible founder ST1065. All other STs were double locus variants (DLV).

Table 3 Sequencing statistics of 24 *Streptococcus uberis* isolates plus the strains 0140 J and EF20

Isolate	sequence size (bp)	Number of contigs (> 500 bp)	Longest contig (bp)	Shortest contig (bp)	GC %
Group A					
Su-01	1,821,188	21	515,871	610	36.5
Su-02	1,818,057	12	1,002,414	800	36.5
Su-03	1,883,847	8	1,008,567	933	36.4
Su-04	1,849,257	21	318,687	933	36.5
Su-05	1,805,858	8	1,000,761	800	36.4
Group B					
Su-06	1,911,851	11	524,326	5 579	36.3
Su-07	1,950,395	14	659,959	891	36.4
Su-08	1,949,859	18	653,595	553	36.3
Su-09	1,950,920	9	1,070,520	933	36.4
Su-10	1,983,343	17	429,199	606	36.4
Group C					
Su-11	1,883,972	8	1,061,615	800	36.4
Su-12	1,875,458	8	1,032,662	548	36.3
Su-13	1,875,842	8	1,032,662	933	36.3
Su-14	1,875,689	9	1,031,954	933	36.3
Su-15	1,876,200	8	1,032,186	933	36.3
Su-16	1,876,102	11	1,031,729	508	36.3
Su-17	1,875,804	8	1,032,437	933	36.3
Su-18	1,982,968	50	922,189	508	36.4
Su-19	1,876,339	9	1,032,186	933	36.3
Su-20	1,896,764	15	1,035,846	567	36.4
Su-21	1,875,591	9	690,884	933	36.3
Su-22	1,875,809	8	1,032,566	933	36.3
Su-23	1,884,146	9	1,061,387	535	36.4
Su-24	1,876,146	8	1,032,828	933	36.3
Reference strains					
0140J	1,852,352	1	–	–	36.6
EF20	1,932,039	17	1,013,731	366	36.3

Virulence factors

In total, 66 putative virulence genes were found. Forty-eight of these occurred in all isolates with 100 % gene coverage. All genes examined were also detected in strain 0140J and strain EF20 likewise contained all but four of these genes. The genes for zinc binding protein (*acdA*), lactoferin binding protein (locus tag in 0140J genome: *SUB0145*; gene name: *lbp*), conserved hypothetical protein (*SUB0159*), putative surface-anchored subtulase family protein (*SUB0826*), collagen-like surface-anchored protein (*SUB1095*; *sclB*), S-ribosylhomocysteinase (*luxS*) and the putative bacteriocins locus tags/genes *SUB0506* and *pedA* were detected in all isolates, but not with complete gene coverage (Fig. 5). Virulence properties in group C isolates were highly homologue and the majority of all isolates matched identical virulence genes. The isolate Su-18 showed a lower coverage for the locus tag *SUB1635*. The isolates Su-11 and Su-23 lacked the capsule gen *hasB2* (*SUB1027*) and the gene for the putative fructan beta-fructosidase precursor (*SUB0135; fruA*) but carried the gene for the glycosyl transferase (*SUB0699*). A total list of all genes with accession numbers and references can be found in Suppl. Table 2.

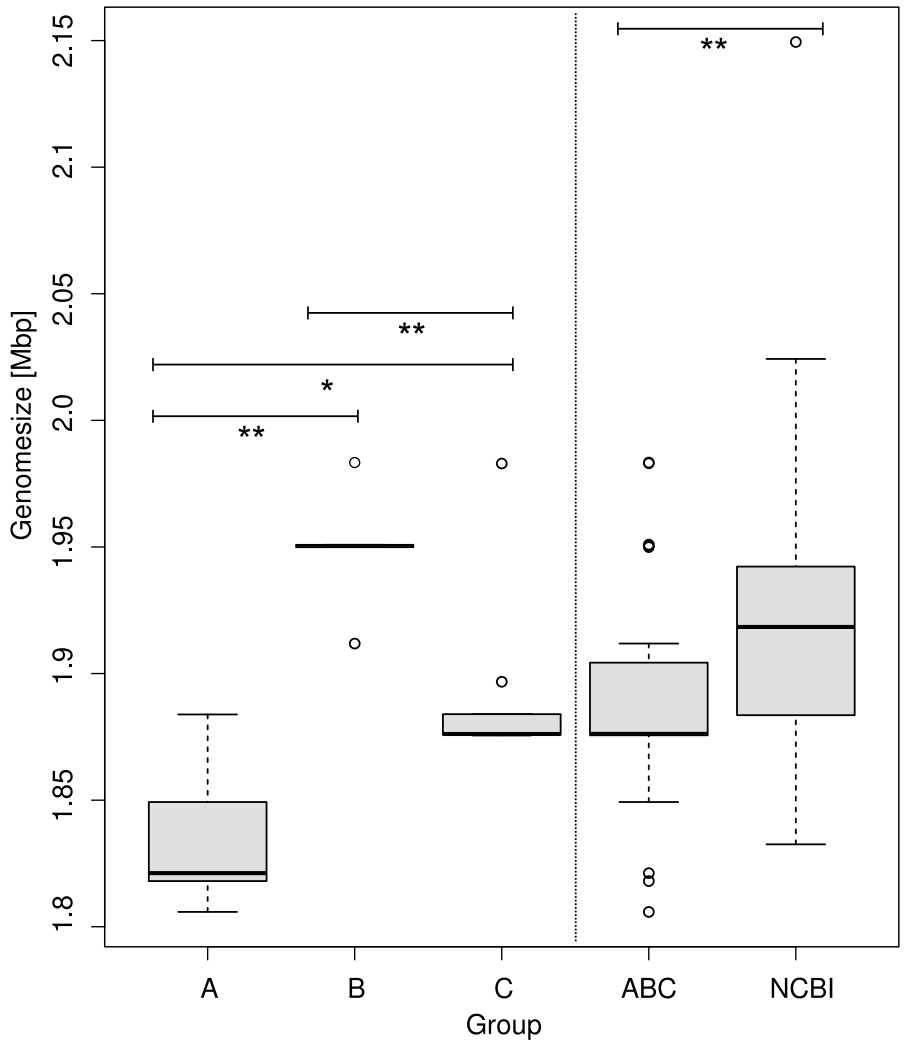


Fig. 1 There is a high statistically significant difference ($p < 0.01$) between groups A and B and between groups B and C. Groups A and C differed significantly ($p < 0.05$). There is a high statistically signifant difference between all genomes currently found in the NBCI database and the total genome size of group A, B and C combined. Boxplot and calculation was created with R-Studio

Antimicrobial resistance genes

In the search for putative AMR genes, those for the penicillin-binding-proteins (*pbp1a, pbp1b, pbp2a* and *pbp2x*) were found in all isolates. In the sequence of *pbp1b* an amino acid substitution $G_{769}S$ occurred in all isolates, with the exception of Su-06 and Su-08, as well as the strain EF20. In addition, the isolates Su-01 to Su-04 were found to contain the substitutions $P_{746}Q$ and $N_{117}K$. In the gene *pbp2a,* the substitution $I_{330}V$ was determined in all isolates, whereas the substitution $Q_{321}H$ could only be detected in isolates Su-11 and Su-23. The isolates Su-02, Su-03, Su-05 as well as the strain EF20 contained an amino acid substitution $K_{43}E$. Whereas $T_{36}A$ was substituted in isolates Su-06 and Su-08. Gene *pbp2x* showed substitutions $E_{381}K$, $Q_{554}E$, and $G_{600}E$ in all isolates except Su-11 and Su-23. All isolates were tested sensitive to

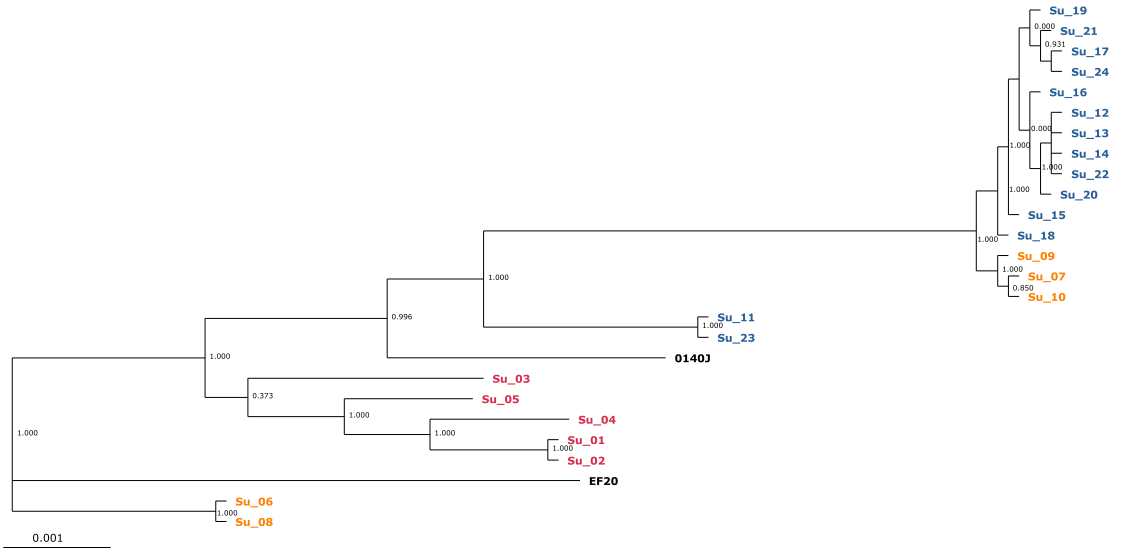


Fig. 2 Core genome phylogenetic tree. Core genes of these genomes were computed in EDGAR 3.0 based on muscle alignment. An approximately-maximum-likelihood phylogenetic tree was calculated using the FastTree software. The core genome analysis was based on 1567 genes per genome in 24 strains plus the reference strains 0140J and EF20. The core had 489,494 amino acid residues/bp per genome, 12,726,844 in total. Bar, 0.001 nucleotide substitutions per site. The values at the branches are Shimodaira-Hasegawa support values. Isolates are highlighted according to group affiliation (Group A: red; Group B: orange; Group C: blue)

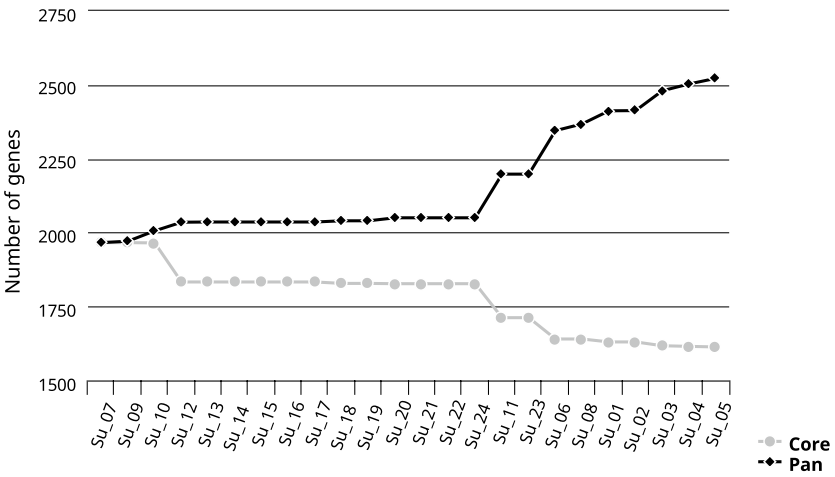


Fig. 3 Pan versus core development plot to gain insight into the stability of the pan- and core-genome of the isolates. Results were obtained using EDGAR 3.0. Starting with the first contig, consecutive numbers for the core and pan genome size were calculated and plotted. The plot shows an increase of the pan-genome (black line) and a decrease of the core-genome (grey line) as more genomes are added

typical β-lactam antibiotics in vitro (Table 2). Further putative resistance genes included hits in Su-06 and Su-08 containing the gene *lnuC*, whereas Su-01, Su-02, Su-04 and Su-05 carried the gene *lnuD,* both of which can lead to resistance to lincosamides. In Su-01 and Su-02, the gene *tetS*, coding for tetracycline resistance, was also confirmed. Furthermore, the gene *qacH* (*SUB0162*), which may leads to resistance to quaternary ammonium compounds, was found in all isolates with a coverage ranging from at

Antonie van Leeuwenhoek (2022) 115:857–870

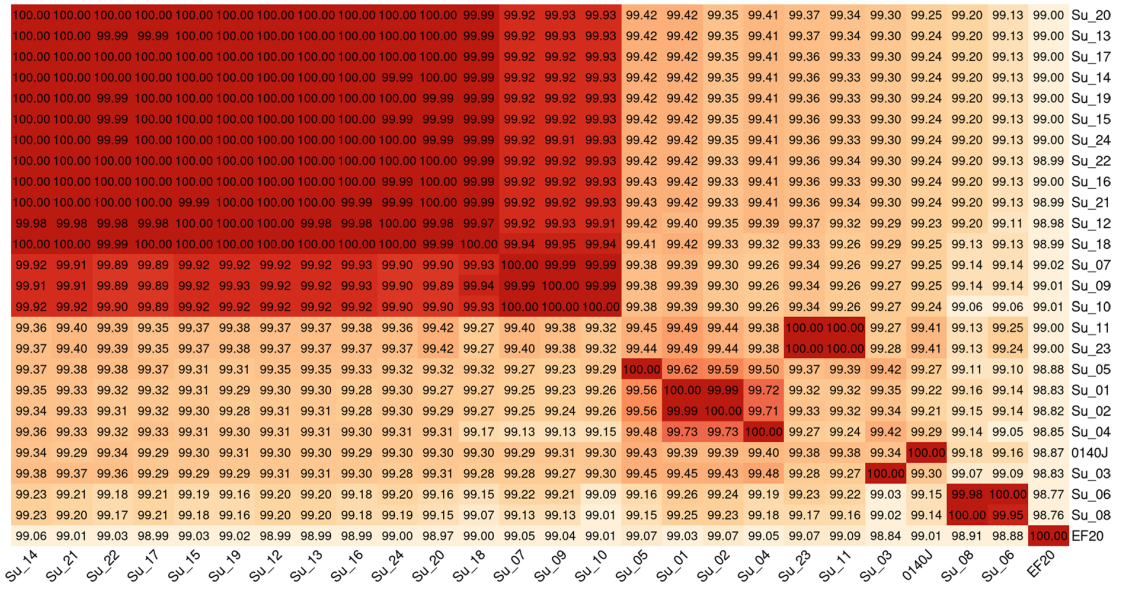


Fig. 4 Overview of the average nucleotide match between genomes. The isolates were grouped according to their match. Within each box, the identity was given in percent. Darker heat map colours indicate higher relatedness. Results were obtained using EDGAR 3.0 based on a BLASTN comparison of the genome sequences

least 50 % to as high as 100 % in the isolates Su-06, Su-08, Su-11 and Su-23.

Phages

A total of 12 different phages were found in all isolates, with the most commonly detected phages being *Streptococcus* phage SMP and *Lactococcus* phage bIL311 (Fig. 6; detailed summary in Suppl. Table 3). The *Streptococcus* phage phiNJ2 (Strept_phiNJ2) was most abundant as an intact region. Overall eight of the 24 isolates had intact prophage regions.

Discussion

The aim of this study was to identify genetic differences and similarities between 24 *Streptococcus uberis* isolates that proved responsible for highly different disease outcomes. Of particular interest was the question, whether there were clearly identifiable reasons as to why the isolates from farm C led to such a severe infection. Recent studies have found a smaller genome size in virulent strains of the closely related bacterial pathogen *S. suis* (Murray et al. 2021) and other studies also hypothesised this for *S. uberis* (Hossain et al. 2015). Despite the evidence for significant differences in the genome size of our isolates, there was no linkage between small genome size and potentially epidemic isolates, associated with higher virulence and a severe course of infection, or higher genome size and isolates that caused sporadic subclinical or clinical mastitis. Although group B isolates displayed significantly higher genome sizes compared to the other two groups, group A had the smallest genome size on average (as measured by total assembled contig size). However, it should be noted that group A and B are prone to relatively small sample sizes and therefore, the results are not necessarily exhaustive. Although the genomes examined in this study have a lower mean genome size than those currently found in the NCBI database, it is not clear whether there is a correlation between genome size and virulence, as not all genomes in the database are clearly labelled as clinical or subclinical isolates. Further studies are needed to fully clarify a potential correlation. Due to the similarity of the isolates from group C, the assumption of a clonal outbreak was reasonable to suspect. The phylogenetic analysis revealed a most closely related population of all

Table 4 MLST results for 24 isolates plus the strains 0140J and EF20

Isolate	Disease status	Allele							ST	CC
		arcC	*ddl*	*gki*	*recP*	*tdk*	*tpi*	*yqiL*		
Group A										
Su-01	C	1	37	4	2	62	1	3	**1375**	–
Su-02	C	1	37	4	2	62	1	3	**1375**	–
Su-03	C	1	1	4	2	**123**	1	3	**1377**	–
Su-04	C	2	33	4	2	**124**	1	3	**1378**	–
Su-05	C	1	37	5	2	62	1	3	**1379**	–
Group B										
Su-06	C	2	1	10	3	2	3	3	931	–
Su-07	C	2	1	5	1	**122**	1	3	**1374**	5
Su-08	C	2	1	10	3	2	3	3	931	–
Su-09	C	2	1	5	1	**122**	1	3	**1374**	5
Su-10	C	2	1	5	1	**122**	1	3	**1374**	5
Group C										
Su-11	C*	2	37	5	1	97	1	3	964	–
Su-12	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-13	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-14	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-15	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-16	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-17	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-18	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-19	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-20	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-21	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-22	C*	2	1	65	1	**122**	1	3	**1373**	5
Su-23	C*	2	37	5	1	97	1	3	964	–
Su-24	C*	2	1	65	1	**122**	1	3	**1373**	5
Reference strains										
0140J	–	1	1	1	1	1	1	1	1	5
EF20	–	15	1	4	3	13	4	3	55	–

In bold are the allele variants and resulting STs that were novel in this study. Isolates marked with asterisks (*), showed the particularly severe disease status

but two isolates from group C, further suggesting a clonal relationship in these isolates that represented the vast majority of the outbreak isolates in this herd. However, two isolates (Su-11, Su-23) from the same group clustered closer together with the strain 0140J, thereby supporting the general opinion of multiple, genetically non-related *S. uberis* within one herd. Interestingly, isolates Su-07, Su-09 and Su-10 from the randomly chosen farm B seemed to be next closely related to the clonal group of the isolates from farm C. The stability of the pan and core genomes within group C confirms the concept of a clonal outbreak and is consistent with the results of the phylogenetic tree.

Future studies should elucidate, whether the supposed clonal outbreak isolates from this study indeed represent a dominant genotype among other *S. uberis* strains that is linked with increased virulence and possibly with further geographic distribution.

With respect to MLST, 17 out of 24 isolates showed novel alleles, reflecting the high diversity in *S. uberis* as has already been mentioned previously in a similar study carried out with dairy cows from Australia (Vezina et al. 2021). Interestingly, the gene for the thymidine kinase (*tdk*) shows the most new variants here, also proven by the fact that the gene locus *tdk* is the most variable allele variant within the pubMLST database with 124 variants, whereas only

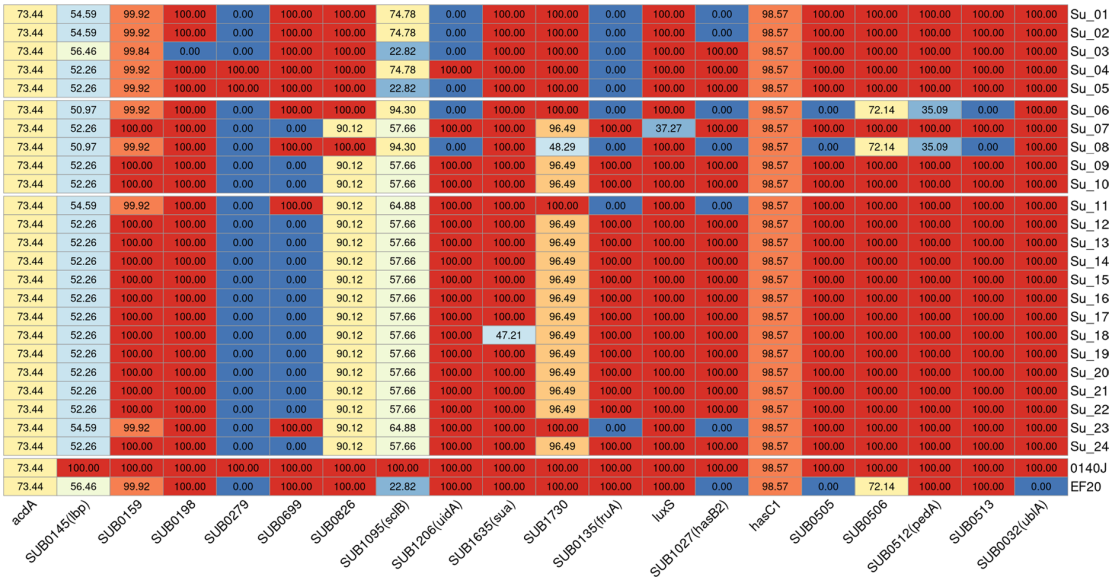


Fig. 5 All virulence genes that were not found in all isolates with a gene coverage of 100 % are listed. The gene-coverage in percent is indicated in the boxes. The heatmap was created with R-Studio from the results generated with ABRicate

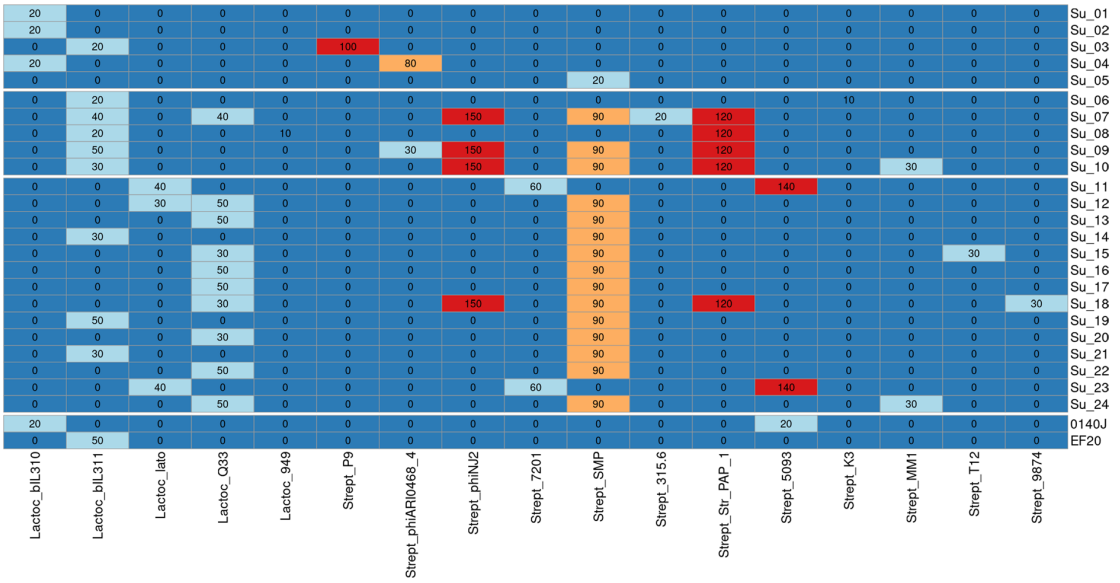


Fig. 6 All prophage sequences found are listed. The score (< 70: incomplete; 70–90: questionable; > 90: intact) for each sequence is indicated in the boxes. The heatmap was created with R-Studio from the results generated with PHASTER

37 to 85 variants are stored for the other gene loci. This could be taken as a reason to believe that this gene is not as stable as other housekeeping genes.

The similarity of the allele profiles in group C as well as their phylogenetic position in the core genome tree confirms once again that most of the isolates are

clones of the same type, which fosters to speculate that the spreading event inside the herd was caused by infected cows rather than from the environment. Nevertheless. it should not be disregarded, that this type could prevail as a dominant type in the environment and thus leads to predominant udder infections. Previous findings on this pathogen have also suggested that 50–100 % of all animals are infected with the same or a very similar strain in some herds (Zadoks et al. 2003). Despite the smaller sample size, this study also shows a great diversity of STs, mostly indicating a heterogeneous than contagious spread of *S. uberis* (Käppeli et al. 2019). However, the almost exclusive occurrence of ST1373 within group C as a new ST, could point to a contagious event with ST1373 as a ST possibly strongly associated with virulence. In addition, the two STs ST1373 and ST1374 belong to the clonal cluster CC5, which is most commonly associated with clinical mastitis (Zadoks et al. 2011).

Considering the equipment with virulence genes, all strains from this study principally appear to have a considerable virulence potential. More than half of all detected putative virulence factors occurred in all 24 isolates with complete gene coverage and almost complete sequence identity, indicating strong conservation of these sequences. As a significant finding contrary to our hypothesis, this would suggest that no significant differences do exist in the repertoire of known virulence genes that predispose isolates for causing severe and possibly epidemic or just local infections. This is also supported by the fact that, as already stated, the three isolates from farm B clustered closely together with all isolates from farm C. However, another possible explanation could indicate that epidemic isolates have a better adaptation to the udder tissue or that the animals were immunocompromised by pretreatments. In line with the traditional concept that *S. uberis* represents a pathogen whose pathogenic potential is highly dependent on environmental factors and host immune system, our results suggest that the fulminant disease progression in farm C was probably more likely being influenced by such factors than by a specific set of genes in the outbreak strain. No systematic data were available for the general health situation in farms A, B and C, but the accumulation of *S. uberis* mastitis in all these herds suggests an a priori sub-optimal immune status that is most often related to deficiencies in animal keeping, nutrition, co-infections or other health issues (Günther et al. 2016).

It is rather difficult to make a clear statement as to whether certain phages contribute to increased virulence here, because intact regions could not be found in any of the used reference strains. In addition, most intact phage regions were found within group B, consisting of isolates from local, rather mild infections. Thus, this will be the reason why in this group the mean genome size is higher than in the other groups. In any case, the diversity of prophage regions in this group may be indicative of phage adaptation to new hosts.

The size of the core genome of all 24 isolates is slightly higher than the results in other studies, whereas the size of the pan genome is smaller (Lang et al. 2009; Hossain et al. 2015; Vezina et al. 2021). This further suggests that *S. uberis* appears to be a genetically highly diverse organism. Six of the 24 isolates carried either *lnuC* or *lnuD* as putative antimicrobial resistance genes. While *lnuD* has already been found in *S. uberis* isolates from cases of mastitis, *lnuC* has so far mostly been reported in *S. agalactiae,* which could be an indication for lateral gene transfer according to (Vezina et al. 2021). The three substitutions found in the gene *pbp2x* could lead to increased resistance to oxacillin, according to the results of McDougall et al. 2020. This is a cause for concern, as β-lactam antibiotics, particularly penicillin G are first choice antimicrobials to treat *S. uberis* infections and are used during infection prophylaxis in dry cows, both of which could promote further mutations and resistance. The substitutions found in the other genes for penicillin binding proteins (*pbp1b*, *pbp2a*) have not yet been linked to β-lactam resistance. Nevertheless, since the substitutions are present in the majority of all isolates, i.e. seem to be evolutionary stable, it should not be completely ruled out that they could contribute to increased resistance. However, all isolates were tested sensitive to penicillin and oxacillin in vitro, which means that functional resistance cannot be confirmed here.

The gene *qacH* (*SUB0162*) has been described in previous studies with 41 % identity to the gene QACH_STASA in *Staphylococcus saprophyticus,* which codes for a quaternary ammonium compound-resistance protein (Ward et al. 2009). Disinfectants containing such compounds are often

routinely used as part of the daily milking procedure and could imply a certain selective pressure towards these substances. The fact that all isolates as well as both reference strains carried this gene should raise concern and warrant in vitro susceptibility testing as well as considering the prudent use and regularly change of other disinfectants in milking hygiene.

Conclusion

Comparison of all isolates revealed no clear set of genes responsible for enhanced virulence. Furthermore, no indubitable differences in overall gene content were found. However, it should be noted that most isolates within group C are highly similar, thus a spreading event from cow to cow and a common course for the specific disease progression is very likely. Furthermore, our study has also shown a highly conserved set of virulence genes. This should open new avenues for the search of potential vaccine targets. It also remains to be observed in more detail whether strains with fewer or no virulence genes are also responsible for mastitis cases or severe infections. To gain further clarity on the pathogenicity of *S. uberis*, one should include also host genomes in addition to bacterial genomes as well as the environmental circumstances. Such further metagenomic studies may reveal factors that can be used to better control the spread of *S. uberis* and thus reduce the risk of mastitis and the high use of antibiotics.

Author contributions IN and TE designed the microbiological study. TS performed the NGS. LF conducted data analyses, interpreted the data and wrote the manuscript. JB was involved in data analysis. All authors critically checked and contributed to the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. Hessian Ministry for the Environment, Climate Change, Agriculture and Consumer Protection supports the Hessian State Laboratory. We acknowledge access to resources financially supported by the BMBF grant FKZ 031A533 within the de.NBI network.

Data availability All genomes sequenced in this study are available at https://pubmlst.org; Ids 2255–2287 and at NCBI under the Bioproject: PRJNA795889. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent for publication All authors gave their consent to publish results from this study and to be listed as a co-author.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS (2016) PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res 44:16–21. https://doi.org/10.1093/nar/gkw387

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A, Lesin V, Nikolenko S, Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev M, Pevzner P (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. https://doi.org/10.1089/cmb.2012.0021

Bradley A (2002) Bovine mastitis: an evolving disease. Vet J 164:116–128. https://doi.org/10.1053/tvjl.2002.0724

Davies PL, Leigh JA, Bradley AJ, Archer SC, Emes RD, Green MJ (2016) Molecular epidemiology of *Streptococcus uberis* clinical mastitis in dairy herds: strain heterogeneity and transmission. J Clin Microbiol 54:68–74. https://doi.org/10.1128/JCM.01583-15

Dieckmann MA, Beyvers S, Nkouamedjo-Fankep RC, Hanel PHG, Jelonek L, Blom J, Goesmann A (2021) EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. Nucleic Acids Res 49:185–192. https://doi.org/10.1093/nar/gkab341

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG (2004) eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 186:1518–1530. https://doi.org/10.1128/JB.186.5.1518-1530.2004

Günther J, Czabanska A, Bauer I, Leigh JA, Holst O, Seyfert HM (2016) *Streptococcus uberis* strains isolated from the bovine mammary gland evade immune recognition by mammary epithelial cells, but not of

macrophages. Vet Res 47:13. https://doi.org/10.1186/s13567-015-0287-8
GVA (German Veterinary Association) (2012) Deutsche Leitlinien Bekämpfung des Mastitis des Rindes, 5th edn. German Veterinary Association, Gießen
Hogeveen H, Huijps K, Lam T (2011) Economic aspects of mastitis: new developments. N Z Vet J 59:16–23. https://doi.org/10.1080/00480169.2011.547165
Hossain M, Egan SA, Coffey T, Ward PN, Wilson R, Leigh JA, Emes RD (2015) Virulence related sequences; insights provided by comparative genomics of *Streptococcus uberis* of differing virulence. BMC Genomics 16:334. https://doi.org/10.1186/s12864-015-1512-6
Kabelitz T, Aubry E, van Vorst K, Amon T, Fulde M (2021) The role of *Streptococcus spp*. in bovine mastitis. Microorganisms 9:1497
Käppeli N, Morach M, Zurfluh K, Corti S, Nüesch-Inderbinen M, Stephan R (2019) Sequence types and antimicrobial resistance profiles of *Streptococcus uberis* isolated from bovine mastitis. Front Vet Sci 6:234. https://doi.org/10.3389/fvets.2019.00234
Lang P, Lefébure T, Wang W, Zadoks RN, Schukken Y, Stanhope MJ (2009) Gene content differences across strains of *Streptococcus uberis* identified using oligonucleotide microarray comparative genomic hybridization. Infection, genetics and evolution. MEEGID 9:179–188. https://doi.org/10.1016/j.meegid.2008.10.015
McDougall S, Clausen L, Ha HJ, Gibson I, Bryan M, Hadjirin N, Lay E, Raisen C, Ba X, Restif O, Parkhill J, Holmes MA (2020) Mechanisms of β-lactam resistance of *Streptococcus uberis* isolated from bovine mastitis cases. Vet Microbiol 242:108592. https://doi.org/10.1016/j.vetmic.2020.108592
Murray G, Charlesworth J, Miller EL, Casey MJ, Lloyd CT, Gottschalk M, Tucker A, Welch JJ, Weinert LA (2021) Genome reduction is associated with bacterial pathogenicity across different scales of temporal and ecological divergence. Mol Biol Evol 38:1570–1579. https://doi.org/10.1093/molbev/msaa323
Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A (2021) Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics 7:000685. https://doi.org/10.1099/mgen.0.000685
Thomas LH, Haider W, Hill AW, Cook RS (1994) Pathologic findings of experimentally induced *Streptococcus uberis* infection in the mammary gland of cows. Am J Vet Res 55:1723–1728
Vezina B, Al-Harbi H, Ramay HR, Soust M, Moore RJ, Olchowy T, Alawneh JI (2021) Sequence characterisation and novel insights into bovine mastitis-associated *Streptococcus uberis* in dairy herds. Sci Rep 11:3046. https://doi.org/10.1038/s41598-021-82357-3
Ward PN, Holden MT, Leigh JA (2009) Evidence for niche adaptation in the genome of the bovine pathogen *Streptococcus uberis*. BMC Genomics 10:54. https://doi.org/10.1186/1471-2164-10-54
Wellenberg GJ, van der Poel WH, Van Oirschot JT (2002) Viral infections and bovine mastitis: a review. Vet Microbiol 88:27–45. https://doi.org/10.1016/s0378-1135(02)00098-6
Wente N, Klocke D, Paduch JH, Zhang Y, Seeth MT, Zoche-Golob V, Reinecke F, Mohr E, Krömker V (2019) Associations between *Streptococcus uberis* strains from the animal environment and clinical bovine mastitis cases. J Dairy Sci 102:9360–9369. https://doi.org/10.3168/jds.2019-16669
Whiley RA, Hardie JM (2015) *Streptococcus*. In: Whitman WB et al (eds) Bergey's manual of systematics of archaea and bacteria. Wiley, Chichester, pp 1–86
Zadoks RN, Gillespie BE, Barkema HW, Sampimon OC, Oliver SP, Schukken YH (2003) Clinical, epidemiological and molecular characteristics of *Streptococcus uberis* infections in dairy herds. Epidemiol Infect 130(2):335–349. https://doi.org/10.1017/s0950268802008221
Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH (2011) Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia 16:357–372. https://doi.org/10.1007/s10911-011-9236-y

7.1.2. BakRep - a searchable large-scale web repository for bacterial genomes, characterizations and metadata

Linda Fenske[1], Lukas Jelonek[1], Alexander Goesmann[1], Oliver Schwengers[1]

[1]Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

Microbial Genomics (2024), DOI: 10.1099/mgen.0.001305

The following contributions are attributed to the thesis author:

Actively involved in formulating the research idea. Developed the analysis workflow, conducted statistical data analyses and interpreted the data Drafted the manuscript and finalized it with the co-authors.

MICROBIAL GENOMICS

RESEARCH ARTICLE

Fenske *et al.*, *Microbial Genomics* 2024;10:001305
DOI 10.1099/mgen.0.001305

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

BakRep – a searchable large-scale web repository for bacterial genomes, characterizations and metadata

Linda Fenske, Lukas Jelonek, Alexander Goesmann and Oliver Schwengers*

Abstract

Bacteria are fascinating research objects in many disciplines for countless reasons, and whole-genome sequencing (WGS) has become the paramount methodology to advance our microbiological understanding. Meanwhile, access to cost-effective sequencing platforms has accelerated bacterial WGS to unprecedented levels, introducing new challenges in terms of data accessibility, computational demands, heterogeneity of analysis workflows and, thus, ultimately its scientific usability. To this end, a previous study released a uniformly processed set of 661405 bacterial genome assemblies obtained from the European Nucleotide Archive as of November 2018. Building on these accomplishments, we conducted further genome-based analyses like taxonomic classification, multilocus sequence typing and annotation of all genomes. Here, we present BakRep, a searchable large-scale web repository of these genomes enriched with consistent genome characterizations and original metadata. The platform provides a flexible search engine combining taxonomic, genomic and metadata information, as well as interactive elements to visualize genomic features. Furthermore, all results can be downloaded for offline analyses via an accompanying command line tool. The web repository is accessible via https://bakrep.computational.bio.

Impact Statement

BakRep is a revolutionary web repository designed to enhance the findability and accessibility of sequenced bacterial genomes stored in public data repositories. By providing a uniformly processed and annotated dataset, BakRep addresses critical challenges in the microbiology field, including data integration, standardization and usability. We truly believe that this large-scale but accessible genome web repository will help microbiology researchers from various fields exploit this vast amount of genomes by compiling subsets as a starting point for their targeted analyses. Its user-friendly interface and extensive dataset make it an invaluable resource for the scientific community, supporting a wide range of comparative and clinical studies and driving new discoveries in microbial genomics.

DATA SUMMARY

The website can be accessed via https://bakrep.computational.bio. The workflow used for data analysis is available at https://github.com/ag-computational-bio/bakrep. Original data were retrieved via https://ftp.ebi.ac.uk/pub/databases/ENA2018-bacteria-661k. The source codes for the website and server design are available at https://github.com/ag-computational-bio/bakrep-web and https://github.com/ag-computational-bio, respectively. The accompanying command line tool is available at https://github.com/ag-computational-bio/bakrep-cli.

Received 25 June 2024; Accepted 19 September 2024; Published 30 October 2024
Author affiliations: [1]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany.
***Correspondence:** Oliver Schwengers, oliver.schwengers@computational.bio.uni-giessen.de
Keywords: annotation; bacteria; big data; computational biology; multilocus sequence typing; taxonomic classification; whole-genome sequencing.
Abbreviations: ENA, European Nucleotide Archive; FAIR, findability, accessibility, interoperability, and reusability; GTDB, genome taxonomy database; MLST, multilocus sequence type; WGS, whole-genome sequencing.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary figures and two supplementary tables are available with the online version of this article.

INTRODUCTION

Bacteria represent a significant portion of Earth's biodiversity, showcasing an astounding variety of habitats. For the past three decades, bacterial whole-genome sequencing (WGS) has provided deep insights into the vast diversity of populations and ecosystems' complexity, just as into the organization and plasticity of single genomes – both fundamental for our perception of microbial life. In particular, WGS of bacterial pathogens has tremendously propelled our understanding of drug resistances, virulence factors and host interactions and has become invaluable for medical microbiology. But simultaneously, the exploration and analysis of less-studied species continuously expand our knowledge of the broad and hard-to-comprehend diversity within the bacterial domain of life. However, the rapid and accelerating generation of WGS data demands substantial storage and analysis capacities. To securely store the raw DNA sequencing data, public databases like the Sequence Read Archive, the DNA Data Bank of Japan or the European Nucleotide Archive (ENA) are primarily considered [1]. Consequently, these data repositories are in a constant state of growth. For example, at the time of writing, more than 4.7 billion sequences are stored in the ENA (https://www.ebi.ac.uk/ena/browser/about/statistics), and the latest GenBank release (v261.0) contains about 3.4 billion WGS records (https://ncbiinsights.ncbi.nlm.nih.gov/2024/06/20/genbank-release-261). Along with these rapidly growing data collections, several challenges arise with regard to the FAIR (findability, accessibility, interoperability, and reusability) principles [2]. *Findability*: to conduct comparative analyses targeting particular sublineages or multilocus sequence types (MLSTs), sequenced samples often need to be processed prior to genome-based screening and filtering steps. *Accessibility*: the sheer amount of raw data needs to be handled and properly processed for analysis, which poses a serious barrier for many researchers lacking necessary IT infrastructure and bioinformatics skills. *Interoperability*: common data formats, vocabularies and ontologies are crucial to facilitate data integration across different platforms. *Reproducibility*: large parts of this data are processed over and over again, introducing adverse variability regarding used analysis tools, parameters and databases. Furthermore, user-provided metadata may be prone to inaccuracies and incompleteness complicating reproducibility and subsequent processing [3, 4]. In conclusion, this situation leads to inflated bioinformatic workloads, increasing analysis costs regarding computational resources and valuable staff time. The analyses of genomes of varying quality, assembled and annotated using different algorithms, ultimately put the usability of this valuable data at stake [5, 6]. In contrast to the large raw data repositories, dedicated initiatives, e.g. EnteroBase [7], conduct consistent data processing procedures comprising targeted and streamlined genome characterizations. However, these platforms typically focus on distinct taxa and thus are of limited general usability. An essential step addressing these challenges was made by a previous study by Blackwell *et al.* following a uniform approach to assemble and characterize all bacterial paired-end WGS datasets retrieved from the ENA as of November 2018 [8]. As a result, 661 405 consistently assembled genomes were made publicly available, facilitating the broader access and utilization of these data for the research community. This study accomplished the systematic and standardized processing of this massive dataset and thus fostered the usability of these genomic data. However, access to these genomes remains limited, as all genome FASTA files are provided as one comprehensive 751 GB single-file archive, thus posing a significant barrier in terms of findability and accessibility for further analyses. Even though assembled genomes are pre-indexed using various search algorithms, it remains challenging for users without sufficient bioinformatics knowledge or command line skills to find and extract genomes of interest. Hence, to fully exploit the huge potential of this highly valuable dataset, researchers would benefit from a user-friendly platform providing streamlined access to this huge amount of data via flexible search capabilities integrating the various information layers, like genome characterizations, taxonomic classifications and subtypings, annotated genomic features and last but not least metadata. Building on these uniformly assembled bacterial genomes, here, we present BakRep, a large-scale comprehensive web repository specifically addressing these challenges. All 661 405 genomes were consistently quality controlled, taxonomically classified, multilocus sequence typed and annotated. In line with the FAIR principles, all information is findable and accessible via an interactive website providing researchers with a versatile search engine integrating genomic and taxonomic information, annotated features and original metadata. Batch downloads of search results can be conducted via an accompanying command line tool. BakRep is publicly available at https://bakrep.computational.bio.

METHODS

Raw data processing

We retrieved 661 405 assemblies and associated metadata published by Blackwell *et al.* from http://ftp.ebi.ac.uk/pub/databases/ENA2018-bacteria-661k, which were assembled using Shovill (v1.0.4) including several pre- and postprocessing steps. Further details can be found in the original publication [8]. For taxonomic classification, the GTDB-Tk (v2.2.6) classify workflow [9] based on the Genome Taxonomy Database (GTDB) release R207 [10] was used, with the '--mash_db' argument set for enabling ANI screening. Contamination and completeness of the assemblies were estimated with CheckM2 (v1.0.1) [11]. Basic statistics of the raw assemblies were collected using assembly-scan (https://github.com/rpetit3/assembly-scan). Determination of MLST was conducted using mlst (v2.23.0) (https://github.com/tseemann/mlst) utilizing the PubMLST database. Furthermore, assemblies were annotated with Bakta (v1.7.0) using the 'full' database version to use all features and the '--keep-contig-headers' flag to preserve the original contig headers of the raw assemblies [12]. Results were stored as JSON files via custom Python scripts. All analyses were implemented as part of a Nextflow [13] workflow executed in the de.NBI consortiums' cloud computing

infrastructure (https://github.com/ag-computational-bio/bakrep). The Metacoder package (v0.3.6) was used for graphical summaries of the taxonomic abundances [14].

Implementation of the web repository

The BakRep web repository is implemented as an HTTP-based API, based on Vert.x, offering public endpoints for search and data access [15]. Elasticsearch is utilized for implementing the search functionality [16]. All genomic data are stored as compressed plain text files in a S3-compatible storage. We provide a publicly available website that retrieves data via the API and visualizes it. The website's graphical user interface is implemented as a single-page application in Vue.js 3 (https://vuejs.org). The services are deployed on a scalable Kubernetes cluster, which is currently hosted and runs within the cloud computing infrastructure of the de.NBI consortium. An additional command line tool for automated large-scale downloads was implemented in Python (https://github.com/ag-computational-bio/bakrep-cli).

RESULTS

Expansion of consistent genome analyses

In this study, we built up on the 661405 bacterial genome assemblies provided by Blackwell *et al.* who uniformly assembled WGS raw data retrieved from the ENA archive as a November 2018 snapshot. We aimed to expand the range of consistent per-genome characterizations and to provide these results as accessible and user-friendly as possible. In this regard, all 661405 assembled genomes (751 GB in total) were quality-checked, and basic assembly statistics were calculated. We then taxonomically classified all genomes using the robust GTDB taxonomy, and where applicable, we further sequence-typed all genomes for which a species-specific multilocus sequence typing scheme existed. Last but not least, we performed a robust annotation of all genomes using Bakta taking advantage of its taxonomically untargeted full database version. From these 661405 input assemblies, 648567 were successfully characterized. To streamline the technical accessibility of all results, output files of all analysis tools were parsed, normalized and serialized in JSON format, generating a total of 3891402 unique files. In addition, annotation results are also available in GenBank format, as well as nucleotide and amino acid FASTA files for all annotated coding sequences. A total of 6.15 TB of genomic information was generated and stored in a cloud-based S3 storage, which is publicly available via an interactive web repository at https://bakrep.computational.bio.

Diversity and bias across the various taxonomic ranks

Given the vast size of public databases, they naturally encompass a variety of species. Nevertheless, certain species receive more frequent attention due to their clinical relevance, ease of cultivation or long-standing usage as model organisms. This inherent bias contributes to taxonomic imbalances in such data repositories. Blackwell *et al.* comprehensively demonstrated an intrinsic taxonomic bias at both the genus and species levels [8]. However, we would like to address one more aspect: to what extent is there either bias or diversity at higher taxonomic ranks? Therefore, we comprehensively explored the distribution across all taxonomic ranks, utilizing a robust, purely genome-based and thus objective taxonomic classification. We used GTDB-Tk, a widely utilized tool in the community, which delineates prokaryotic taxa based on systematic criteria and phylogenetic relationships using domain-specific marker genes in combination with mutual ANI-based genome distances. At the species level, and in line with former results, our analysis revealed that the 24 most prevalent species constitute 90% of all genomes. The most abundant species were as follows: *Salmonella enterica* (27.10%), *Escherichia coli* (13.52%), *Streptococcus pneumoniae* (7.80%), *Mycobacterium tuberculosis* (7.43%) and *Staphylococcus aureus* (7.28%). At the genus level, the most prevalent genera were as follows: *Salmonella* (27.99%), *Escherichia* (13.82%), *Streptococcus* (12.89%), *Mycobacterium* (8.6%) and *Staphylococcus* (7.92%). However, despite these over-represented species and genera, the genomes contained in this repository exhibit a notable degree of diversity at higher taxonomic ranks, comprising 66 distinct phyla divided into 132 classes, 345 orders, 722 families, 2466 genera and 8207 species. In comparison, the genome sequence-based GTDB database counts 175 phyla, divided into 538 classes, 1840 orders, 4870 families, 23 112 genera and 107 235 species (https://gtdb.ecogenomic.org), and the literature-based Bacterial Diversity Metadatabase (BacDive) lists 42 phyla divided into 106 classes, 255 orders, 648 families, 3801 genera and 21203 species (https://bacdive.dsmz.de/dashboard). Thus, this repository covers 37% and 157% of phyla, 24% and 124% of classes, 18% and 135% of orders, 14% and 111% of families, 10% and 64% of genera and 7% and 38% of species available in the genome-based GTDB and described in the literature-based BacDive databases, respectively. To illustrate both the diversity and bias of this repository, a taxonomic tree weighted by aggregated genome counts along all ranks was created (Fig. 1). For better visualization, taxa were clipped and aggregated at the family level. A more detailed version including all ranks is available in the supplemental data (Fig. S1, available in the online version of this article). Notably, 1634 assemblies (0.25%) could not be assigned to any species epithet, of which 122 (0.02%) could not be assigned to a genus. A closer examination of the unclassified genomes revealed that those lacking a genus assignment exhibit an average estimated completeness of only 46.50%. Genomes lacking a species epithet classification exhibited a higher average completeness of 67.02%, albeit with increased variability (Fig. S2).

Various comparative analyses begin with the selection of suitable genomes from public repositories. Here, the reliability of pre-assigned taxa is of utmost importance, immediately impacting the outcome of comparative studies. Unfortunately, user-provided

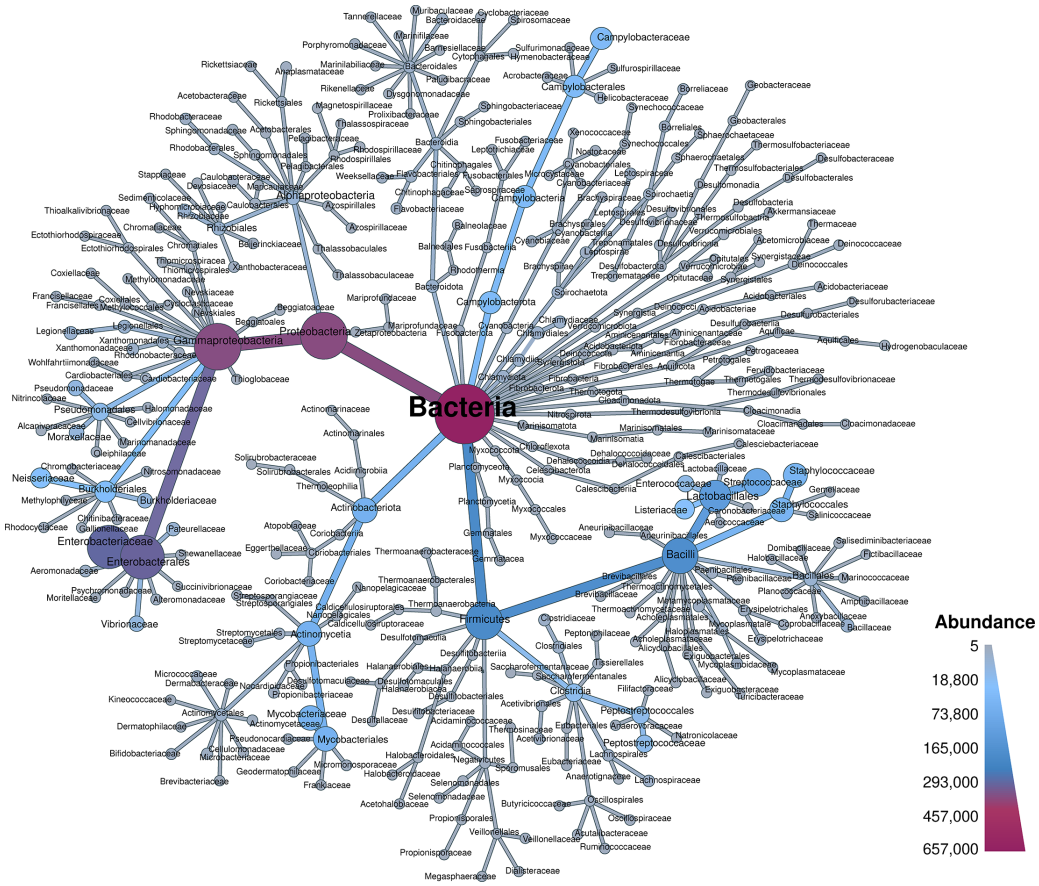


Fig. 1. Overview of the taxonomic composition at the family level. Nodes and branches are coloured and sized by aggregated genome counts at each taxonomic rank. The figure was created using the Metacoder package.

taxonomic information stored as metadata in raw sequencing data archives is known to be error-prone and often does not correspond to genome-based taxonomic classifications. To address this issue, we compared the scientific taxonomic names associated with the raw data in the ENA with the genome-based taxonomic classifications conducted with GTDB-Tk. For 45 275 (6.98%) genomes, we observed discrepancies at the species epithet. Variations at the genus level occurred in 25 913 (4.0%) genomes. In 21 349 (3.29%) cases, both the genus and species epithets differed. On further review, a substantial portion of these discrepancies (54.96%) is attributed to the genus *Shigella*, which was consistently classified as *Escherichia*. Frequent inconsistencies were also evident for *Mycobacteroides abscessus,* designated as *Mycobacterium abscessus* in 2675 cases (10.32%), and *Burkholderia pseudomallei* classified as *Burkholderia mallei* in 1763 cases (6.80%). Among the 2774 (10.71%) species discrepancies within the *Salmonella* genus, variations arose from assigning distinct subspecies, designated as full species names by GTDB-Tk. Considering these examples, 7106 (33.28%) cases remained for which neither the genus nor the species epithets matched (Table S1).

Distribution of genome-based key metrics

In the next-generation sequencing (NGS) era, a multitude of sequencing platforms as well as constantly evolving bioinformatic methods and implementations contribute to a variety of assembly approaches. To quickly assess biological key features and the technical quality of assembled genomes, several metrics have evolved as gold standard indicators. For instance, the mere size of a genome alone can provide important information regarding its completeness. Also, the GC content is widely used as a rough proxy for the nucleotide composition of a genome that is typically found in a narrow range specific to a particular bacterial species. We

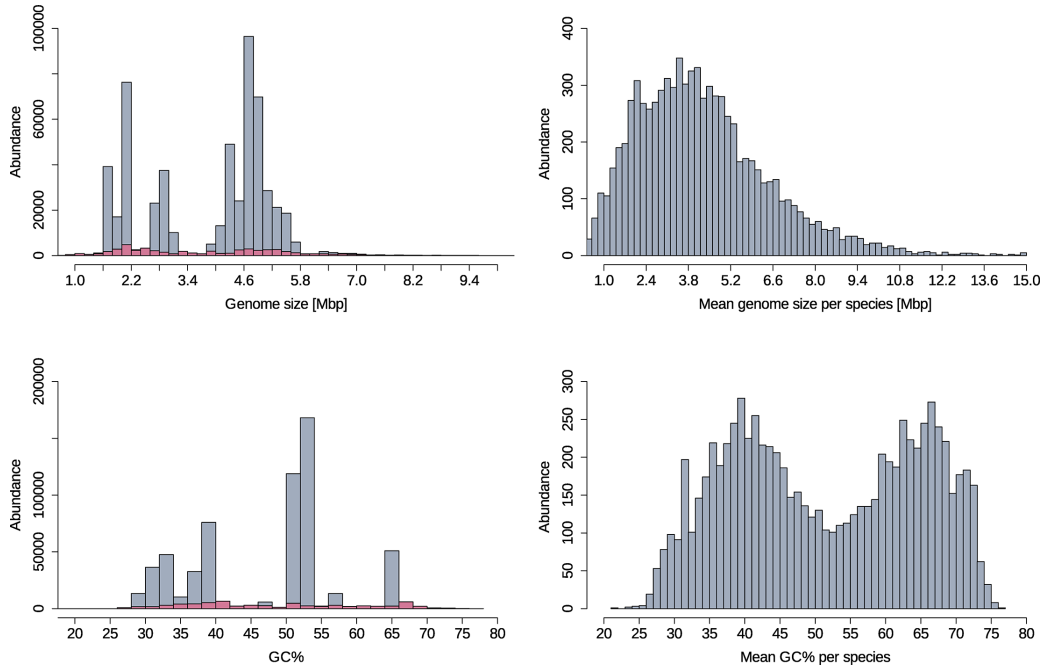


Fig. 2. Distribution of genomic metrics in the repository. The genome size (top left) and the GC content (bottom left) are displayed for all genomes. In comparison, the mean genome size (top right) and the mean GC content (bottom right) per species are shown. The magenta-highlighted region illustrates the distribution excluding the 24 most abundant species.

used the available information stored in BakRep to get an overview of the distribution of some of the most important and widely used metrics for this repository. To better understand the extent of variation within the bacterial diversity, we summarized the overall distribution of the genome size and GC content. The total genome size ranges from a minimum of 100943 bp to a maximum of 20285777 bp, with a mean value of 3901303 bp and a median of 4379349 bp. To account for the observed taxonomic biases, we excluded the 24 most abundant species accounting for 90% of all genomes. For this taxonomically clipped set of genomes, the maximum and minimum genome sizes remain unchanged, while the mean genome size increased to 3962704 bp, and the median decreased to 3853294 bp. To further mitigate the influence of over-represented regions in the genome size distribution, primarily attributed to the *Enterobacteriaceae* in the range of 4.5–5.5 Mbp, the *Mycobacteriaceae* in the range of 1.5–2.0 Mbp and the *Vibrionaceae* and *Neisseriaceae* in the range of 2.5–3.5 Mbp, we calculated the mean genome size per species. In contrast, this reveals a notably homogeneous distribution with a peak at ~3.8 Mbp, followed by a rapid decline extending to a maximum of 15 Mbp (Fig. 2). The GC content of all genomes ranges from a minimum of 23.6% to a maximum of 76.5%, with a mean value of 47.2% and a median of 50.7%. Distinct peaks are observed at ~40%, within the 50–55% range and at 65%, mostly attributed to the 24 over-represented species. The GC content was likewise normalized based on species, resulting in a bimodal distribution that peaks at 40% and between 60 and 70% (Fig. 2).

In addition to the genome size and GC content, further metrics evolved to quickly assess the technical quality of a sequenced and assembled genome, e.g. the number of contigs and the well-known N50 metric. However, actual values for these metrics can vary widely not only between sequencing platforms and assembly approaches but also between species due to biological factors like the existence and abundance of sequence repeats and mobile elements. Furthermore, due to the lack of common guidelines, it is often far from obvious which actual values are acceptable for a given metric. Hence, we leveraged the robust taxonomic classifications and vast size of this repository to aid with the provision of potential guidelines for acceptable value ranges of these key metrics per species. Hence, we examined the distributions of the aforementioned key metrics for each of the most prevalent species, including many of significant medical relevance. As anticipated, we observed substantially varying value ranges for these metrics across species (Fig. 3). Additionally, the distribution ranges within individual species also showed considerable variability. For instance, for *Klebsiella pneumoniae*, we observed notable downward deviations, with some isolates exhibiting a minimum genome size ranging from 2.8 to 4.0 Mbp, while the mean is 5.5 Mbp. However, upon closer observation, most of these outliers were identified as several isolates from the same study that utilized transposon-directed insertion-site sequencing, suggesting

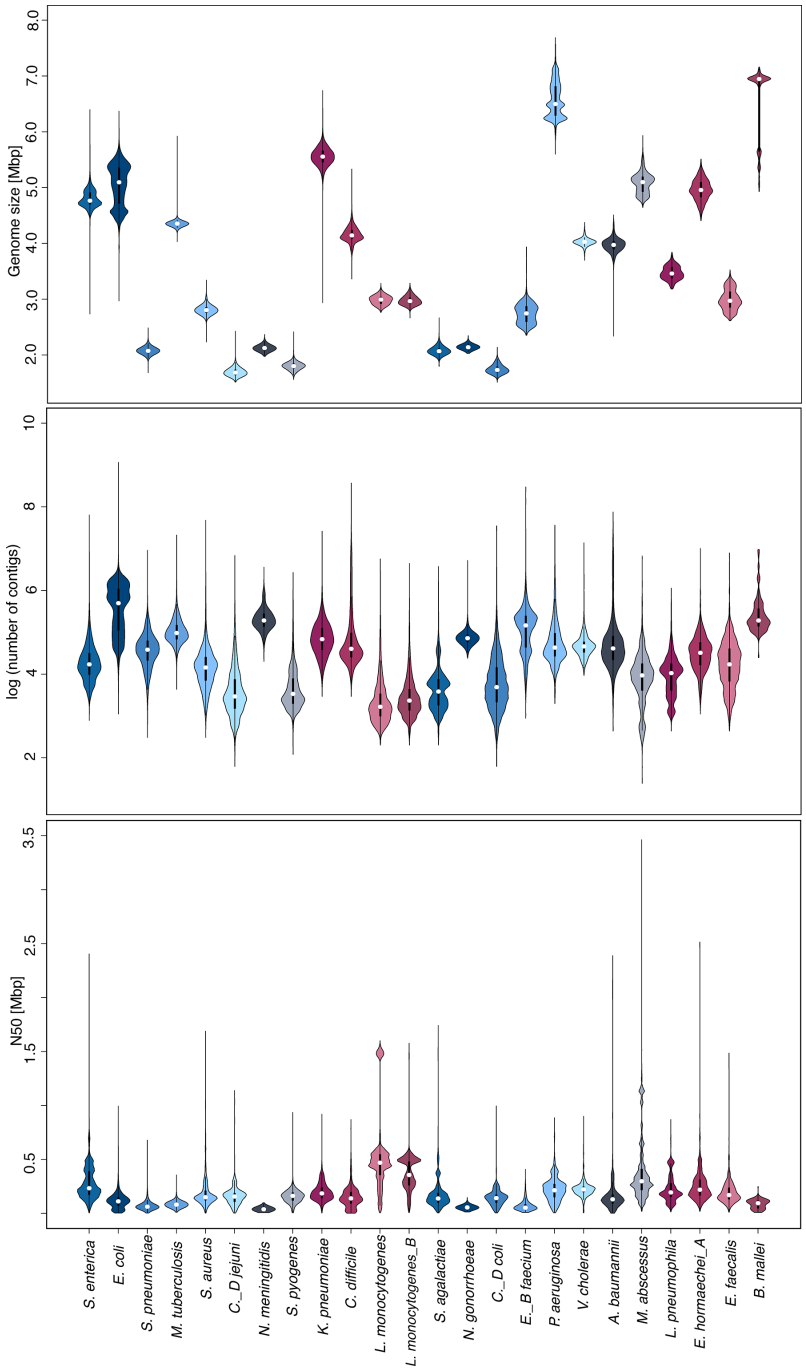


Fig. 3. Distribution of genome assembly metrics for the 24 most abundant species. The genome size (top), the number of contigs (middle) and the N50 values (bottom) are displayed. White points indicate medians; bold black bars represent interquartile ranges and thin black lines represent outliers. Genomes were filtered for 95% completeness and less than 1% contamination.

that these samples were not whole-genome sequenced. Discrepancies also exist for *B. mallei,* with likewise noticeable downward deviations for which no clear explanation could be found within the metadata. Despite some outliers, core ranges of these key metric values might help to establish guidelines for quality assessment.

The demonstrated varying ranges in genome sizes in the preceding section are outcomes of different habitats and evolutionary mechanisms constantly introducing and removing genes. Due to the intricate and diverse set of ecosystems, bacterial genomes exhibit significant variability in size and complexity, encompassing a fluid continuum between compact genomes and those with larger and more elaborate structures. As a rule of thumb, it is accepted as common knowledge that bacterial gene lengths average ~1 kbp per gene. To assess this assumption, we juxtaposed the mean genome sizes with the mean number of genes per species. A regression analysis revealed a slope of ~915 genes per 1 Mbp, resulting in a mean gene length of 1093 bp, roughly validating but specifying this assumption with a deviation of 9.3% and a determination coefficient (R^2) of 0.98, confirming the postulation of a linear relationship between genome size and the number of coding genes. Besides, non-coding RNA features also play pivotal roles in bacterial genomes and cellular processes, like for example, non-coding RNAs (ncRNAs), recognized for their regulatory functions, and transfer and ribosomal RNAs (tRNAs/rRNAs) as essential components of the protein synthesis machinery. Hence, we likewise compared the numbers of annotated non-coding RNA features to the mean genome size per species. Here, tRNAs showed a linear correlation, however, with significant variability (R^2 = 0.56). In contrast, for ncRNAs (R^2 = 0.35) and rRNAs (R^2 = 0.15), no clear linear trend could be observed (Fig S3).

Interactive and command line access via a searchable web repository

A major part of research would be constrained or rendered infeasible without accessible data. While a high level of standardization is crucial, it is equally essential to consider the ease of data findability and accessibility. For example, in outbreak analyses, for which the presence of specific antibiotic resistance genes is pivotal, it is essential to systematically search for genomes of a particular species characterized by distinct features such as MLST or virulence factors. To ensure the accessibility of our results, all data were stored in a public S3 bucket. To further ensure the findability of genomes of interest, we developed and provided an interactive web page that is publicly available at https://bakrep.computational.bio. It offers diverse search and filter options, allowing and streamlining the compilation of customized cohorts. To obtain an initial comprehensive overview, all available genomes can be browsed by the GC content, number of contigs, genome size, and estimated completeness and contamination levels. To conduct comprehensive and detailed large-scale searches, BakRep offers an advanced search engine that enables robust scalable queries flexibly combining various information like genome size, GC content, number of contigs, sequence type, different taxonomic ranks and annotated gene symbols or protein product descriptions. Furthermore, and in addition to genome analysis-based information, users also have access to quality-controlled metadata associated with each dataset upon the initial raw data submission to the ENA. Therefore, the repository supports the filtering of genomes based on various metadata, including isolation source and time, associated host species and project affiliation, enabling targeted searches by criteria such as country of origin, isolation period or host organism. A more detailed list including all possible search tags is available in the supplemental data (Table S2). To name an example, in one of our ongoing research projects, we utilized this search engine to identify all *Streptococcus agalactiae* genomes that met specific quality criteria, were isolated from humans, belonged to sequence type 17 and contained the penicillin-binding proteins *pbp1a*, *pbp1b*, *pbp2a* or *pbp2X* (Fig. 4). A summary of the particular search results can be exported in the tab-separated values (TSV) format. All individual genomes are displayed in human-readable formats such as a summary table, a feature table and an igv.js-based genome browser [17], and provided cross-links to databases such as the GTDB [10], RefSeq [18] or UniProt [19]. Each analysis result can be accessed and downloaded per genome via the website. To facilitate extensive analyses with the download of larger genome cohorts, we offer access to the download backend through a dedicated command line tool accessible via https://github.com/ag-computational-bio/bakrep-cli.

DISCUSSION

With today's rapid and cost-effective sequencing technologies, vast amounts of bacterial sequence data are generated daily. Submitting genomic data to public repositories is now essential for supporting research and ensuring open access to valuable information. While these databases hold extensive bacterial WGS data with great potential, inconsistencies and lack of standardization can hinder research. Challenges stem from varied assembly methods and quality control, introducing potential batch artefacts in large-scale analyses. To this end, Blackwell *et al.* assembled 661405 bacterial WGS samples from the ENA as of November 2018 in a standardized manner. Building on this, BakRep adds further analyses – such as assembly metrics, taxonomic classifications, MLST subtyping, genome annotations and original metadata – while providing streamlined access via an interactive website with robust search capabilities.

The sequencing of bacterial genomes has become routine, significantly reshaping our understanding of the bacterial world with information gleaned from tens of thousands of genomes. Nevertheless, this quantity exhibits a notable skew towards specific phyla housing, e.g. particular model organisms [20]. This taxonomic bias of just a few species making up the majority of genomic data may be due to various factors, such as the over-representation of well-researched and easily cultivable species, leading to gaps in the representation of the lesser researched or uncultivable microbial diversity. Furthermore, many sequencing projects focus

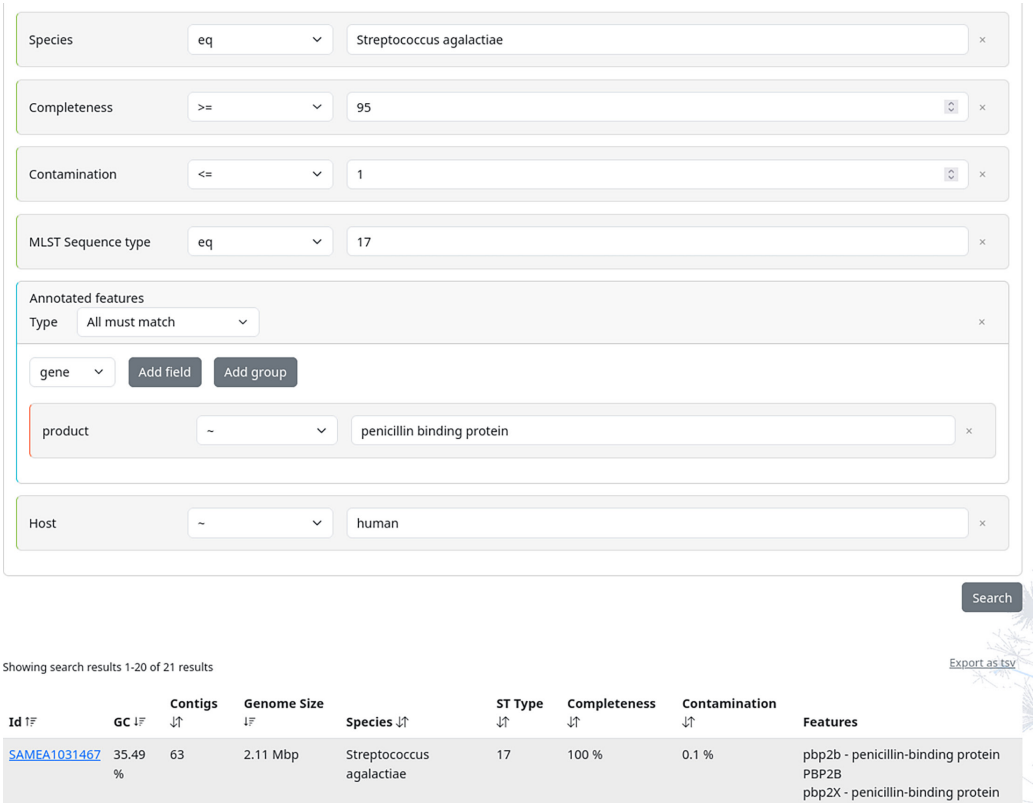


Fig. 4. Overview of the search function of the BakRep web repository. Advanced queries for genomes with specific characteristics such as species, completeness, contamination levels, sequence type, annotated features or host species are possible. The search outcomes are presented in a concise summary table. Details of each dataset are provided on a separate page.

on certain pathogens or organisms with global significance. For example, the GenomeTrakr network represents the inaugural distributed collaboration of laboratories employing WGS for pathogen identification [21] or the '10000 Salmonella genomes project', which sequenced more than 10000 *Salmonella* isolates [22]. This shows the impact of funding and scientific emphasis on the diversity of sequences. In contrast to the approach employed by Blackwell *et al.*, our study presents a more intricate portrayal of the taxonomic distribution through systematic species assignment utilizing the GTDB. They already acknowledged that certain aspects of sequence diversity within the assemblies might have been overlooked due to constraints inherent in the Kraken 2 database, which they used for taxonomic assignment and abundance estimation [8]. In contrast, the taxonomic classification in this study was conducted using the GTDB, employing a normalized genome-based classification derived from phylogenetic trees. These trees were constructed using a concatenated protein phylogeny, serving as the foundation for bacterial taxonomy. This approach conservatively eliminates polyphyletic groups and normalizes taxonomic ranks based on relative evolutionary divergence [23]. However, the GTDB currently enumerates 175 phyla, divided into 538 classes, 1840 orders, 4870 families, 23112 genera and 107235 species. This indicates that our dataset covers 37% of those phyla, 24% of classes, 18% of orders, 14% of families, 10% of genera and 7% of species, highlighting its limited scope and underscoring that it encompasses only a fraction of the extensive bacterial diversity. There is a need to shift emphasis from a strong focus on known pathogens in sequencing projects towards underrepresented and unknown species. This approach is crucial for a more comprehensive understanding of the patterns within bacterial diversity.

Examining several assembly metrics provides valuable insights into bacterial genomes, aiding in understanding their genetic diversity, evolutionary relationships, functional roles and taxonomic classification. The bias of over-represented species in the repository is also evident regarding these metrics. Nevertheless, the absence of prominently discernible gaps in the distribution of the mean genome size per species instils confidence that this snapshot may nevertheless encapsulate

a substantial portion of bacterial diversity. However, while bacteria can attain genome sizes of up to 16 Mbp [24], the upper ranges are only poorly represented here. A recent study mentions a connection between the distribution of genome sizes and ecosystem type or associations with hosts, and it also discusses the ongoing challenge of precisely defining the distribution of genome size beyond the confines of laboratory settings [25]. Another study postulated an indirect mechanism of natural selection whereby ancient adaptations have induced alterations in the bacterial genome, contributing to a bimodal distribution pattern of genomic GC, which we also observed here [26]. While the genome size of a species may show some variability, caution should be exercised when encountering pronounced outliers. Substantial deviations from the mean literature value may indicate potential issues with quality, possible contamination or the sequencing of partial segments rather than the entire genome, as exemplified in *K. pneumoniae*. Knowledge gained from a comprehensive and standardized analysis of numerous bacterial genomes has the potential to contribute valuable insights, aiding in the formulation of robust guidelines specific to certain species. Empirical values derived from diverse biological samples might offer more reliable guidelines than solely relying on literature values established over the years only using a few type strains or reference genomes. By encompassing a broader array of datasets, our repository helps to generate such guidelines. This extensive collection allows for more robust analysis and comparisons. Consequently, researchers can develop more nuanced and reliable guidelines that better reflect the complexity of bacterial genomes.

Genome fragmentation is a prevalent issue associated with short-read sequencing technologies. This challenge stems from the generation of shorter DNA fragments during the sequencing process, leading to genome assemblies typically consisting of an increased number of contigs. A recent article mentioned that the quality of these genome sequences may suffice for most analyses but needs to be more practical for comparative genomics [27]. Given the typical size of a bacterial genome, a genome with a high number of contigs would result in smaller contig sizes. Referring to the average gene size of 1093 bp, which we have calculated here, smaller contigs may contain at most one complete gene, with fragmented genes may frequently appear at contig boundaries. Therefore, genomes with a particularly large number of contigs should be approached with caution, which may limit the usefulness of such an assembly. Due to the fact that the underlying assemblies in this study are based solely on Illumina sequencing, to improve the dataset, it would be beneficial to include other sequencing techniques. Another important aspect in this context is which genome annotation pipeline to choose in order to predict and annotate bacterial genes. In this study, we chose Bakta [12] in favour of other state-of-the-art software tools, due to its favourable combined performance in terms of comprehensive functional annotations and wallclock runtimes – the latter being a very important property regarding the large number of processed genomes. Other genome annotation pipelines might, however, provide better results regarding structural gene predictions and pseudogene detections, which in turn would have a beneficial effect, also on the average gene size calculation. Unfortunately, due to the vast amount of genomes processed in this repository, a direct comparison could not be conducted.

The presence of taxonomic misclassified species in public repositories is of significant concern to researchers, as it can introduce inaccuracies into various analyses, thus impacting the reliability of the findings. Furthermore, classification errors can propagate over time as incorrectly labelled genomes are used as references to identify novel sequences. Specific errors may stem from taxonomic naming inconsistencies or the frequent reclassification of organisms prompted by new discoveries. In our study, this applies, e.g. to the discrepancies found with the genus *Shigella,* as *Shigella* species were reclassified as later heterotypic synonyms of *E. coli* in the GTDB [28]. The variations in the nomenclature of *Burkholderia* species can be similarly explained, given that *B. mallei* can be characterized as a recently evolved, host-adapted clonal lineage derived from *B. pseudomallei* [29]. This may also explain the observable variations in the genome size. During host adaptation, *B. mallei* experienced considerable genome reduction [30]. Given that *B. mallei* and *B. pseudomallei* share over 99% of genetic homology, taxonomic transitions between them can be fluid [31]. As GTDB-Tk uses an operational average nucleotide identity-based approach relying on type strains, only a few different genes will not lead to species differentiation. Unfortunately, we were initially unaware of the extensiveness of taxonomic discrepancies in the dataset, and thus, we decided to use species information associated as metadata for our genome annotation processes. This will certainly be addressed in future versions by using GTDB-Tk species classifications, ensuring accurate and consistent species listings down to annotation result files.

Adherence to the FAIR principles – Findability, Accessibility, Interoperability and Reusability – is crucial for advancing genomic research. With our public web repository, we ensure that genomes are easily findable with persistent identifiers being accessible to a wide range of users across different platforms. By providing genome annotations in common file formats such as GenBank, we foster compatibility with various bioinformatic tools for targeted downstream analyses, thus reducing technical barriers, increasing efficiency and supporting the reproducibility of research results. Furthermore, we provide streamlined access to the valuable raw assemblies of Blackwell *et al.*, ensuring that these results can be used for further studies.

Conclusion

The BakRep web repository provides a consistent and comprehensive characterization of one of the largest collections of bacterial genomes comprising assembly metrics, robust taxonomic classifications, MLST subtypings, genome annotations and original metadata. Its implementation and underlying cloud infrastructure facilitate scalability and allow for swift adjustments to extended analyses and expanding datasets. Our long-term plan includes the addition of more genomes and further analyses to our repository, aiming for the continuous expansion of this standardized dataset. We envision BakRep as a high-quality open resource for microbial researchers worldwide helping to streamline targeted large-scale genome analyses.

Funding information
This work was supported by the Justus Liebig University Giessen, Germany, as well as the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Acknowledgements
We gratefully thank Frank Förster for setting up the cloud project and for all the technical support. We also would like to thank Sonja Diedrich for her code contributions and the support with image editing. We gratefully thank Grace Blackwell, John Lees and Zamin Iqbal for fruitful discussions, feedback and support.

Author contributions
O.S. designed and supervised the study. L.F. conducted data analyses, interpreted the data and wrote the manuscript. L.J. developed the website and the servers. A.G. supervised the study and was responsible for funding. All authors critically checked and contributed to the final version of the manuscript.

Conflicts of interest
The authors declare that they have no conflicts of interest.

References

1. Blaxter M, Danchin A, Savakis B, Fukami-Kobayashi K, Kurokawa K, *et al*. Reminder to deposit DNA sequences. *Science* 2016;352:780.
2. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, *et al*. The FAIR guiding principles for scientific data management and Stewardship. *Sci Data* 2016;3:160018.
3. Bagheri H, Severin AJ, Rajan H. Detecting and correcting misclassified sequences in the large-scale public databases. *Bioinformatics* 2020;36:4699–4705.
4. Keck F, Couton M, Altermatt F. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Mol Ecol Resour* 2023;23:742–755.
5. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, *et al*. Extensive error in the number of genes inferred from draft genome assemblies. *PLOS Comput Biol* 2014;10:e1003998.
6. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* 2019;20:92.
7. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study Group, *et al*. The enterobase user's guide, with case studies on salmonella transmissions, yersinia pestis phylogeny, and escherichia core genomic diversity. *Genome Res* 2020;30:138–152.
8. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, *et al*. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLOS Biol* 2021;19:e3001421.
9. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315–5316.
10. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–D794.
11. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;20:1203–1212.
12. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, *et al*. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* 2021;7:000685.
13. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, *et al*. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–319.
14. Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput Biol* 2017;13:e1005404.
15. Parviainen T. *Real-Time Web Application Development Using Vert.x 2.0*. Packt Publishing, 2013, p. 122
16. Gormley C, Tong Z. *Elasticsearch: The Definitive Guide,* 1st ed. O'Reilly Media, Inc, 2015, p. 724.
17. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. IGV.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* 2023;39:btac830.
18. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, *et al*. Reference Sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–D745.
19. Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, *et al*. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–D531.
20. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, *et al*. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genom* 2015;15:141–161.
21. Timme RE, Sanchez Leon M, Allard MW. Utilizing the public genometrakr database for foodborne pathogen traceback. *Methods Mol Biol Clifton NJ* 2019;1918:201–212.
22. Achtman M, Zhou Z, Alikhan N-F, Tyne W, Parkhill J, *et al*. Genomic diversity of *Salmonella enterica* -the UoWUCC 10K genomes project. *Wellcome Open Res* 2020;5:223.
23. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, *et al*. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;36:996–1004.
24. Garcia R, Gemperlein K, Müller R. *Minicystis rosea* gen. nov., sp. nov., a polyunsaturated fatty acid-rich and steroid-producing soil myxobacterium. *Int J Syst Evol Microbiol* 2014;64:3733–3742.
25. Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, *et al*. A genomic perspective across earth's microbiomes reveals that genome size in archaea and bacteria is linked to ecosystem type and trophic strategy. *Front Microbiol* 2021;12:761869.

26. Wenkai T, Bin L, Mengyun C, Wensheng S. Genomic legacies of ancient adaption illuminate the GC-content evolutionin bacterial genomes. *Microbiol Spectr* 2019;11:e02145-22.
27. Smits THM. The importance of genome sequence quality to microbial comparative genomics. *BMC Genom* 2019;20:662.
28. Parks DH, Chuvochina M, Reeves PR, Beatson SA, Hugenholtz P. Reclassification of *Shigella* species as later heterotypic synonyms of *Escherichia coli* in the genome taxonomy database. *Microbiology* 2021.
29. Appelt S, Rohleder A-M, Jacob D, von Buttlar H, Georgi E, *et al*. Genetic diversity and spatial distribution of *Burkholderia mallei* by core genome-based multilocus sequence typing analysis. *PLoS One* 2022;17:e0270499.
30. Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, *et al*. Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biol Evol* 2010;2:102–116.
31. Hatcher CL, Muruato LA, Torres AG. Recent advances in *Burkholderia mallei* and *B. pseudomallei* research. *Curr Trop Med Rep* 2015;2:62–69.

7.1.3. Evidence of a novel sublineage of *Streptococcus agalactiae* in elephants from zoo populations in Germany

Linda Fenske[1], Elita Jauneikaite[2,3], Maria Getino[2], Yu Wan[2,4], Alexander Goesmann[1], Tobias Eisenberg[5]

[1]Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany
[2]NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Department of Infectious Disease, Imperial College London, London, UK
[3]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK
[4]David Price Evans Global Health and Infectious Diseases Research Group, University of Liverpool, Liverpool, UK
[5]Hessian State Laboratory, Department of Veterinary Medicine, Giessen, Germany

Microbial Genomics (2025), DOI: 0.1099/mgen.0.001489

The following contributions are attributed to the thesis author:

Actively involved in planning and coordination of the study. Conducted bioinformatic data analyses and interpreted the data. Drafted the manuscript and finalized it with input from the co-authors. Corresponded with the journal.

MICROBIAL GENOMICS

RESEARCH ARTICLE

Fenske *et al.*, *Microbial Genomics* 2025;11:001489
DOI 10.1099/mgen.0.001489

Evidence of a novel sublineage of *Streptococcus agalactiae* in elephants from zoo populations in Germany

Linda Fenske[1,*], Elita Jauneikaite[2,3], Maria Getino[2], Yu Wan[2,4], Alexander Goesmann[1] and Tobias Eisenberg[5]

Abstract

Streptococcus agalactiae research primarily centres on investigating human and bovine infections, although this pathogen also can be carried and cause infections in a wider range of animal species. Moreover, infections with *S. agalactiae* are posing significant health implications, and recent studies furthermore are highlighting a potential zoonotic risk. Despite the relatively frequent isolation of *S. agalactiae* from elephants, only a few reports document infections in wild and zoo populations. We performed a comparative genomic analysis of 24 elephant isolates from three different zoos in Germany to achieve a comprehensive characterization. Elephant isolates showed pronounced phylogenetic divergence from isolates of other host species, while also forming clusters based on zoo of origin and their genotypes (MLST profiles). Capsular serotypes could not be predicted for the majority of the isolates (*n*=20/24). Several genes, exclusively associated with the elephant host, may underlie the pathogen's capacity to improve its survival and virulence across varied ecological niches. This study not only deepens our understanding of *S. agalactiae* across diverse species and environments but also represents the first whole-genome sequencing characterization of *S. agalactiae* isolates from elephants, helping to expand our knowledge about infections in animals.

Impact Statement

This study provides the first whole-genome characterization of *Streptococcus agalactiae* isolates from elephants, revealing their distinct phylogenetic divergence from group B streptococcus (GBS) isolates of other host species. Our findings suggest that elephant-derived GBS may represent a unique sublineage, with potential adaptations to their host and environment. The absence of identifiable capsular serotypes in most isolates and the presence of host-associated genes highlight the need for further research on the pathogen's evolution, virulence and zoonotic potential. Expanding genomic studies to include isolates from broader geographic regions will be crucial for understanding the role of GBS in exotic animals and its potential impact on both wildlife and public health.

DATA SUMMARY

Raw whole-genome sequencing data for elephant isolates used in this study are available in the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/browser/home) project accession number PRJEB28328, and individual GBS isolate accession numbers have been listed in Table S1, available in the online Supplementary Material. Reads and assemblies from Oxford Nanopore MinION sequencing for two GBS isolates, 161002207-5 and 161002207-6 have been deposited to ENA project accession number PRJNA1244849.

Received 31 March 2025; Accepted 03 August 2025; Published 04 September 2025
Author affiliations: [1]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, Germany; [2]NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Department of Infectious Disease, Imperial College London, London, UK; [3]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK; [4]David Price Evans Global Health and Infectious Diseases Research Group, University of Liverpool, Liverpool, UK; [5]Department of Veterinary Medicine, Hessian State Laboratory, Giessen, Germany.
***Correspondence:** Linda Fenske, linda.fenske@computational.bio.uni-giessen.de
Keywords: capsule typing; elephants; MLST; penicillin-binding proteins; prophage regions; *Streptococcus agalactiae*; virulence; whole-genome sequencing.
Abbreviations: AMR, antimicrobial resistance; GBS, group B streptococcus; MLST, multilocus sequence typing; ONT, Oxford Nanopore Technologies; PBP, penicillin-binding protein.
Two supplementary figures and three supplementary tables are available with the online version of this article.

INTRODUCTION

Streptococcus agalactiae, commonly known as group B streptococcus (GBS), is a facultative anaerobic, non-motile, chain-forming, Gram-positive, catalase-negative bacterium that can cause infections in a wide range of hosts. As a human pathogen, GBS plays a significant role as the primary cause of neonatal infections, usually transmitted from mother to newborn during childbirth, provoking conditions such as pneumonia, meningitis or septicaemia [1]. Though GBS is also known to be carried asymptomatically within the urogenital tract and the gut, it can also cause infections in adults, especially the elderly. Examples of such infections are urinary tract infections, diabetic foot infections, osteomyelitis and, in rare cases, even more severe conditions like toxic shock syndrome, necrotizing fasciitis, disseminated intravascular coagulopathy and renal dysfunction [2, 3]. GBS is also well known as a pathogen in dairy farming, where it was first reported as a causative agent for the majority of mastitis cases in 1927 [4]. The implementation of the five-point hygiene plan in the 1960s, which included measures such as rapid identification and treatment of infected cows, whole-herd antibiotic dry cow therapy, post-milking teat disinfection, culling chronically infected cows and routine disinfection of milking equipment, significantly reduced the prevalence of GBS infections but has not eliminated its relevance in the dairy industry today. Although GBS research predominantly focuses on human and bovine diseases, GBS has also been reported to cause infections in a variety of animal species including dolphins [5], camels [6], rats [7], fish, seals and other aquatic species [8, 9] and elephants [10].

One of the key virulence factors in GBS is the capsular polysaccharide, which is involved in host immune response evasion [11]. Based on the antigenic properties of the polysaccharide capsule, a serotyping system for GBS was developed [12]. Currently, ten major serotypes (Ia, Ib and II-IX) are recognized, encoded by the capsular locus (*cps*), which comprises 16–18 genes [13, 14]. It is known that the prevalence and distribution of serotypes vary across geographic regions, host species and clinical presentations [15], and some of the serotypes are also associated with different virulence potential [16, 17]. Most studies on GBS focus on human isolates, and though studies in animals have been published, there is a lack of detailed genomic description of GBS isolated from elephants. The earliest report of GBS in elephants dates back to 1997, when the bacterium was identified in pododermatitis lesions of African elephants [18], followed by its detection in Asian elephants in 2008 [19]. Since then, only one other study reported GBS isolated from African and Asian elephants in zoos in Germany [10]. The latter study employed molecular techniques to characterize and compare the identified GBS isolates from elephants to those in other animals and humans, highlighting the potential genetic diversity of GBS causing infections in elephants, including the main finding that most of the GBS from elephants were non-typable using the standard sera or capsular locus typing by multiplex PCR. The non-typable state of GBS isolates is of interest as disease-causing GBS are typically encapsulated, though some studies have reported isolates from humans lacking a capsular locus [20], and GBS isolates from animals have been reported as non-typable [21] or not clearly typable due to previously unknown serotype variants. This could be due to the multiplex PCR designed for human-specific GBS serotypes that does not account for the potential capsular loci diversity of GBS from animals [22].

In the present study, we conducted a comparative genomic analysis of 24 GBS isolates obtained from African and Asian elephants between 2010 and 2023 from zoos in Germany. We aimed to characterize in detail the genomes of these GBS isolates by determining their virulence factors, genotypes based on multilocus sequence typing (MLST), capsular serotypes and potential host-specific genes. This provides valuable insights into the evolution and pathogenicity of GBS populations in less studied host species such as elephants. To our knowledge, this is the first study characterizing GBS isolates from elephants using whole-genome sequencing (WGS).

METHODS

Bacterial isolates

A total of 24 GBS isolates from African and Asian elephants in German zoos were whole-genome sequenced and analyzed. Of these, 23 isolates were obtained during routine bacteriological investigations from elephants in two different zoos (A+B; ~200 km apart) between 2010 and 2016. Twelve of these isolates were obtained from the previous study by Eisenberg *et al*. [10], indicated in Table S1. One isolate (IHIT53690) was obtained from an elephant in Zoo C (~400 km apart from Zoo A; ~200 km apart from Zoo B) in 2023 as part of a routine investigation carried out by the Institute of Hygiene and Infectious Diseases of Animals, Giessen. Metadata of all draft genomes used in this study are provided in Table S1.

Identification by MALDI-TOF MS

Bacterial isolates were selected from the culture plates and then transferred to steel targets according to the manufacturer's instructions (Bruker Biotyper, Bruker Daltonics, Bremen, Germany). Isolates were prepared using the direct smear method and analysed by MALDI-TOF MS using Biotyper (v3.3.1.0).

WGS, raw reads processing and assembly

GBS isolates were streaked on Columbia blood agar plates (Oxoid, Basingstoke, UK) and incubated at 37 °C, 5% CO_2 overnight. Genomic DNA was extracted using the GenElute bacterial Genomic DNA kit (Sigma-Aldrich, Burlington, MA, USA) following the manufacturer's instructions for Gram-positive bacteria with modifications as follows: GBS was lyzed in 180 µl G+lysis solution with added 20 µl mutanolysin (Sigma-Aldrich, USA; prepared at 3,000 U ml^{-1}) and 20 µl lysozyme (Sigma-Aldrich, USA; prepared at 100 mg ml^{-1}) added prior to incubation at 37 °C for 1 h. Subsequent steps followed the manufacturer's instructions. DNA was quantified using a NanoDrop spectrophotometer (Thermo, Waltham, MA, USA).

For short-read sequencing, multiplexed DNA library preparation was conducted according to the Illumina protocol, and WGS was performed on a HiSeq X Ten system (Illumina, USA) with 150-cycle paired-end mode. Raw reads were trimmed and filtered with fastp (v0.23.2) [23], and reads were checked for quality with FastQC (v0.11.9) (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmed reads were used for *de novo* assembly into contiguous sequences using Unicycler (v0.5.0) [24]. The draft assemblies were purged from possible errors with Polypolish (v0.5.0) [25] and POLCA from the MaSuRCA (v4.1.0) (Maryland Super Read Cabog Assembler) genome assembly and analysis toolkit [25].

Two isolates (161002207-5, 161002207-6) were selected for long-read sequencing using an Oxford Nanopore Technologies (ONT, UK) MinION Mk1B device. Genomic DNA libraries were prepared using the Rapid Barcoding Sequencing Kit (SQK-RBK004; ONT, UK) and loaded into an R9.4.1 flow cell (FLO-MIN106; ONT, UK). Basecalling was conducted using Guppy (v6.5.7) (ONT, UK) under its super-accuracy mode. Long reads were checked for quality using Filtlong (v0.2.1) (https://github.com/rrwick/Filtlong) and assembled using Trycycler (v0.5.5) [26] followed by short-read polishing using Polypolish and pypolca (v0.3.1) [27, 28] as recommended in Ryan Wick's guide to bacterial genome assembly [29]. Polished assemblies were annotated using Bakta (v1.7.0) [30]. Contamination and completeness of all genome assemblies were estimated with CheckM2 (v1.0.1) [31], and furthermore, a taxonomic verification with GTDB-Tk (v2.2.3) was conducted [32].

Population analysis

MLSTs were assigned using mlst (v2.23.0) (https://github.com/tseemann/mlst) utilizing the PubMLST *S. agalactiae* database [33] (https://pubmlst.org/organisms/streptococcus-agalactiae). Comparative genome analyses for determination of the pan and core genome, as well as singleton genes, were performed with EDGAR (v3.2) [34]. For the creation of the phylogenetic tree, the core genes of all genomes were computed. In the following step, alignments of each core gene set are generated using MUSCLE, and the alignments are concatenated to one huge alignment. The tree was constructed with FastTree using the approximately maximum likelihood method. For visualization of the core genome phylogeny, iTOL (v7) was used [35]. For phylogenetic placement of the elephant-derived isolates, a subset of GBS genomes from different host species was selected and included for comparison. Up to 23 representative genomes for each of the different host species were selected from a database of confirmed GBS genomes [36]. These included genomes from rats (n=5), dogs (n=4), dolphins (n=1), fish (n=23), frogs (n=2), seals (n=4), camels (n=16), bovines (n=17) and humans (n=23) (Table S1). To validate the findings of the phylogenetic analysis, a target-free split k-mer analysis and single-linkage clustering were conducted using SKA (v1.0) [37]. Specifically, k-mer files (k=15) were generated with the fasta subcommand under default parameters; pairwise SNP distances were calculated, and SKA clusters were defined at a threshold of 10 SNPs, if they met the minimum identity cutoff of 0.9.

Antimicrobial resistance genes and mobile genetic elements

Antimicrobial resistance (AMR) genes were detected using AMRFinderPlus [38], the Resistance Gene Identifier software [39] and abriTAMR (https://github.com/MDU-PHL/abritamr). Putative resistance markers to β-lactam antibiotics were analysed using EDGAR. Therefore, orthologous genes encoding the penicillin-binding proteins PBP1A, PBP1B, PBP2A, PBP2B and PBP2X were identified in the draft genomes. The amino acid sequences of these penicillin-binding protein (PBP) genes were visualized and manually arranged in Jalview [40] based on sequence similarities and subsequently compared to the penicillin-susceptible strain 2603 V/R (GenBank: NC_004116). For reconstruction and typing of potential plasmids, MOB-suite was used (v3.1.8) [41], and the PHASTEST web tool was used to identify prophage regions within the draft genomes [42]. For 12 isolates, capsular serotyping and antimicrobial susceptibility testing results from Eisenberg *et al*. [10] were incorporated in Table S2.

Analysis of the capsular locus genes

A combination of GBS-SBG [43], srst2 (v0.2.0) [44], seq_typing (v2.3.0) (https://github.com/B-UMMI/seq_typing) and KMA (v1.4.12) [45] was used to try to determine the capsular serotype based on the capsular loci genes present. As the majority of isolates remained untypable despite utilizing various tools, an in-depth analysis of the *cps* locus was conducted. The *cps* loci of the elephant-derived GBS isolates were extracted with *in silico* PCR utilizing SnapGene (v8.0.2) (https://www.snapgene.com/). The primers used were introduced in a previous study [22]. If no binding sites for these primers were detected within the genomes, alternative primers flanking the *cps* locus at more distant regions were used [20].

Genome-wide association studies

A gene-enrichment analysis, focused on GBS genomes from elephants, was performed using Scoary (v1.6.16) [46] from the gene presence/absence matrix generated by Panaroo (v1.5.0) [47]. The latter was generated from the GFF3 files annotated with Bakta. The host groups were defined as binary phenotypes with a value of 1 assigned to genomes from the elephant host group and 0 to those GBS genomes from all other species.

RESULTS

Sequencing statistics

All 24 genomes were classified as *Streptococcus agalactiae* according to the GTDB taxonomy. The estimated completeness of all draft genomes was above 99.99% with a contamination rate lower than 0.07%. The combined lengths of the assembled contigs range from 1,863,022 to 2,041,196 bp with a G+C content between 35.4 and 35.5 mol%. To evaluate how the genome size and G+C content of the elephant isolates used here integrate into the overall context of GBS genomes, we compared the elephant draft genomes to all GBS genomes currently included in the bacterial web repository BakRep [48]. The BakRep v1 contains 10,359 genomes classified as *S. agalactiae* (as of October 2024). After filtering available genomes for completeness of >95% and contamination rate <1%, 9,984 genomes remained for comparison. As none of the genomes included in the repository listed the elephant as the host species, no further subdivision of the data was made. The GBS genomes included in BakRep had a genome size ranging from 1,798,114 to 2,667,783 bp (only two genomes were larger than 2.5 Mbp, i.e. 2.53 and 2.67 Mbp, respectively, in the whole collection) with a G+C content ranging from 34.9 to 40.3 mol% (Fig. 1). This highlights that the GBS draft genomes from elephants have a mean genome size of 1,930,136 (SD±515,960.30) that is below the genomes contained in BakRep with a mean genome size of 2,067,641 (SD±6,982.26).

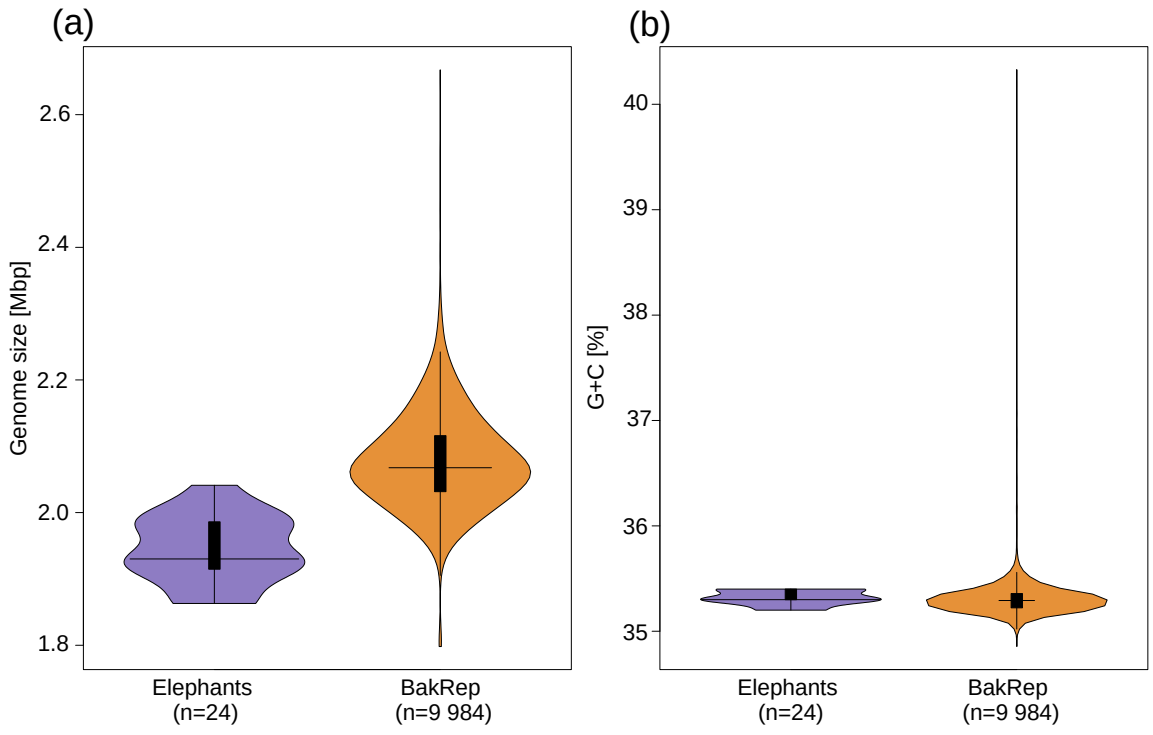


Fig. 1. Comparison between genome size (a) and G+C content (b). GBS isolates from elephants used in this study (*n*=24, purple) and all GBS genomes included in BakRep (*n*=9,984, orange). Horizontal black lines in the middle indicate medians; bold black bars represent interquartile ranges; vertical black lines represent outliers.

Characteristics of GBS genomes from elephants: genotypes, capsular types and AMR

Out of 24 GBS isolates from elephants, 13 (54.17%) were assigned to ST2019, 10 isolates were assigned to sequence type ST2304 and 1 isolate was assigned to ST2305. The latter two STs were first reported in this study and submitted to the *S. agalactiae* MLST database (Table 1). Interestingly, all ST2019 GBS isolates (*n*=13) came from Zoo A, all ST2304 GBS (*n*=11) came from Zoo B and the one GBS ST2305 isolate came from Zoo C; of note, ST2305 was a single-locus variant (SLV) of ST2304. None of the identified STs could be assigned to a clonal complex.

We knew from a previous study [10] that 12 of our elephant-derived GBS isolates were reported as non-typable for their capsular serotype using molecular capsular typing methods. The remaining 12 isolates were not routinely serotyped in the laboratory. After WGS, 4 out of 24 (16.7%) GBS isolates were assigned to serotype Ia. For all other isolates, no clear assignment could be made even after using various accepted GBS capsular typing methods (Table 2).

To analyse the *cps* locus in greater detail, *cps* loci of all elephant-derived isolates were extracted. Binding sites for the primer sequences used by Crestani *et al.* [22] could only be found in the isolates 151006965-2, 151010127, 161002207-6, 161002208-5 and IHIT53690. All these isolates, except IHIT53690, matched the serotype Ia reference sequence (GenBank: LT671983.1) with high similarity, aligning with the results from other applied methods. The region extracted using primer sequences from Creti *et al.* [20] exhibited a deletion in the remaining 19 isolates, similar to the one previously reported by Creti *et al.* (Fig. 2). For IHIT53690, no clear serotype assignment could be made due to several gaps and mismatches (Fig. S2).

Analysis of AMR determinants showed that all 24 GBS isolates likely remain susceptible to most antibiotics due to no acquired resistance genes detected in the genomes, bar *mprF* gene. In PBP1A, all isolates shared a V742A substitution and a deletion of four amino acid residues (739–742) previously described but not associated with reduced susceptibility to β-lactams [49]. In PBP1B, all isolates had an A95D substitution, already detected in GBS from Japan associated with reduced penicillin resistance [50], and in 13 out of 24 (54.17%) isolates, a V64I substitution was identified (not previously described). In the PBP2A, 13 out of 24 (54.17%) isolates had the E18K substitution, and 10 out of 24 (41.76%) had the S394G substitution (not previously described). All isolates showed the V80A substitution in the PBP2B, which was previously detected in other studies [49–51], while 13 out of 14 (54.17%) isolates had a D572N substitution in the PBP2X gene (not previously described). Notably, the 13 isolates with the E18K substitution in PBP2A and the D572 substitution in PBP2X all originated from Zoo B, while the 10 isolates with the S394G substitution came from Zoo A. However, phenotypic testing done in the previous study by Eisenberg *et al.* [10] confirmed high susceptibility to penicillin G and other β-lactams, including those with identified PBP mutations. Additionally, 16 of the 24 tested isolates exhibited low-level phenotypic resistance to gentamicin (Table S2). All substitutions and results of previously done antimicrobial susceptibility testing are listed in Table S2.

Table 1. MLST results for 24 isolates of *S. agalactiae* from elephants analyzed in this study utilizing the MLST scheme of Jones *et al.* [74]. The newly identified STs in this study are highlighted in bold

Isolate	Allele							ST
	adhP	*pheS*	*atr*	*glnA*	*sdhA*	*glcK*	*tkt*	
10-7-D-02041; 141000096/2; 141014875; 151001913-2; 151002450-1; 151002450-2; 151003337-1; 151003337-2; 151003441-1; 151003441-2; 151003441-3; 151008126; 151010216	20	13	15	1	15	9	5	2019
161002207-5; 161002207-6; 161002208-5; 151004802; 151005628-1; 151005628-2; 151006965-1; 151006965-2; 151006967; 151010127	16	17	4	2	4	9	5	**2304**
IHIT53690	16	23	4	2	4	9	5	**2305**

Table 2. Serotyping results for 24 isolates of *S. agalactiae* from elephants obtained with different *cps* typing tools

Isolate	Tool		
	GBS-SBG (best match)	Seq_typing (best match)	KMA (best match)
161002207-5	NT	NT	VI (score: 15264; coverage: 1.30%)
161002207-6	Ia (coverage: 99.98%)	Ia (sequence covered: 95.74%)	Ia (score: 1111688; coverage: 85.85%)
161002208-5	III-3 (coverage: 78.85%)	Ia (sequence covered: 97.03%)	Ia (score: 1305804; coverage: 87.78%)
10-7-D-02041	NT	NT	VI (score: 12116; coverage: 1.28%)
141000096/2	NT	NT	VI (score: 13035; coverage: 1.27%)
141014875	NT	NT	VI (score: 13505; coverage: 1.28%)
151001913-2	NT	NT	II (score: 12165; coverage: 2.35%)
151002450-1	NT	NT	VI (score: 15747; coverage: 1.29%)
151002450-2	NT	NT	VI (score: 19190; coverage: 1.28%)
151003337-1	NT	NT	VI (score: 14669; coverage: 3.12%)
151003337-2	NT	NT	VI (score: 12245; coverage: 1.26%)
151003441-1	NT	NT	VI (score: 16389; coverage: 1.29%)
151003441-2	NT	NT	VI (score: 11571; coverage: 1.29%)
151003441-3	NT	NT	VI (score: 16338; coverage: 1.34%)
151004802	NT	NT	VI (score: 14957; coverage: 1.29%)
151005628-1	NT	NT	VI (score: 17042; coverage: 1.29%)
151005628-2	NT	NT	VI (score: 15494; coverage: 1.34%)
151006965-1	NT	NT	VI (score: 13236; coverage: 1.29%)
151006965-2	Ia (coverage: 99.99%)	Ia (sequence covered: 97.03%)	Ia (score: 1127495; coverage: 85.73%)
151006967	NT	NT	VI (score: 12073; coverage: 1.27%)
151008126	NT	NT	VI (score: 18202; coverage: 1.29%)
151010127	Ia (coverage: 100%)	Ia (sequence covered: 96.65%)	Ia (score: 1139070; coverage: 86.66%)
151010216	NT	NT	VI (score: 15068; coverage: 1.27%)
IHIT53690	VII (coverage: 56.55%)	IX (sequence covered: 65.85%)	IV (score: 995859; coverage: 80.59%)

Mobile genetic elements: prophages identified in GBS isolates

There were neither complete nor partial plasmid sequences identified in any of the 24 isolates. The prophage analysis suggested that 13 out of 24 (54.17%) isolates had predicted intact phages: 7 out of 24 (29.17%) isolates had *Streptococcus* phage T12 (NC_028700), 2 out of 24 (8.33%) isolates had *Streptococcus* prophage 315.1 (NC_004584.1) and 1 out of 24 (4.17%) had an intact *Staphylococcus* phage SPbeta-like (NC_029119.1), while another isolate carried a questionable version of the same phage. Two out of twenty-four (8.33%) isolates had Bacteriophage Shelly (NC_041909.1), 1 out of 24 (4.17%) had *Bacillus* phage G (NC_023719.1) and 1 out of 24 (4.17%) isolates had incomplete *Streptococcus* phage 5093 (NC_012753.1) (Fig. 3).

Pan-genome analysis and phylogenetic analysis of GBS genomes from elephants

Analysis of the gene content of the 24 elephant-derived GBS isolates revealed a pan-genome comprising 2,352 genes, with a core genome of 1,629 genes. When the representative GBS genomes from other host species (as described in the 'Method' section) were included in the analysis, the pan-genome expanded to 5,485 genes, while the core genome decreased to 1,325 genes. Within the phylogenetic analysis, GBS isolates clustered into two broader clusters: one comprising lineages previously reported only in a single host species, such as camel-associated CC609, bovine-associated CC67 and the elephant sequence types identified in this study, and another cluster comprising lineages reported in multiple host species, including human-associated CC23, fish-associated CC7 and rat-associated CC10, and similar consistent with prior findings (Fig. 4) [52]. To further explore the genetic relationships among GBS isolates from elephants, a phylogenetic analysis was conducted exclusively on the 24 elephant-derived isolates (Fig. 5).

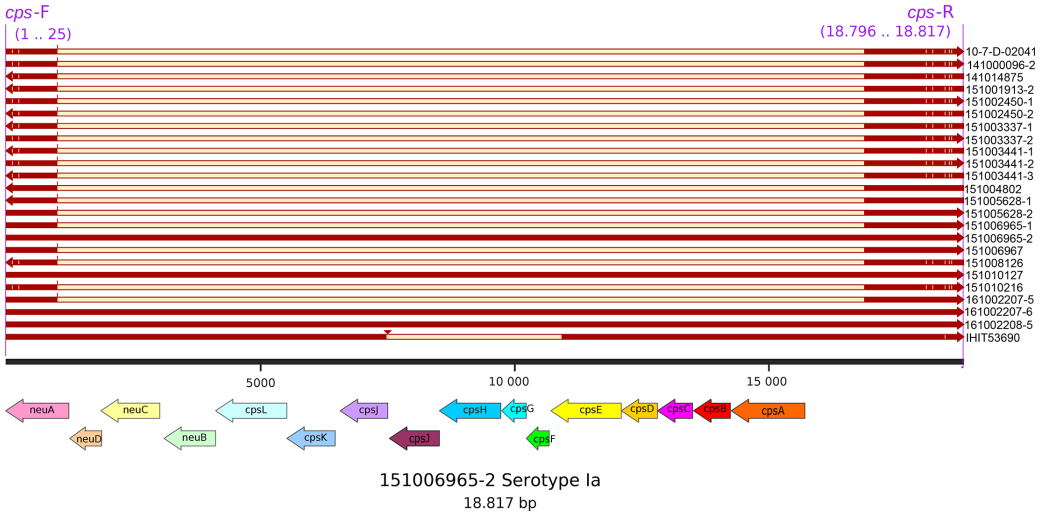


Fig. 2. Multiple-sequence alignment of the extracted *cps* locus of the elephant-derived GBS isolates. The regions between the primers (*cps*-F and *cps*-R) are indicated in dark red, while the deletion is shown in beige. Coloured arrows at the bottom depict the genetic organization of the serotype Ia gene sequence from isolate 151006965-1, which was used as reference. Primer binding sites are indicated at the top in purple. The figure was created with SnapGene (v8.0.2), and annotation of the genes was performed with Bakta (v1.7.0).

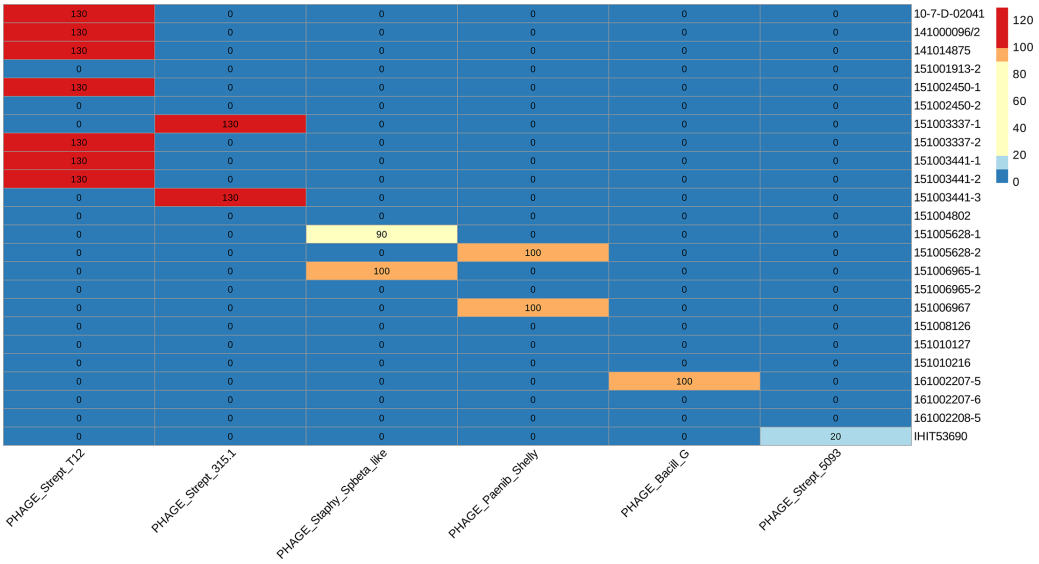


Fig. 3. Heatmap of the identified prophage regions. The score (<70: incomplete; 70–90: questionable; >90: intact) for each isolate is indicated in the boxes. The heatmap was created with R from the results generated with PHASTEST.

The isolates generally clustered by geographical origin, corresponding to the zoo from which the host was sampled. An exception was the most recent isolate from an African elephant in Zoo C (collected in 2023), which shared a common ancestor with four GBS isolates from Asian elephants in Zoo A. Notably, a clear clustering pattern emerged based on both the zoo of origin and the collection year. In Zoo B, all isolates were from African elephants and shared the same sequence type (ST2019). These isolates lacked an identifiable capsular serotype and were primarily isolated in 2015. However, this clade also included two isolates from

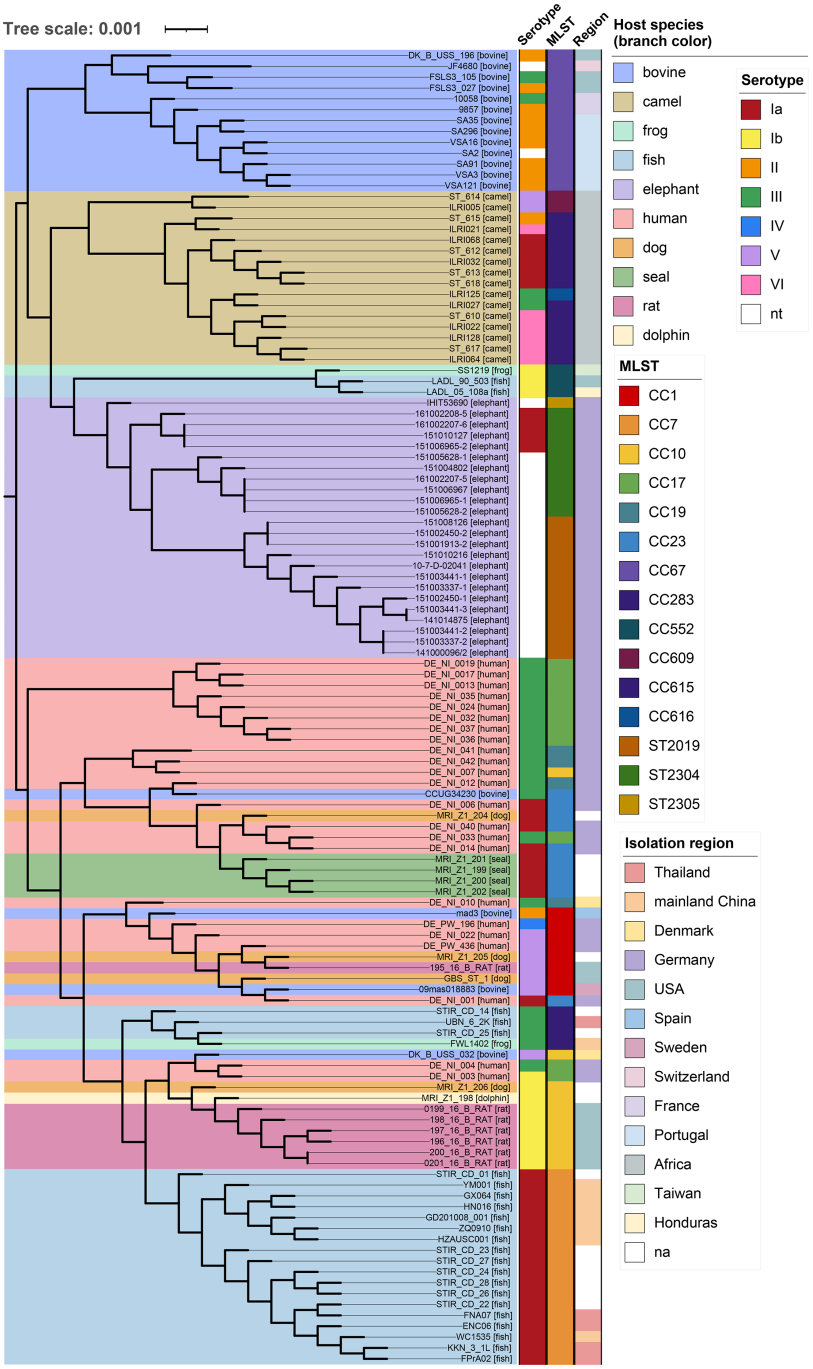


Fig. 4. Phylogenetic analysis of different GBS isolates (*n*=121). The tree is colour-coded according to the respective host species, with the host species also indicated in square brackets. The scale bar corresponds to the number of amino acid substitutions per site, with branch lengths reflecting the relative genetic divergence among genomes. Annotation of the tree was done using iTOL.

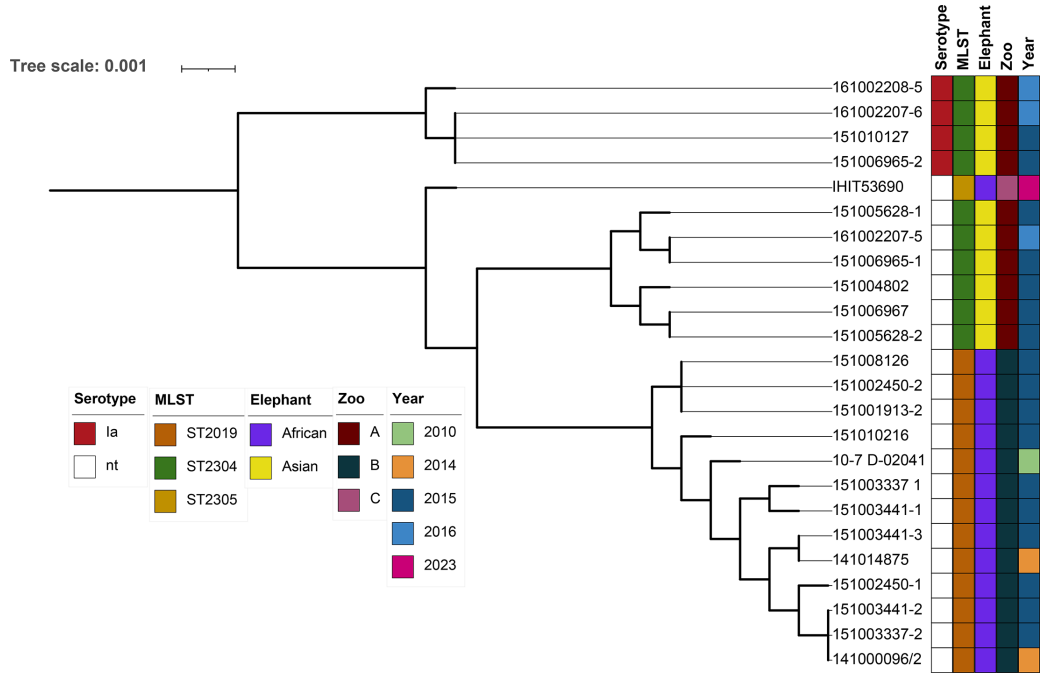


Fig. 5. Core genome phylogenetic tree of the 24 elephant isolates. Coloured blocks showing capsular serotype, MLST, elephant species, zoo of origin and isolation year, as indicated by the legend. The scale bar corresponds to the number of amino acid substitutions per site, with branch lengths reflecting the relative genetic divergence among genomes. The tree was annotated using iTOL.

2014 and one from 2010 that clustered very closely together, suggesting potential persistence of this GBS strain within the zoo population over multiple years.

With regard to Zoo A, isolates 161002208-5, 151010127, 151006965-2 and 161002207-6 formed a distinct clade separate from the other Zoo A isolates. Notably, these were the only isolates with an identifiable capsular serotype, all classified as serotype Ia. As expected, major clades aligned with MLST types: ST2019 isolates clustered together, while ST2304 clustered with the closely related ST2305 isolate, IHIT53690, as ST2305 is an SLV of ST2304 (Fig. 5).

Comparison dataset cluster attribution and genome-wide association investigation

SKA identified a total of 96 clusters among all GBS genomes (n=121), with the elephant-derived isolates distributed across six distinct clusters (Table S3). However, the elephant isolates did not cluster with any isolates from other host species (Fig. S1). The median SKA SNP distance was 7,147.5 SNPs across all GBS isolates and 2,344.4 SNPs specifically for the elephant-derived isolates. Within Zoo A, the median SNP distance was 182.5 SNPs, while in Zoo B, it was 5.2 SNPs.

A genome-wide association study was conducted using a pan genome approach to investigate genes that are unique to GBS isolated from elephants. We have found that the *maR* gene was significantly associated with the elephant host, found in all elephant-derived isolates but not in any of the isolates from the other host species. Two additional proteins uniquely associated with all elephant isolates were the Maf family protein and the three acyl-CoA thioester hydrolase/BAAT C-terminal domain-containing proteins. The Lin0465 protein was found in all but one of the elephant isolates (IHIT53690) and was absent in all but two isolates from other host species (P-value: 8.17×10^{-15}).

DISCUSSION

S. agalactiae is a known multi-host pathogen, primarily associated with infections in humans and bovines. However, it is far more frequently isolated as a harmless colonizer, with many strains representing asymptomatic carriage rather than disease. Despite this, GBS is also relatively frequently isolated in endangered species such as wild and zoo elephants. However, to date, no studies have investigated these GBS isolates from elephants using WGS.

The phylogenetic analysis demonstrates that the core genome of GBS isolates from elephants is distinct from GBS genomes isolated from different host species, supporting the niche specificity and potentially host-restricted lineage of GBS [52, 53].

There were challenges in defining the elephant capsule type. In our study, serotype Ia was defined only in four GBS isolates, with the rest of the isolates classed as non-typable due to a deletion of multiple genes in the capsular locus. Recently, it has been reported that a human GBS isolate had deletions of multiple capsular genes [20], the same as identified in our study. Authors of the mentioned study have proposed that a recombination event led to the loss of the whole capsule locus, and large capsular recombination events are known to happen in the GBS population [54, 55]. Given that elephants are a rarely reported host for GBS, as indicated by the limited studies available and, especially, only a small number of elephant GBS genomes available, further genetic and experimental data would be required to confirm if elephant GBS isolates are more likely to be acapsular as part of their species-specific adaptations. It is important to note that GBS isolates investigated here came from infections in elephants, and carriage isolates are not sought after; hence, we are not able to comment on the potential diversity of disease-causing vs asymptomatic or carriage strains of GBS in elephant populations. Notably, current development of GBS vaccines is primarily focused on human GBS vaccines based on ten capsular serotypes or specific surface proteins [56], and current vaccines for fish are also based on serotypes [57]. There is a concern that serotype-based vaccines, when implemented widely, will drive the evolution of non-vaccine serotype and acapsular GBS clones to emerge, similar to what has happened to pneumococcal clones [58]. The risk of this happening may increase if there is horizontal transfer of capsular operons between strains or even across different host species that occur more frequently [22, 59].

Eleven of 24 GBS isolates from elephants had a novel ST2304 (n=10) and ST2305 (n=1), highlighting the genetic diversity in the GBS isolated from animals that is still unexplored. Notably, isolate IHIT53690 from Zoo C differed only by one allele variant from the isolates originating from Zoo B. The remaining GBS isolates from elephants were classified as ST2019 and were found only in Zoo B. Within the PubMLST database, only one other isolate of ST2019 has been reported, which was also reported as non-typable for its capsular serotype and has been isolated from elephants in the UK (personal communication, Dr Jauneikaite, manuscript in preparation). One of the limitations of our study is that there is potential confounding between host species and geographical location where the elephants were present at the time of sampling, as GBS isolates were available only from one species of elephants present in each of the zoo, potentially highlighting the genotype specificity linked to either elephant species or the zoo. Only one elephant species was present, and it is closely associated with the zoo of origin. Consequently, it is not possible to determine whether the observed STs are influenced primarily by host species or geographic and environmental factors. Furthermore, the dataset spanning over a time period of 13 years with irregular sampling intervals adds another layer of complexity to the interpretation of phylogenetic relationships and strain persistence. These limitations should be taken into account when interpreting findings related to host specificity, transmission patterns and the evolutionary trend of GBS in elephants and other species.

We have found the *mprF* gene present in all elephant-derived GBS genomes. This gene has been reported to play a role in bacterial resistance to cationic antimicrobial peptides, including agents like gentamicin or daptomycin, the latter classified as an antibiotic of last resort used in cases of multidrug-resistant infections [60]. It is known that GBS displays intrinsic resistance to gentamicin due to the low permeability of its cell wall to large molecules such as aminoglycosides [61]. The presence of the *mprF* gene may further contribute to reduced susceptibility, as it has been reported to affect the bacterial cell membrane and decrease susceptibility to certain antibiotic classes. In *Staphylococcus aureus*, for example, *mprF* is known to modulate susceptibility to cationic antibiotics, including the glycopeptide vancomycin, the aminoglycoside gentamicin and moenomycin [62, 63].

PBPs, the enzymes essential for the synthesis of peptidoglycan as components of the cell wall in Gram-positive bacteria, have also been investigated in these isolates and found no mutations that had previously been reported to be linked with decreased susceptibility to penicillin or β-lactam antibiotics. Previous reports indicate that there are specific point mutations in PBPs, primarily PBP2x, that are linked to reduced susceptibility to β-lactam antibiotics [49, 64]. We have found the A95D mutation in PBP1B and the V80A mutation in PBP2B in all elephant isolates, which has been previously reported, although they have only been tentatively linked to penicillin-non-susceptible GBS [49, 50, 64]. None of the other substitutions in PBPs identified in this study have been associated with β-lactam resistance to date, and from limited phenotypic susceptibility data available to us, we were able to show that 12 GBS isolates investigated in our study did not show any evidence for resistance to penicillin [10].

Bacteriophages, including prophages, play a crucial role in the evolution of bacterial genomes and are often associated with enhanced infectivity, pathogenicity and virulence across various bacterial species [65, 66]. Seven GBS isolates in our study carried intact prophage regions with most gene hit counts for *Streptococcus* phage T12. This phage is a prototypic temperate phage first associated with *Streptococcus pyogenes* that carries the *speA* gene coding for erythrogenic toxin A [67]. However, there are no reports of phage T12 in GBS, and none of the elephant isolates in the pan-genome analysis contained the *speA* gene or any related toxin.

When examining the presence of host-specific genes, only one gene was consistently found across all elephant GBS isolates in comparison to other GBS isolates. This gene encodes the MprF enzyme, responsible for the unique synthesis of a cationic glycolipid, Lys-Glc-DAG, which aids in the invasion of human endothelial cells, suggesting a potential role in enhancing

bacterial entry into host cells and promoting disease progression [68]. However, it remains to be confirmed whether it fulfils the same function in elephants. The *maR* gene found exclusively in all elephant isolates encodes the MarR family transcriptional regulator, which plays a key role in controlling the oxidative stress response, a vital environmental sensing function, particularly for pathogenic bacteria [69]. A recent study postulated that the broad host range and ability of GBS to colonize different tissues may be due to the ability of its regulatory systems to recognize and respond to external stimuli, including oxidative and aerobic stress [70]. Maf-like proteins have been proposed to belong to a family of house-cleaning nucleotide hydrolyzing enzymes, which prevent the incorporation of noncanonical nucleotides into cellular DNA [71]. This could also offer an advantage in adapting to different hosts and environmental conditions. Lin0465, which encodes an intracellular PfpI protease in *Listeria*, plays a role in stress response [72, 73]. A study postulated that lin0464, lin0465 and their homologues contribute more to the environmental fitness of these bacteria rather than to their virulence [73]. In summary, this set of genes unique to the elephant isolates examined here collectively confers advantages that enhance adaptability to various hosts and environmental conditions. By enabling GBS to respond effectively to diverse physiological and ecological challenges, these genes likely play a role in its ability to adapt to a host like the elephant and its survivability in different environments.

Conclusion

GBS plays an important role in bacterial infections not only in human and bovine diseases, but also in less often studied host species like elephants. Comparative genomic analysis revealed that, in several aspects, the elephant GBS isolates differ from those of other host species, despite the geographical limitations of our study. Capsular typing of the elephant-derived isolates revealed a high proportion of non-typable strains due to a deletion in the *cps* locus, highlighting their genetic differentiation from serotypes predominantly characterized in human GBS isolates. To further clarify whether elephants may represent a distinct sublineage of GBS, future studies should include more isolates from elephants across diverse geographic regions. All of this could contribute to a better understanding of the zoonotic potential and pathogenic properties of GBS.

Funding information
This work was funded by The Rosetrees Trust and The Stoneygate Trust Imperial College Research Fellowship (M683) awarded to Dr Elita Jauneikaite. This work was also supported by the Justus Liebig University Giessen, Germany, as well as the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (FKZ 031A532–031A540, W-de.NBI-010). Furthermore, the Hessian Ministry of Agriculture and Environment, Viticulture, Forestry, Hunting and Homeland supports the Hessian State Laboratory.

Acknowledgements
E.J., M.G. and Y.W. are affiliated with the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Imperial College London in partnership with the UK Health Security Agency, in collaboration with Imperial Healthcare Partners, University of Cambridge and University of Warwick. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or the UK Health Security Agency. Y.W. is a research fellow funded by the David Price Evans endowment to the University of Liverpool. Microbiological work and sample preparation for whole-genome sequencing of GBS isolates were undertaken at the Colebrook Laboratory, a facility supported by the NIHR Imperial Biomedical Research Centre (BRC). Part of the bioinformatics analysis was performed on equipment purchased as part of the MRC CARP fellowship award MR/T005254/1. We would also like to thank Sophie Aurich from the Institute of Hygiene and Infectious Diseases of Animals, Giessen, for providing isolate IHIT53690.

Author contributions
T.E. and E.J. designed and supervised the study. E.J., M.G. and Y.W. performed the WGS and analyzed the data. L.F. conducted bioinformatic analysis and interpreted the data. Bioinformatic analyses were supervised by A.G. L.F. wrote the first draft of the manuscript. L.F. and E.J. wrote the next versions of the manuscript draft. A.G. and E.J. acquired funding for the study. All authors critically checked and contributed to the final version of the manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Le Doare K, Heath PT. An overview of global GBS epidemiology. *Vaccine* 2013;31 Suppl 4:D7–12.
2. Farley MM, Strasbaugh LJ. Group B streptococcal disease in nonpregnant adults. *Clin Infect Dis* 2001;33:556–561.
3. Sendi P, Johansson L, Norrby-Teglund A. Invasive group B streptococcal disease in non-pregnant adults. *Infection* 2008;36:100–111.
4. Ruegg PL. A 100-year review: mastitis detection, management, and prevention. *J Dairy Sci* 2017;100:10381–10397.
5. Evans JJ, Pasnik DJ, Klesius PH, Al-Ablani S. First report of *Streptococcus agalactiae* and *Lactococcus garvieae* from a wild bottlenose dolphin (*Tursiops truncatus*). *J Wildl Dis* 2006;42:561–569.
6. Seligsohn D. Pure white gold: subclinical mastitis in dairy camels in Kenya with a special focus on *Streptococcus agalactiae*. Acta Universitatis Agriculturae Sueciae, Swedish University of Agricultural Sciences 2021. https://pub.epsilon.slu.se/23725/ [accessed 8 September 2021].
7. Shuster KA, Hish GA, Selles LA, Chowdhury MA, Wiggins RC, *et al*. Naturally occurring disseminated group b streptococcus infections in postnatal rats. *Comp Med Februar* 2013;63:55–61.
8. Delannoy CMJ, Crumlish M, Fontaine MC, Pollock J, Foster G, *et al*. Human *Streptococcus agalactiae* strains in aquatic mammals and fish. *BMC Microbiol* 2013;13:41.
9. Numberger D, Siebert U, Fulde M, Valentin-Weigand P. Streptococcal infections in marine mammals. *Microorganisms* 2021;9:350.
10. Eisenberg T, Rau J, Westerhüs U, Knauf-Witzens T, Fawzy A, *et al*. Streptococcus agalactiae in elephants - a comparative study with

isolates from human and zoo animal and livestock origin. *Vet Microbiol* 2017;204:141–150.

11. Avci FY, Kasper DL. How bacterial carbohydrates influence the adaptive immune system. *Annu Rev Immunol* 2010;28:107–130.
12. Wilkinson HW, Moody MD. Serological relationships of type I antigens of group B streptococci. *J Bacteriol* 1969;97:629–634.
13. Cieslewicz MJ, Chaffin D, Glusman G, Kasper D, Madan A, *et al*. Structural and genetic diversity of group B streptococcus capsular polysaccharides. *Infect Immun* 2005;73:3096–3103.
14. Kapatai G, Patel D, Efstratiou A, Chalker VJ. Comparison of molecular serotyping approaches of *Streptococcus agalactiae* from genomic sequences. *BMC Genomics* 2017;18:429.
15. Bianchi-Jassir F, Paul P, To K-N, Carreras-Abad C, Seale AC, *et al*. Systematic review of Group B Streptococcal capsular types, sequence types and surface proteins as potential vaccine candidates. *Vaccine* 2020;38:6682–6694.
16. Burcham LR, Spencer BL, Keeler LR, Runft DL, Patras KA, *et al*. Determinants of Group B streptococcal virulence potential amongst vaginal clinical isolates from pregnant women. *PLoS One* 2019;14:e0226699.
17. Liu Y, Liu J. Group B streptococcus: virulence factors and pathogenic mechanism. *Microorganisms* 2022;10:2483.
18. Keet DF, Grobler DG, Raath JP, Gouws J, Carstens J, *et al*. Ulcerative pododermatitis in free-ranging African elephant (Loxodonta africana) in the Kruger National Park; 1997. https://repository.up.ac.za/handle/2263/20739 [accessed 25 July 2024].
19. Aupperle H, Reischauer A, Bach F, Hildebrandt T, Göritz F, *et al*. Chronic endometritis in an Asian elephant (*Elephas maximus*). *J Zoo Wildl Med* 2008;39:107–110.
20. Creti R, Imperi M, Pataracchia M, Alfarone G, Recchia S, *et al*. Identification and molecular characterization of a *S. agalactiae* strain lacking the capsular locus. *Eur J Clin Microbiol Infect Dis* 2012;31:233–235.
21. Sørensen UBS, Klaas IC, Boes J, Farre M. The distribution of clones of *Streptococcus agalactiae* (group B streptococci) among herdspersons and dairy cows demonstrates lack of host specificity for some lineages. *Vet Microbiol* 2019;235:71–79.
22. Crestani C, Seligsohn D, Forde TL, Zadoks RN. How GBS got its hump: genomic analysis of group B *Streptococcus* from camels identifies host restriction as well as mobile genetic elements shared across hosts and pathogens. *Pathogens* 2022;11:1025.
23. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
24. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* 2017;13:e1005595.
25. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, *et al*. The MaSuRCA genome assembler. *Bioinformatics* 2013;29:2669–2677.
26. Wick RR, Judd LM, Cerdeira LT, Hawkey J, Méric G, *et al*. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biol* 2021;22:266.
27. Wick RR, Holt KE. Polypolish: Short-read polishing of long-read bacterial genome assemblies. *PLoS Comput Biol* 2022;18:e1009802.
28. Bouras G, Judd LM, Edwards RA, Vreugde S, Stinear TP, *et al*. How low can you go? Short-read polishing of Oxford Nanopore bacterial genome assemblies. *Microbial Genomics* 2024;10:001254.
29. Wick RR. Choose-your-own-adventure guide to bacterial genome assembly. Zenodo. 2021. https://zenodo.org/records/7471199 [accessed 25 July 2024].
30. Schwengers O, Jelonek L, Dieckmann M, Beyvers S, Blom J, *et al*. Bakta: rapid & standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genomics* 2021.
31. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;20:1203–1212.
32. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315–5316.
33. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
34. Dieckmann MA, Beyvers S, Nkouamedjo-Fankep RC, Hanel PHG, Jelonek L, *et al*. EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. *Nucleic Acids Res* 2021;49:W185–W192.
35. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49:W293–W296.
36. Rothen J, Sapugahawatte DN, Li C, Lo N, Vogel G, *et al*. A simple, rapid typing method for *Streptococcus agalactiae* based on ribosomal subunit proteins by MALDI-TOF MS. *Sci Rep* 2020;10:8788.
37. Harris SR. SKA: Split Kmer analysis toolkit for bacterial genomic epidemiology. *Genomics* 2018.
38. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, *et al*. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 2019;63:e00483-19.
39. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, *et al*. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2023;51:D690–D699.
40. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;25:1189–1191.
41. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics* 2018;4:e000206.
42. Wishart DS, Han S, Saha S, Oler E, Peters H, *et al*. PHASTEST: faster than PHASTER, better than PHAST. *Nucleic Acids Res* 2023;51:W443–W450.
43. Tiruvayipati S, Tang WY, Barkham TMS, Chen SL. GBS-SBG - GBS serotyping by genome sequencing. *Microb Genom* 2021;7:000688.
44. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, *et al*. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014;6:90.
45. Clausen P, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* 2018;19:307.
46. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* 2016;17:238.
47. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.
48. Fenske L, Jelonek L, Goesmann A, Schwengers O. BakRep – a searchable large-scale web repository for bacterial genomes, characterizations and metadata. *Microb Genomics* 2024;10.
49. van der Linden M, Mamede R, Levina N, Helwig P, Vila-Cerqueira P, *et al*. Heterogeneity of penicillin-non-susceptible group B streptococci isolated from a single patient in Germany. *J Antimicrob Chemother* 2020;75:296–299.
50. Li C, Sapugahawatte DN, Yang Y, Wong KT, Lo NWS, *et al*. Multidrug-resistant *Streptococcus agalactiae* strains found in human and fish with high penicillin and cefotaxime non-susceptibilities. *Microorganisms* 2020;8:1055.
51. Nagano N, Nagano Y, Kimura K, Tamai K, Yanagisawa H, *et al*. Genetic heterogeneity in pbp genes among clinically isolated group B Streptococci with reduced penicillin susceptibility. *Antimicrob Agents Chemother* 2008;52:4258–4267.

52. Crestani C, Forde TL, Bell J, Lycett SJ, Oliveira LMA, *et al*. Genomic and functional determinants of host spectrum in Group B Streptococcus. *PLOS Pathog* 2024;20:e1012400.
53. Gori A, Harrison OB, Mlia E, Nishihara Y, Chan JM, *et al*. Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific genes associated with virulence and niche adaptation. *mBio* 2020;11:00728–20.
54. Neemuchwala A, Teatero S, Athey TBT, McGeer A, Fittipaldi N. Capsular switching and other large-scale recombination events in invasive sequence type 1 group b streptococcus. *Emerg Infect Dis* 2016;22.
55. Khan UB, Jauneikaite E, Andrews R, Chalker VJ, Spiller OB. Identifying large-scale recombination and capsular switching events in *Streptococcus agalactiae* strains causing disease in adults in the UK between 2014 and 2015. *Microbial Genomics* 2022;8:000783.
56. Pena JMS, Lannes-Costa PS, Nagao PE. Vaccines for *Streptococcus agalactiae*: current status and future perspectives. *Front Immunol* 2024;15.
57. Wong KY, Megat Mazhar Khair MH, Song AAL, Masarudin MJ, Loh JY, *et al*. Recombinant lactococcal-based oral vaccine for protection against Streptococcus agalactiae infections in tilapia (*Oreochromis niloticus*). *Fish Shellfish Immunol* 2024;149:109572.
58. Brueggemann AB, Pai R, Crook DW, Beall B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLOS Pathog* 2007;3:e168.
59. Palmeiro JK, De Carvalho NS, Botelho ACN, Fracalanzza SEL, Madeira HMF, *et al*. Maternal group B streptococcal immunization: capsular polysaccharide (CPS)-based vaccines and their implications on prevention. *Vaccine* 2011;29:3729–3730.
60. Caliot E, Firon A, Solgadi A, Trieu-Cuot P, Dramsi S. Lipid lysination by MprF contributes to hemolytic pigment retention in group B Streptococcus. *Res Microbiol* 2024;175:104231.
61. Zakerifar M, Kaboosi H, Goli HR, Rahmani Z, Peyravii Ghadikolaii F. Antibiotic resistance genes and molecular typing of *Streptococcus agalactiae* isolated from pregnant women. *BMC Pregnancy Childbirth* 2023;23:43.
62. Ernst CM, Staubitz P, Mishra NN, Yang S-J, Hornig G, *et al*. The bacterial defensin resistance protein MprF consists of separable domains for lipid lysinylation and antimicrobial peptide repulsion. *PLOS Pathog* 2009;5:e1000660.
63. Ernst CM, Peschel A. Broad-spectrum antimicrobial peptide resistance by MprF-mediated aminoacylation and flipping of phospholipids. *Mol Microbiol* 2011;80:290–299.
64. Hu Y, Kan Y, Zhang Z, Lu Z, Li Y, *et al*. New mutations of penicillin-binding proteins in *Streptococcus agalactiae* isolates from cattle with decreased susceptibility to penicillin. *Microb Drug Resist* 2018;24:1236–1241.
65. Davies EV, Winstanley C, Fothergill JL, James CE. The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol Lett* 2016;363:fnw015.
66. Lichvariková A, Soltys K, Szemes T, Slobodnikova L, Bukovska G, *et al*. Characterization of clinical and carrier *Streptococcus agalactiae* and prophage contribution to the strain variability. *Viruses* 2020;12:1323.
67. Weeks CR, Ferretti JJ. The gene for type A streptococcal exotoxin (erythrogenic toxin) is located in bacteriophage T12. *Infect Immun* 1984;46:531–536.
68. Joyce LR, Manzer HS, da C Mendonça J, Villarreal R, Nagao PE, *et al*. Identification of a novel cationic glycolipid in Streptococcus agalactiae that contributes to brain entry and meningitis. *PLoS Biol* 2022;20:e3001555.
69. Wilkinson SP, Grove A. Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. *Curr Issues Mol Biol* 2006;8:51–62.
70. Wang R, Li L, Huang Y, Luo F, Liang W, *et al*. Comparative genome analysis identifies two large deletions in the genome of highly-passaged attenuated *Streptococcus agalactiae* strain YM001 compared to the parental pathogenic strain HN016. *BMC Genomics* 2015;16.
71. Galperin MY, Moroz OV, Wilson KS, Murzin AG. House cleaning, a part of good housekeeping. *Mol Microbiol* 2006;59:5–19.
72. Harter E, Wagner EM, Zaiser A, Halecker S, Wagner M, *et al*. Stress survival islet 2, predominantly present in *Listeria monocytogenes* strains of sequence type 121, is involved in the alkaline and oxidative stress responses. *Appl Environ Microbiol* 2017;83:e00827-17.
73. Hein I, Klinger S, Dooms M, Flekna G, Stessl B, *et al*. Stress survival islet 1 (SSI-1) survey in Listeria monocytogenes reveals an insert common to listeria innocua in sequence type 121 *L. monocytogenes* strains. *Appl Environ Microbiol* 2011;77:2169–2173.
74. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan M-S, *et al*. Multilocus sequence typing system for group B streptococcus. *J Clin Microbiol* 2003;41:2530–2536.

7.1.4. Towards a holistic epidemiology of *Streptococcus agalactiae* using the BakRep repository

Linda Fenske[1], Oliver Schwengers[1], Alexander Goesmann[1]

[1]Bioinformatics and Systems Biology, Justus Liebig University, Giessen, Germany

bioRχiv (2026), DOI: 10.64898/2026.03.02.709001

The following contributions are attributed to the thesis author:

Formulated the research idea and planned and coordinated the study. Conducted bioinformatic data analyses, as well as statistical analyses and interpreted the data. Drafted the manuscript and finalized it with input from the co-authors.

Towards a holistic epidemiology of *Streptococcus agalactiae* using the BakRep repository

Linda Fenske[1*], Oliver Schwengers[1], Alexander Goesmann[1]

[1]Bioinformatics and Systems Biology, Justus Liebig University Giessen, Germany

***Correspondence:** linda.fenske@computational.bio.uni-giessen.de

Abstract

Streptococcus agalactiae is a versatile multi-host pathogen that can cause major neonatal disease in humans, as well as mastitis in dairy animals. Its ability to infect a wide range of hosts is largely driven by its high genomic plasticity and the acquisition of distinct accessory genes. The global population of *S. agalactiae* is characterized by multiple of capsular serotypes and clonal complexes that differ in their propensity to cause invasive disease, including hypervirulent CC17 (often serotype III) associated with neonatal meningitis, whereas CC1/CC19/CC23 are more often colonizing lineages. Although widely studied, most research is limited to particular regions or single outbreak events, offering only fragmented snapshots instead of a comprehensive global picture. To move beyond region- or outbreak-limited studies, this work has analyzed 37970 *S.agalactiae* genomes from BakRep, integrating serotypes, MLST, AMR genes, lineage-specific genes, and descriptive metadata to map current trends and identify potential gaps in public data. The dataset largely matched the known population structure with serotype III, Ia and V most common and stable serotype/clonal complex lineages (e.g. III-2/CC17, Ia/CC23, CC1/V), while also rising serotype diversity. Lineages differed in their accessory-gene profiles, with III-2/CC17 being enriched for virulence and adhesion genes, while other groups showed either greater genomic plasticity (mobile/phage genes) or niche specialization. AMR was widespread with very high tetracycline resistance (>80%), frequent MLSB resistance determinants, and emerging aminoglycoside resistance in some genomes. But overall it became evident that the associated metadata contained substantial gaps. Missing or incomplete information limits biological interpretation, underscoring that rigorously curated, structured metadata is essential for maximizing the value of ongoing sequencing efforts.

Keywords: *Streptococcus agalactiae,* serotypes, multilocus sequence typing, clonal complexes, antimicrobial resistance, metadata, GWAS

Introduction

Streptococcus agalactiae, also known as Group B Streptococcus (GBS), is a Gram-positive bacterium, capable of causing infections in various host species. In human medicine, GBS is a common cause of neonatal infections, primarily transmitted from mother to newborn during childbirth (1). An estimated 20-30% of pregnant women are colonized with GBS, with around 50% of their newborns acquiring colonization and approximately 1% developing invasive disease (2,3). Asymptomatic GBS colonization is also commonly observed in both young adults (4) and elderly individuals (5). In the veterinary context, GBS is primarily known as a contagious udder pathogen linked to mastitis. While it is most commonly associated with mastitis in cattle, cases have also been reported in other dairy species, including goats (6,7), sheep (8,9), and camels (10). Although most published GBS research centers on human and bovine cases, GBS has also been detected in several other terrestrial (11,12) and aquatic species (13,14). The broad host range of GBS has been linked to its high genomic plasticity, enabling the integration of accessory genes into its chromosome, which allows strains to colonize diverse hosts and ecological niches, providing an evolutionary advantage (15).

The outer layer of GBS consists of a polysaccharide capsule, whose antigenic properties serve as the basis for further subclassification into currently ten major serotypes (Ia, Ib, II-IX) (16) and four subtypes of serotype III (III-1, III-2, III-3, III-4) (17). The prevalence and distribution of serotypes are known to vary depending on host species, levels of virulence or geographic region (18). Additionally, multilocus sequence typing (MLST) classifies GBS isolates into distinct sequence types (STs), which can be further grouped into clonal complexes (CCs). This can offer insights into the pathogenicity of isolates. For example, CC17, is known as a hyper-virulent cluster, increasingly associated with neonatal infections (19,20). Specifically the combination of serotype III and CC17 is considered to be responsible for the majority of neonatal meningitis cases (21) which is likely attributable to specific genes encoding secreted and surface proteins, which facilitate interactions between the bacterium and host cells (19). In contrast, CC1, CC19 and CC23 are common colonizers of pregnant women and appear well adapted to the vaginal mucosa, with relatively limited invasive potential in neonates (22,23).

As in general, antibiotic resistance (AMR) is also a growing concern in GBS strains. Intrapartum antibiotic prophylaxis (IAP) is administered to pregnant women based on risk factors or screening guidelines, making continuous monitoring of resistance rates essential (24). In mastitis cases, particularly in middle-income countries with higher GBS prevalence, widespread use of antimicrobials, including broad-spectrum and critically important antibiotics, contributes to antimicrobial resistance in multi-host pathogens, reducing treatment options for both animals and humans (25).

Numerous studies have analyzed GBS, though they are often restricted to specific geographic regions or isolated outbreak events. As a result, only limited subsets are analyzed rather than a more holistic overview of the entire population. In 2024, BakRep was released (26), featuring extensive genome characterizations based on 661405 uniformly processed assemblies of all short-read sequence data from the European Nucleotide Archive (ENA) up to November 2018 (27). Up to now, the collection of uniformly assembled data was expanded to 2440377 assemblies (28), which have since been

characterized and made accessible through the BakRep web repository. With its vast collection of genomic information and associated metadata, this repository provides an ideal resource for obtaining a comprehensive overview of publicly available sequence data, enabling insights beyond individual studies.

In the present study, we conducted a comparative analysis of a total of 37970 GBS genomes contained in BakRep v2. We characterized all isolates by MLST, capsular serotypes, AMR genes, putative lineage-specific genes, and complemented this with metadata on geographic origin, host species, and disease patterns. This integrated analysis helps to evaluate current trends in the GBS study landscape, to highlight well-studied areas, and finally to identify understudied gaps with only scarce public data.

Methods

Export data from the BakRep repository

All genome data and associated metadata were retrieved from BakRep (https://bakrep.computational.bio/) (26). For this, the repository was searched for entries where the species exactly matches “Streptococcus agalactiae” (operator “eq”) and all search results were exported into a TSV file. Additional files provided by BakRep were downloaded via the BakRep CLI tool (https://github.com/ag-computational-bio/bakrep-cli) by providing the dataset ids of the identified GBS genomes.

Data analysis

Capsular-serotypes were determined using KMA (v1.4.12) using the FASTA files as input (29). AMR genes were identified using AMRFinderPlus (v4.0.23) with default parameter settings (30).

Bakta-generated GFF3 annotation files (31) were used as input for Panaroo (v1.5.0) (32). The resulting gene presence/absence matrix, combined with a binary trait file of selected phenotypes, was then used to run a pan-GWAS in scoary2 (v0.0.15) using default settings (33). The metadata retrieved from BakRep and all additional results have been analyzed using R (v4.4.2).

Results

Sequence statistics of GBS genomes

BakRep v2 contained 37970 genomes classified as *Streptococcus agalactiae*. The average estimated completeness of all GBS genomes was 99.94%, whereby only 98 (0.26%) genomes were below a completeness of 99%. The average estimated contamination was 0.49%, consisting of 139 (0.37%) genomes, which had a contamination lower than 50%, and a total of 529 (1.39%) genomes which had a contamination lower than 1% (Supplemental Figure 1).

The assembled contigs of all GBS genomes had total lengths ranging from 105259 to 12119634 base pairs (bp), with an average genome size of 2095432 bp. In the NCBI database, GBS genomes exceeding 2.6 Mbp are flagged as unusually large for this species. This size is exceeded by 206 genomes in BakRep. However, genomes exceeding this size threshold also exhibit a significantly higher rate of estimated contamination (data not shown). The GC content varied between 33.25% and

58.16%, averaging to 35.33%. The number of contigs per assembly ranged from 10 to 6016 with a mean of 51.59 (Figure 1).

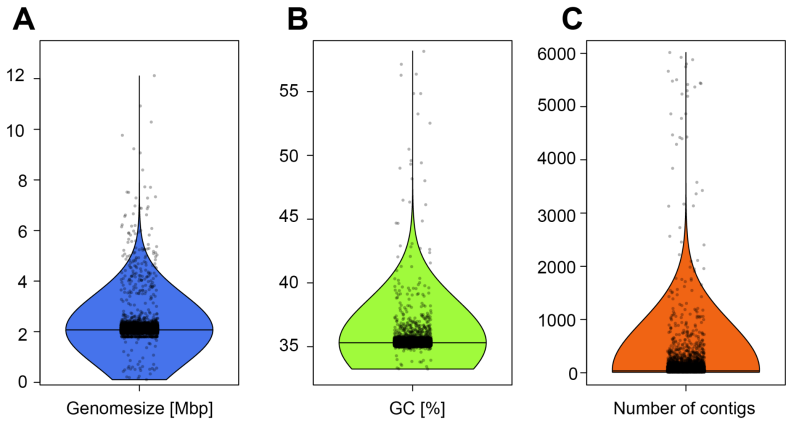


Figure 1: Comparison of genome size (**A**), GC content (**B**), and number of contigs (**C**) of all GBS genomes contained in BakRep v2 (n=37970). Individual dots represent single genomes, with horizontal jitter added to avoid overplotting. Horizontal black lines indicate medians; bold black bars represent interquartile ranges; vertical black lines represent outliers.

Metadata of GBS genomes

The metadata associated with the GBS genomes in BakRep v2 indicated that the genomes originate from 51 countries across six continents. The majority of genomes derive from North America (49.73%; n=18884), with the United States alone accounting for 48.59% (n=18451) of all genomes. No isolation region was provided for 14747 (38.84%) genomes (Figure 2).

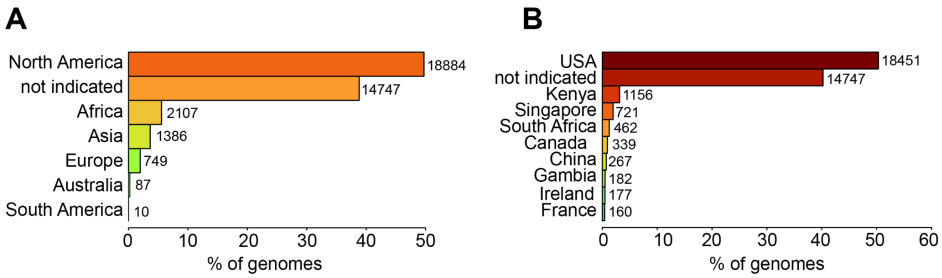


Figure 2: Distribution of GBS genomes contained in BakRep v2 across different continents (**A**) and the ten most common countries (**B**). The total number of genomes for each country and continent, respectively, is displayed right to the bars.

The majority of genomes were declared as human isolates (58.41%; n=22179), followed by 61 bovine-derived genomes (0.16%) and 48 genomes obtained from different fish species (0.12%). Three (0.01%) genomes were listed as isolates from dogs, while individual (0.004%, each) genomes were obtained from a cat, a llama, and a crocodile, respectively. A host species was not provided for 41.28% (n=15675) of genomes. For 98.35% (n=37342) of the genomes, no information on disease

status was provided; 0.88% (n=334) were reported as diseased, 0.60% (n=226) as healthy, and 0.16% (n=60) as carriage isolates. The collection date spans from 1972 to 2024. Regarding the yearly distribution of isolates, a gradual increase is observed over time. Between 1972 and 2005, genome counts remained very low, but with a noticeable increase from 2006 onwards. The majority of genomes were collected between 2015 and 2022, with a slight decline observed after 2020 and very few genomes available for 2023 and 2024 (Supplemental Figure 2). For 39.05% (n=14829) of genomes no information for the isolation year was provided.

Serotype, Sequence Type and Clonal complex distribution

Some genomes had capsular serotype information provided in the associated metadata. However, since the methodology used for serotype determination was unclear and 98.02% (n=37629) genomes lacked serotyping, capsular serotypes were *in silico* reanalyzed for all genomes. The analysis of the serotype distribution revealed that most GBS genomes belonged to one of the nine predominant GBS serotypes plus all four subtypes of serotype III. Only 97 (0.26%) genomes were non-typable. Serotype III was the most common, accounting for 24.41% (n=9270) of all genomes, followed by serotype Ia at 21.07% (n=8000), and serotype V at 16.25% (n=6172) (Figure 3A).

MLST analysis identified 1015 unique STs, which covers 40.39% of the STs currently listed in PubMLST. ST23 was most prevalent at 14.56% (n=5530) of all genomes, followed closely by ST1 at 14.31% (n=5434) and ST17 at 11.62% (n=4414). The STs could be grouped into eleven CCs. The main CCs among all GBS genomes were CC23 (18.99%; n=6935), CC1 (18.25%; n=6669), and CC19 (15.48%; n=5654) (Figure 3B).

The analysis of the associations between CCs and serotypes revealed several connections. CC23 was strongly associated with serotype Ia, which accounted for 93.73% (n=6500) of those genomes. CC17 was mainly associated with serotype III-2, which accounted for 92.38% (n=4929) of those genomes. CC459 was nearly exclusively linked to serotype IV (98.60%; 2331), while CC22 was similarly linked to serotype II (99.53%; n=1479). CC1 showed a predominance of serotype V (55.56%; n=3705), followed by substantial proportions of serotypes Ib (14.78%; n=986), II (14.42%; n= 962) and VI (7.67%; n=512), indicating a more moderate serotype diversity within this lineage. CC19 presented an even more heterogeneous serotype distribution, with serotype III-1 (47.40%; n=2680), II (31.52%; n=1782) and V (18.84%; n= 1065) co-occurring (Figure 3C). Although most serotypes were associated with multiple STs, serotype IX was found predominantly in ST130 (93.10%; n=108). Exceptions included a single genome each in ST1208 and ST1216, both single-locus variants (SLVs) of ST130, as well as one occurrence in ST24, which belongs to CC452. Similarly, serotype VI was largely restricted to CC1, except for two occurrences in CC12 and 16 occurrences in ST889. Serotype VII was likewise primarily associated with CC1, except for one occurrence in ST103 and one occurrence in ST1438. All detailed numbers are listed in Supplementary Table S1.

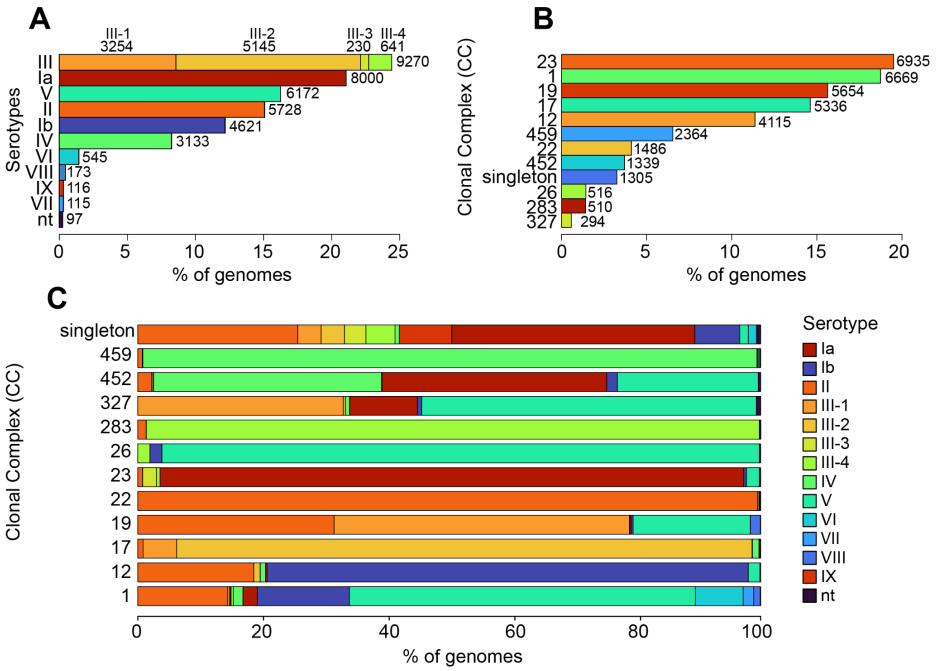


Figure 3: Serotype **(A)** and CC **(B)** distribution, and serotype allocation **(C)** among the CCs of all GBS genomes contained in BakRep v2. The total number of genomes for each serotype and CC, respectively, is displayed right to the bars.

Continental and temporal variation in serotypes and clonal complexes

Serotype and CC distributions showed geographic variation. In North America, serotypes Ia (20.74%; n=3917), III (18.43%; n=3480), distributed across all III subtypes, and II (17.90%; n=3380) were most frequently observed (Figure 4A). In contrast, serotype III (42.76%; n=901) predominated in Africa, particularly subtypes III-1 and III-2 (Figure 4B). Asia showed a distinct pattern, with serotype III-4 (28.14%; n=390) being most prevalent, a subtype that was rare in other regions (Figure 4C). In Europe, serotypes III-2 (24.03%; n=180), Ia (20.83%; n=156), and II (18.02%; n=135) were among the most common, whereas in Australia, serotype Ib (29.89%; n=26) was the most frequently identified (Figure 4D+E). Regarding clonal complexes, CC1 (20.68%; n=3763) and CC23 (19.78%; n=3599) were dominant in North America. In Africa and Europe, CC17 (30.75%; n=643 and 32.86%; n=230) was the most prevalent, while Asia was characterized by a high proportion of CC283 (28.11%; n=377) and CC1 (27.89%; n=374). Notably, both serotype III-4 and CC283 were found almost exclusively in Asia.

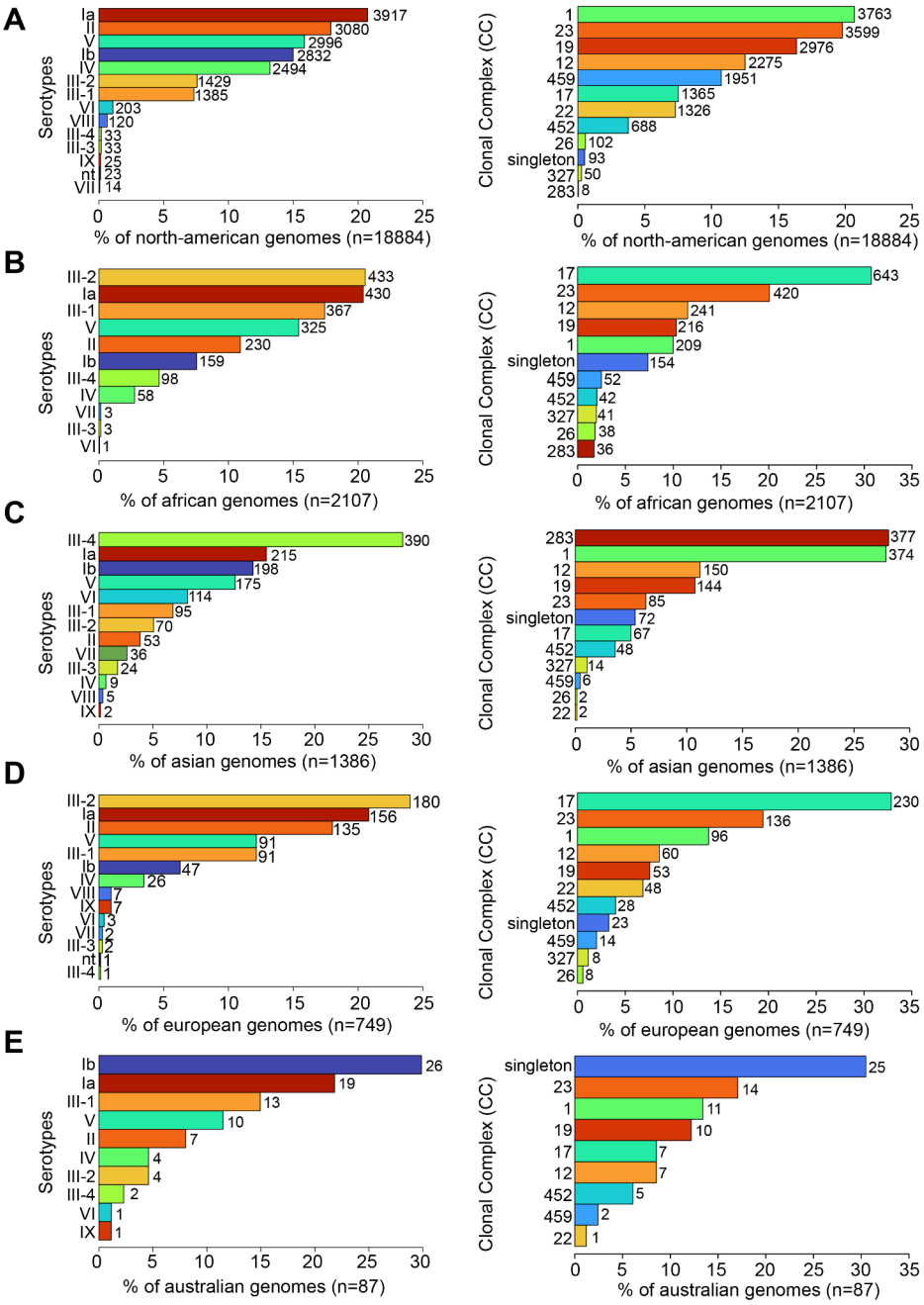


Figure 4: Serotype (left) and CC (right) distribution by continent. (**A**) North-American genomes, (**B**) African genomes, (**C**) Asian genomes, (**D**) European genomes and (**E**) Australian genomes. The total number of genomes is displayed right to the bars.

Examining the temporal distribution of serotypes revealed that serotype Ia has remained consistently prevalent over the years, comparable to serotype V, which, however, shows greater fluctuations. Serotype Ib and II showed a gradual increase, whereas III-1 appeared to have slightly declined. In contrast, serotypes III-3, VIII and IX have consistently remained at very low frequencies (Figure 5A).

The temporal distribution of CCs showed an increase in diversity and abundance over time. Prior to the early 1990s, CCs were largely restricted to CC1. From the mid-1990s onwards, additional CCs such as CC17, CC19, CC23 and CC283 emerged, although at relatively low frequencies. From 2004 onwards there was a concurrent presence and rapid increase of multiple dominant lineages, including CC1, CC12, CC17, CC19 and CC23 (Figure 5B). Detailed numbers for each year are listed in Supplementary Table S2.

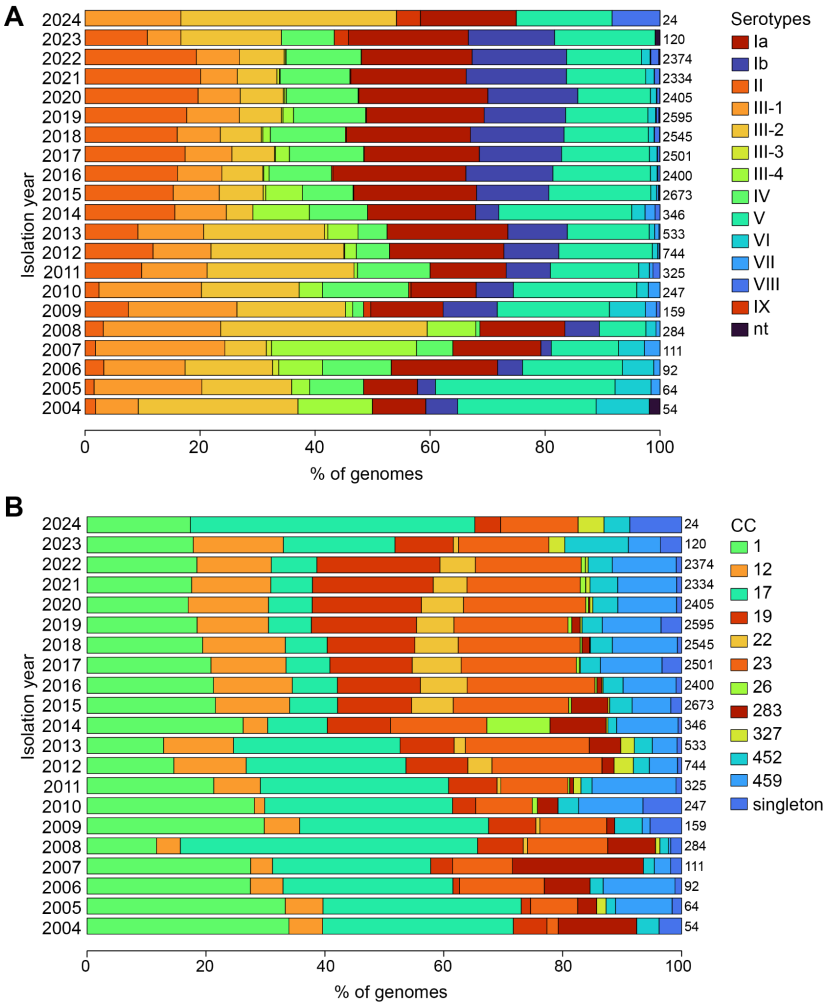


Figure 5: Serotype **(A)** and CC **(B)** distribution per year. Displayed are the years 2004-2024. The total number of genomes for each isolation year is displayed right to the bars.

Resistome of GBS genomes

A total of 88.10% (n=33452) GBS genomes harbored at least one AMR gene. In total 132 known AMR gene determinants were identified conferring to resistance against 24 drug classes. Tetracycline resistance genes (a total of 16 different) were the most prevalent AMR determinants, observed in 84.24% (n=31985) of all genomes. Of these, the most common genes were *tet(M)* identified in 76.99% (n=29232) of genomes and *tet(O)* identified in 8.42% (n=3196) of genomes. The second most common were genes conferring for macrolide-lincosamide-streptogramin B resistance (MLSB) present in 17.66% (n=6707) of all GBS genomes. Of that, most common were *erm(B)* observed in 16.38% (n=6220) of all genomes, *erm(A)* observed in 13.11% (n=4976) of genomes and *erm(T)* observed in 1.3% (n=492) of genomes. 4.15% (n=1574) of all genomes carried aminoglycoside resistance genes. Only 0.14% (n=55) of all genomes carried beta-lactam resistance genes.

More than one AMR gene hit was found for 41.34% (n=15695) of the genomes, including 447 genomes with more than five hits and 11 genomes with more than ten hits. In total, 40.57% (n=15403) of genomes carried resistance genes for multiple drug classes, and 2.89% (n=1098) harbored genes for three or more classes, classifying them as multidrug-resistant. Notably, in one genome, resistance genes against as many as seven different drug classes were detected (ID: SAMEA112769107).

Since erythromycin and tetracycline resistance genes are often located on the same transposon, we examined their co-occurrence to determine in how many genomes both genes were present simultaneously and how much only contained one gene variant individually. The combination of a erythromycin and tetracycline resistance gene was present in 27.62% (n=10485) of genomes, while a tetracycline resistance gene alone was detected in 56.62% (n=21500) and an erythromycin resistance gene alone in 2.92% (n=1109). The combination of *erm(B)* and *tet(M)* was the most common combination at 1.1% (n=4180) of all genomes followed by the combination of *erm(A)* and *tet(M)* at 1.04% (n=3963), and the combination of *erm(B)* and *tet(O)* at 0.58% (n=2209). The combination of *tet(M)* and *mef(A*) was detected in 7.86% (n=2979) of all genomes and the combination of *tet(O*) and *mef(A)* was present in 0.48% (n=183) of all genomes.

Furthermore, macrolide resistance in streptococci is mediated by an efflux system, with *mef(A)* encoding the transmembrane channel and *msr(D)* the ATP-binding domains (34). We therefore analyzed the presence of *mef(A)* and *msr(D)* across all genomes, identifying both genes in 8.52% (n=3236) of genomes, while *mef(A)* alone was found in 0.04% (n=16) and *msr(D)* alone in 0.01% (n=4). Detailed numbers for all AMR genes and classes are listed in Supplemental Table 3.

Serotype and CC associated resistome patterns

Among the eleven CCs, CC1 harbored the highest number of distinct AMR genes (n=72). *Tet(M)* was the most prevalent gene across all CCs, with the exception of CC459. CC459 stood out with *erm(A)* as its most common gene, showing a markedly higher frequency compared to other CCs (Figure 6F). Aminoglycoside resistance genes, such as *ant(6)-Ia* and *aph(3')-IIIa*, were detected in all CCs, with the highest abundance observed in CC17 (Figure 6D). CC17 also harbored the highest number of hits for the trimethoprim resistance gene *dfrF*. Several Vancomycin resistance genes were exclusively found in CC1 (*vanS-G, vanY-G1, vanR-G, vanU-G*). The Macrolide and Strepogramin resistance gene *msr(D)* was by far the most common found in CC23 (Figure 6A). Serotype II showed the highest AMR gene diversity, encompassing 72 unique resistance determinants, with *tet(M)* being universally present. Notably, in serotype VIII, *tet(O)* was more prevalent than *tet(M)*. Similar to CC459, serotype IV was characterized by a high prevalence of *erm(A)*. In contrast, serotype Ia exhibited fewer *erm(A)* and *erm(B)* hits but a higher occurrence of *mef(A)* and *msr(D)* (Figure 7). The highest proportion of genomes without any AMR hits was observed in CC283 (58.82%), followed by CC452 (31.76%). Among serotypes, serotype IX showed the largest fraction without detected AMR genes (95.69%) followed by serotypes VIII (78.03%), VI (61.16%), and III-4 (51.34%).

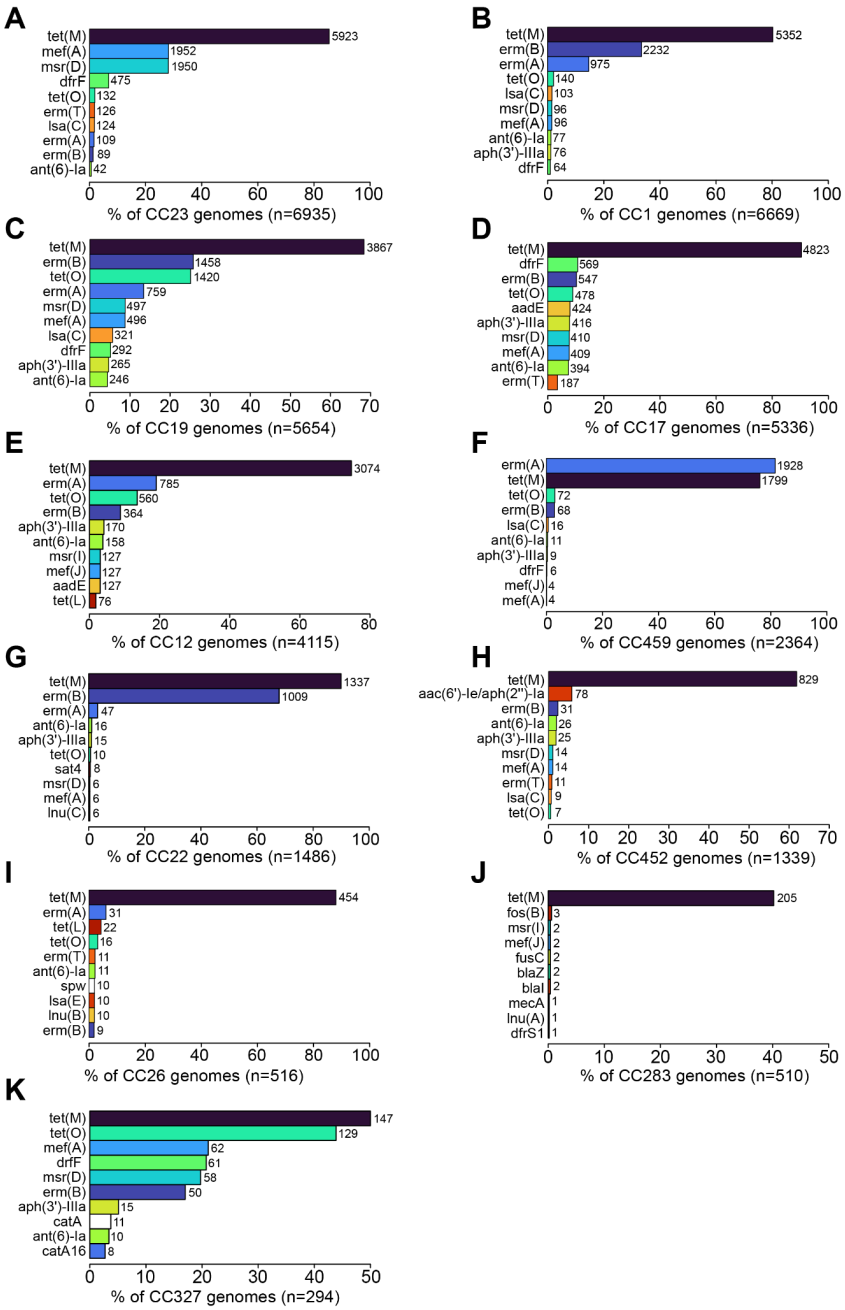


Figure 6: Distribution of the ten most prevalent AMR genes per CC. (**A**) CC23, (**B**) CC1, (**C**) CC19, (**D**) CC17, (**E**) CC12, (**F**) CC459, (**G**) CC22, (**H**) CC452, (**I**) CC26, (**J**) CC283, (**K**) CC327. The total number of genomes for each AMR gene is displayed right to the bars.

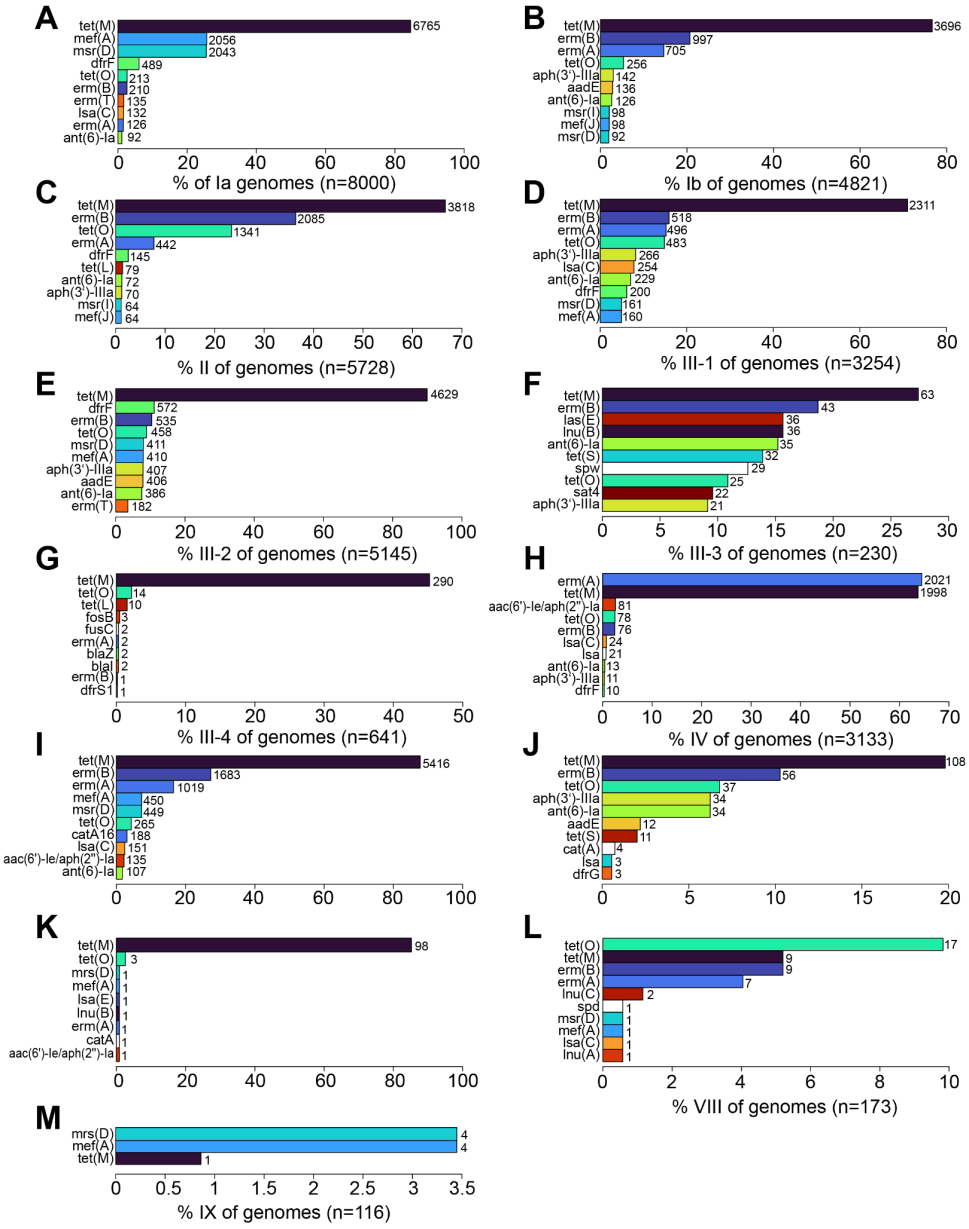


Figure 7: Distribution of the ten most prevalent AMR genes per serotype. (**A**) Ia, (**B**) Ib, (**C**) III-1, (**D**) III-2, (**E**) III-3, (**F**) III-3, (**G**) III-4, (**H**) IV, (**I**) V, (**J**) VI, (**K**) VII, (**L**) VIII, (**M**) IX. The total number of genomes for each amr gene is displayed right to the bars.

Regional resistome patterns

Additionally, the geographical distribution of AMR genes was analyzed. *Tet(M)* emerged as the most widespread resistance gene, detected across all continents. North America exhibited the highest diversity and prevalence of AMR genes (n=59), especially those conferring resistance to macrolides and tetracyclines (Figure 8A). In Africa, several less common resistance genes were identified, including *dfrF*, *aac(6')-Ie/aph(2'')-Ia*, *msr(I)*, *mef(J)*, and *catA16* (Figure 8B). The resistance gene profiles in Asia and Europe were largely similar, albeit with slightly lower overall prevalence (Figure 8C+D).

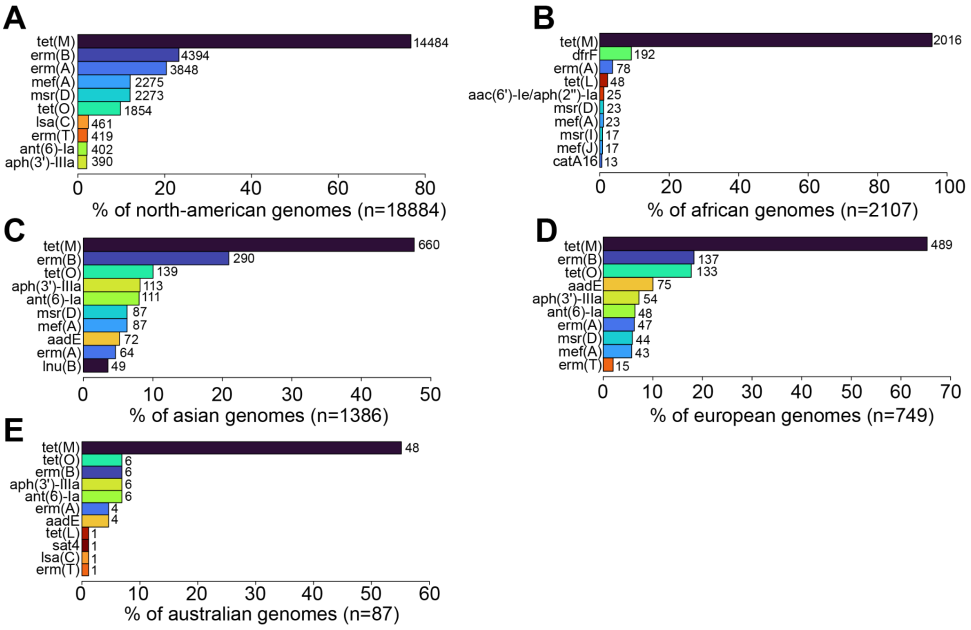


Figure 8: Distribution of the ten most prevalent AMR genes per continent. (**A**) North-America, (**B**) Africa, (**C**) Asia, (**D**) Europe, (**E**) Australia. The total number of genomes for each amr gene is displayed right to the bars.

Temporal resistome variations

AMR genes were largely absent in isolates collected prior to the 1990s, with only sporadic occurrences of certain tetracycline and macrolide resistance genes appearing in the late 1980s. *Tet(M)* remained the most prevalent gene throughout the past two decades but since 2008, the prevalence of *tet(O)* has also risen markedly. MLSB genes like *erm(A)* and *erm(B)* became more frequent from 2000 onward, showing a steady increase. For aminoglycoside resistance genes, there was no consistent upward trend over time, instead only sporadic peaks are observed. Over all, AMR genes have shown a clear upward trend, with a gradual increase in diversity and frequency over the last 20 years (Figure 9). Detailed numbers for each year are listed in Supplementary Table S2.

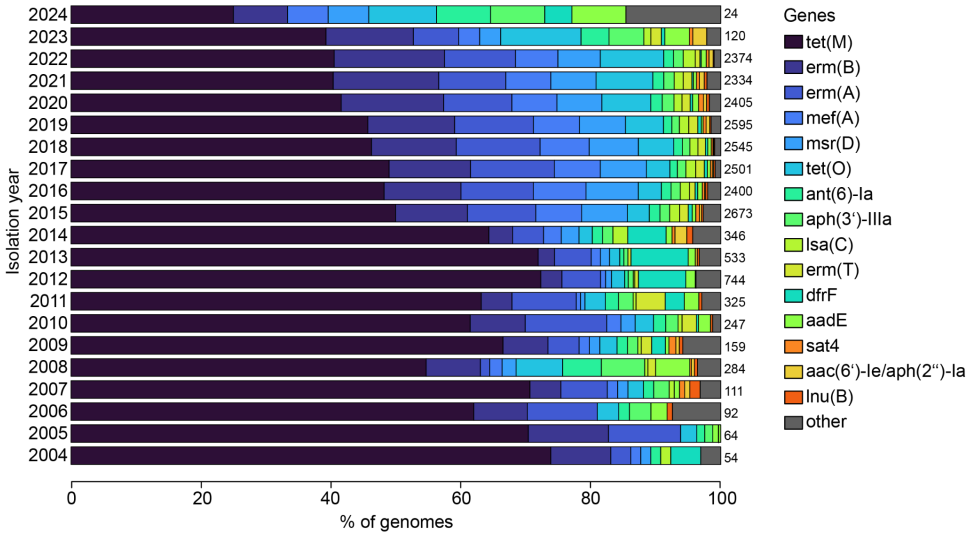


Figure 9: AMR gene distribution per year. Displayed are the years 2004-2024. The total number of genomes for each isolation year is displayed right to the bars.

Pan genome analysis and genome-wide association investigation

Gene-content analysis of the 37970 GBS genomes identified a pan genome of 274610 genes and a core genome of 1398 genes. The distribution of genes displayed that most genes occurred in only a small fraction of genomes, with more than 99% of genes present in fewer than 15% of isolates (cloud genes). Accordingly, we filtered the gene presence-absence matrix out of Panaroo to retain genes present in more than 0.25% and less than 95% of genomes, resulting in 4101 remaining genes (Figure 10). The phylogenetic analysis done by scoary2 showed, although the genomes cluster partly according to serotype and CC, there are also areas with very diverse distribution (Figure 11).

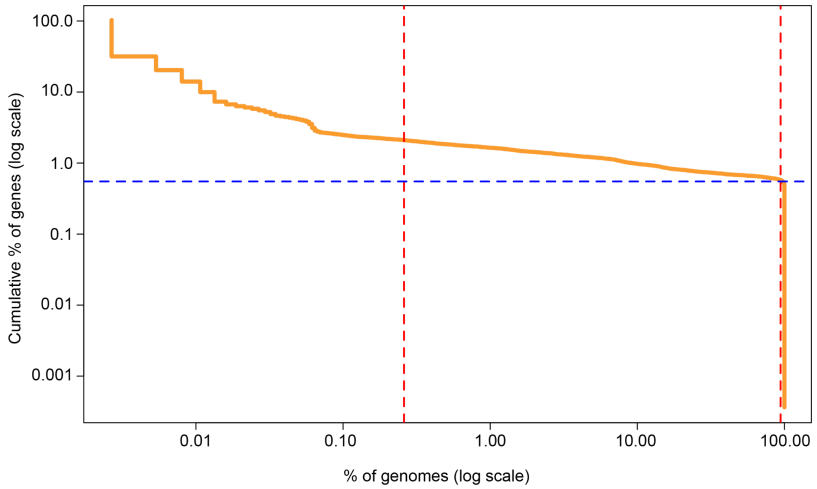


Figure 10: Pan-genome gene frequency distribution. The x-axis represents the percentage of genomes containing a given gene, and the y-axis represents the cumulative percentage of genes, both on logarithmic scales. Vertical red dashed lines indicate the lower and upper frequency cutoffs (0.25% and 95% of genomes), while the blue dashed horizontal line marks the core genome.

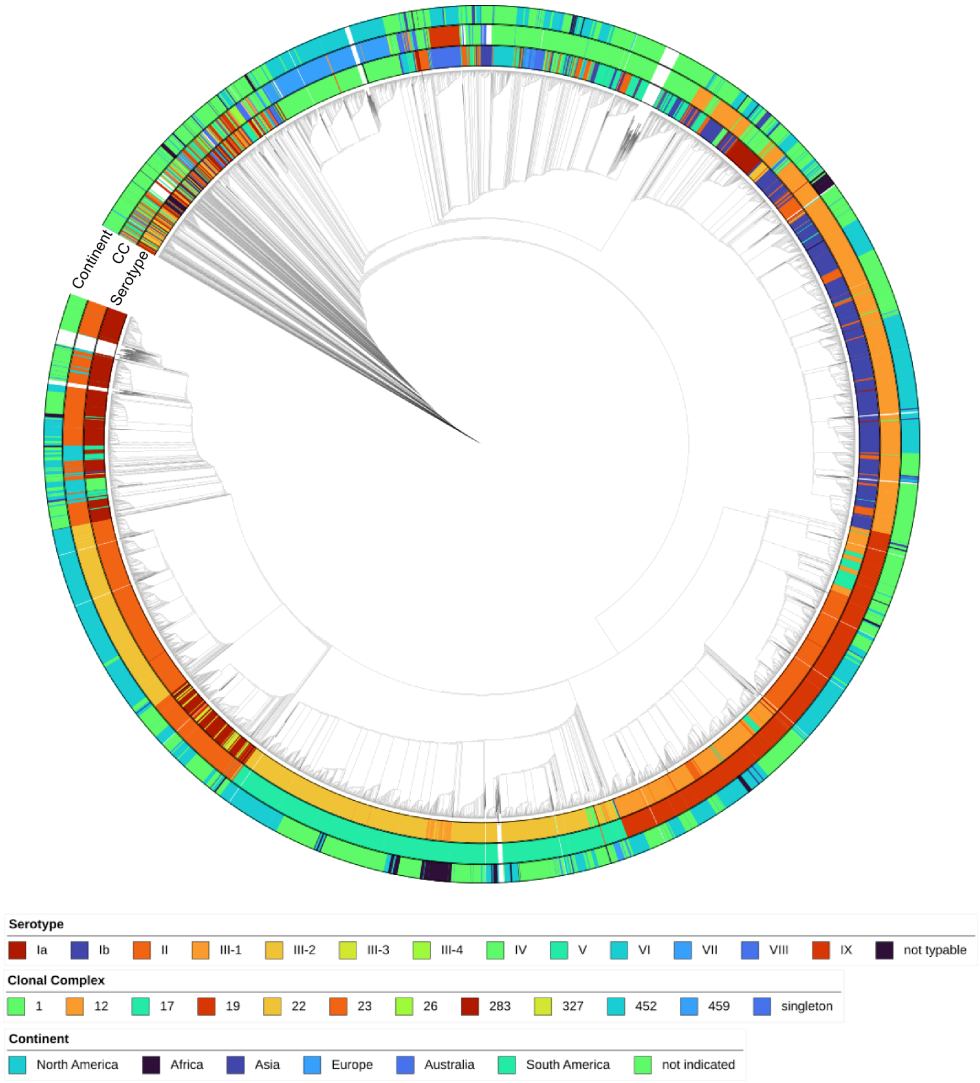


Figure 11: Phylogenetic tree of 37970 GBS genomes. The tree was constructed with scoary2. Colored blocks show serotype, clonal complex and continent of isolation. Annotation of the tree was done using iTol.

Scoary2 analyses were performed on the five most common serotype/clonal-complex combinations (Ia/23, II/22, III-2/17, IV/459, V/1) to identify genes predominantly associated with each combination. Multiple genes showed positive and negative associations with each group. Specifically, 60, 54, 41, 86, and 71 genes were positively associated with Ia/23, II/22, III-2/17, IV/459, and V/1, respectively, while 31, 12, 17, and 1 genes were negatively associated, with more than 70% specificity and sensitivity. Notably, group III-2/17 contained the highest number of positively associated genes with a

specificity and sensitivity greater than 90% (n=30) followed by group II/22 (n=17). Group Ia/23 harbored several mobile genetic elements, numerous phage-associated genes and loci involved in metal homeostasis and heavy-metal resistance. In contrast, groups II/22 and III-2/17 encoded more functionally specialized genes. II/22 was enriched for genes related to metal uptake, including siderophore biosynthesis, whereas III-2/17 showed a marked accumulation of virulence and adhesion determinants, particularly components of the accessory secretion system (SecA2/SecY2 and Asp proteins). Groups IV/459 and V/1 were dominated by phage-associated genes. Table 1 lists the top 3 positive and negative hits per group. All other genes are listed in Supplementary Table S4.

Table 1: Scoary2 results are reported for the most significant positively (+) and negatively (-) associated genes of the five groups ((Ia/23, II/22, III-2/17, IV/459, V/1). The top 3 hits positive and negative are displayed. **SE:** Sensitivity of using the presence/absence of this gene as a diagnostic test to determine trait-positivity. **SP:** Specificity of using the absence/precence of this gene as a diagnostic test to determine trait-negativity. **P-value:** The *p*-value of the post-hoc permutation for the best gene.

Phenotype	pos(+) /neg (-)	Gene	Annotation	SE	SP	*p*-value
Ia/23 (n=6497)	+	group_40999	Phosphoribosylaminoimidazole carboxylase ATPase subunit	100.0	91.71	0.0020
	+	group_51961	DUF6287	100.0	91.69	0.0020
	+	*purN*	Phosphoribosylglycinamide formyltransferase	99.98	88.82	0.0020
	-	group_9603	GNAT family N-acetyltransferase	99.45	94.18	0.0020
	-	*aprE*	S8 family serine peptidase	99.46	85.12	0.0020
	-	group_11313	hypothetical protein	99.86	73.96	0.0020
II/22	+	group_41231	DUF6261	100.0	98.61	0.0020
	+	group_52008	Membrane protein	100.0	98.60	0.0020
	+	group_51962	Phage protein	100.0	98.60	0.0020
	-	group_6272	Cingulin	100.0	81.76	0.0020
	-	group_23506	Tandem five-TM protein	100.0	73.78	0.0020
	-	group_19754	KTSC domain-containing protein	100.0	73.67	0.0020
III-2/17	+	group_6225	Cell wall surface anchor protein	99.98	90.46	0.0020
	+	spaA~~~inlB	Sortase	99.96	90.31	0.0020
	+	group_10458	S8 family serine peptidase	99.94	96.21	0.0020
	-	group_41244	HIT family protein	99.96	74.90	0.0020

	-	group_9833	Lipoprotein	99.88	87.97	0.0020
	-	group_9933	TIGR04197 family type VII secretion effector	99.88	71.98	0.0020
IV/459	+	group_9154	Integrase	99.96	98.95	0.0020
	+	*padR*	Transcriptional regulator	99.96	98.95	0.0020
	+	group_41269	N-acetyltransferase	99.96	98.95	0.0020
	-	group_11745	Relaxase	99.96	77.27	0.0020
V/1	+	neuA~~~wcaA~~~cpsJ	Glycosyltransferase (Capsular polysaccharide biosynthesis protein)	98.49	91.80	0.0020
	+	group_25544	Phage protein	98.22	76.39	0.0020
	+	group_24863	Phage protein	98.22	76.20	0.0020

Other works reported the *scpB-lmb* transposon mainly associated with the human host (35,36). Because most genomes were decelerated as human isolates, we examined how frequently these genes occurred in the dataset. The *scpB* gene was detected in 99.91% (n=37934) of genomes and *lmb* in 94.85% (n=36013). However, 55 distinct *scpB* hits were found by Panaroo, whereas *lmb* was represented by a single variant.

Discussion

Streptococcus agalactiae is a known multi-host pathogen, primarily linked to infections in humans and cattle. Although often studied, most investigations focus on specific regions or outbreaks, leaving a lack of global perspective on GBS diversity. BakRep, with its data size and integration of genomic data and metadata, offers an opportunity to examine study trends on a broader scale and supports more holistic analyses of the GBS population as a whole. Based on this genome repository, in total 37970 GBS genomes were analyzed.

The predominance of genomes originating from North America reflects the region's leading role in clinical trials, holding 44.19% of the market in 2024 (37). More remarkable, however, is the absence of an isolation location in over a third of the data. The sharp rise in available genomes after 2006 likely reflects the introduction of the first commercially available DNA sequencing platform in 2005, which enabled large-scale sequencing efforts (38). Yet, similar to the location, the isolation date remains unknown for nearly two-fifths of the genomes. To enhance the quality and FAIR usability of sequence data, the International Sequence Database Collaboration (INSDC) introduced new standards requiring the inclusion of spatio-temporal metadata in all new ENA submissions. However, this measure has only been in place since May 2023, and although this is mandatory by now, it is still possible to report missing values (39). Same goes for the host species and the disease status (40). Public repositories are vital for advancing science by providing access to vast datasets. Nevertheless, the surge in data submissions, driven by cheaper and faster sequencing technologies, has outpaced the ability to rigorously check metadata quality (41). But without proper metadata, meaningful comparisons and global context are impossible, diminishing the value of rapidly generated sequence data. Yet, effective data sharing is essential for transparency, reproducibility, and enabling others to build on existing work. Without improved practices, the true utility of sequence data remains in question (42).

Although overall GBS colonization rates are similar worldwide, the distribution of serotypes and sequence types varies by region, host species, and clinical presentation. Two global reviews from 2023 and 2012 reported serotype III as the most prevalent, followed by serotype Ia and V (1,18). When serotype III subtypes are combined, this pattern closely mirrors the distribution identified here. Serotype III is predominantly linked to neonatal meningitis (21,43), while serotypes Ia and V are more often associated with invasive infections in non-pregnant adults (44,45). Given that over half of the genomes analyzed here originate from humans, this distribution is expected. However, it remains unclear whether the same pattern holds across non-human hosts globally. Although serotype IX was only recently identified (46), it already occurs more frequently than some other serotypes in certain regions. Increasing attention should also be given to the rising prevalence of serotype IV. Recent studies highlight this serotype as a hotspot for genetic recombination, evidenced by its occurrence across diverse genetic backgrounds (47). The growing diversity of GBS serotypes, combined with the potential for capsular switching, poses a major challenge for vaccine development, since capsular polysaccharide based vaccines may exert selective pressure that enables virulent genotypes to escape coverage. Different studies already demonstrated such switching from serotype III to IV within

the hypervirulent CC17 lineage, underscoring that even highly conserved clones can alter one of their main vaccine targets (48,49). Several studies consistently report CC23, CC1 and CC19 as the most frequent lineages in global GBS collections, which indicates that these CCs represent stable and widely distributed lineages within the global GBS population (22,50,51).

Previous research found CC17 is almost exclusively linked to subtype III-2 and identified III-2/ST17 as a particularly virulent GBS clone, specifically associated with late-onset neonatal disease (50,52–55). Our investigations confirm the strong association between serotype III-2 and CC17, although no conclusions about disease status can be made due to missing metadata. Likewise, the III-1/ST19 combination is well known, and appears particularly in early-onset neonatal disease, often associated with colonizing isolates (43,52,56). The strong linkage between CC23 and serotype Ia is also well recognized, particularly in cases of invasive disease in non-pregnant adults (49,57). CC459 stood out due to its high prevalence of the macrolide resistance gene *erm(A)*, which is clinically relevant since this lineage is strongly associated with serotype IV. A 2015 study reported that ST459 strains dominate the serotype IV population responsible mainly for adult disease (58). Molecular analyses also confirm the strong link between CC1 and serotype V, with most serotype V isolates classified as ST1, while ST19 constitutes the main non-ST1 background (59,60).

Such strong serotype specificity suggests stable clonal lineages and preservation of capsule types, which can serve as useful markers for targeted surveillance. In contrast, some lineages such as CC1 and CC19 are highly heterogeneous, with CC1 showing the broadest serotype diversity. This variation has been proposed to result from an already diverse ancestral population rather than ongoing recombination (61). But most importantly, such heterogenous lineages make vaccination strategies more difficult and may harbor non-vaccine strains.

Although some studies describe serotype III as the predominant serotype in Africa, detailed data on the most common subtype remain lacking (62–64). Nevertheless, across serotype III isolates overall, III-1 and III-2 are the most frequently observed subtypes (52). Serotype III-4 and ST283 has been linked to disease in fish in several studies from Asia (13,65). Yet, more detailed metadata about the host species is needed to determine whether this represents a fish-associated lineage or possibly a case of zoonotic transmission. A study from 2021 described serotype Ib as more prevalent in Asia than in other regions of the world (66). This cannot be directly confirmed, as serotype Ib occurs at similar proportions in North America and even predominates in Australia, though the very low overall number of genomes there must be considered.

AMR is an increasingly relevant concern in GBS, impacting both treatment options and future prevention strategies. Penicillin remains the first-line treatment for colonized pregnant women, with clindamycin, erythromycin, and vancomycin recommended as alternatives in cases of penicillin allergy (67). However, resistance to second-line antibiotics like erythromycin and clindamycin has continued to rise, leading to updated treatment guidelines that now recommend cephalosporins or vancomycin

as alternative therapies for adults (68). Only a small number of genomes contained genes potentially associated with beta-lactam resistance. However, since the primary mechanism for beta-lactam resistance in GBS involves amino acid substitutions in penicillin-binding proteins (PBPs) (69), this aspect was not addressed in the present analysis. Notably, MLSB resistance genes, particularly *erm(B)* and *erm(A)*, represented the second most prevalent drug class in this dataset, highlighting an important trend that should be carefully considered in future research on GBS antimicrobial resistance. Severe invasive GBS infections are often treated with a combination of penicillin and gentamicin (68). Among the genomes displaying aminoglycoside resistance, 340 specifically harbored genes conferring gentamicin resistance. In light of such emerging resistance trends, alternative antibiotic classes should be considered.

Over 80% of the analyzed genomes harbored genes for tetracycline resistance. This resistance is well known (70,71) and largely stems from the widespread use of these antibiotics since their introduction in 1948, which likely promoted the clonal expansion of resistant GBS strains. This expansion has, in turn, contributed to the emergence of lineages better adapted to human colonization and infection, most notably the dissemination of the hypervirulent ST17 clone (24,72). Tetracycline resistance genes are often located on the same mobile genetic elements as erythromycin resistance genes, especially with *erm(B)* frequently co-occurring with *tet(M)* across multiple species (71) which is also evident in this data set. While *tet(M)* is the predominant tetracycline resistance gene in GBS, mobile elements carrying other determinants like *tet(O)* or *tet(S)* have also been reported (24,73). Similarly, other erythromycin resistance determinants are frequently linked to *tet* genes and conjugative transposons capable of horizontal transfer, which can spread rapidly among GBS lineages, increasing the risk of widespread antimicrobial resistance (73). The combination of specific resistance genes should be closely monitored, as their co-occurrence may indicate the spread of multidrug-resistant lineages. The increasing spread of AMR genes in the last decades poses a serious problem and underscores the need for continuous monitoring in order to adapt prevention strategies in a targeted manner. But, the detection of AMR genes does not automatically imply phenotypic resistance. Resistance requires gene expression, and in some cases, the presence of an AMR gene may only partially reduce antibiotic susceptibility without reaching clinical resistance thresholds. Conversely, isolates may gain or lose resistance through other mutations. Therefore, genotypic results should be interpreted with caution and ideally supported by phenotypic testing. Moreover, standardized and comprehensive metadata, such as isolation source and disease type, are essential to enable accurate interpretation of resistance trends and to facilitate global comparisons.

Distinct clonal complex/serotype combinations are associated with characteristic gene repertoires and span a continuum from flexibility to niche-specific specialization. These differences likely contribute to lineage-specific invasive potential across host groups, exemplified by the hypervirulence of CC17/III-2 in neonates and the comparatively low neonatal invasiveness of CC23/Ia and CC459/IV and the tendency to cause disease in adults. All identified gene patterns point to distinct evolutionary and ecological strategies among the groups. The enrichment of mobile genetic elements and

phage-associated genes suggests an enhanced genomic plasticity, likely reflecting frequent exposure to horizontal gene transfer. By contrast, II/22 shows a focused enrichment of siderophore and metal-uptake genes, indicating specialization for iron-limited host niches, while III-2/17 accumulates virulence and adhesion genes, especially the accessory SecA2/SecY2 system, pointing to enhanced host interaction and invasiveness. Beyond the canonical secretion (Sec) pathway, GBS carries an accessory SecA2/SecY2 secretion system. This alternative locus is apparently restricted to the hypervirulent ST17 lineage that encodes the related Srr2 adhesin. Srr2 is specific for ST17 strains and mediates high-affinity binding to fibrinogen and plasminogen, thereby enhancing adhesion, resistance to phagocytic killing and persistence in a murine meningitis model (74). The presence of these genes in group III-2/17 supports the reliability of our analysis. Together, these results indicate that even a small set of accessory genes may strongly influence the adaptation of GBS isolates.

Conclusion

The aim of this study was to assess the overall state of research on GBS. However, due to substantial gaps in the available metadata, a comprehensive evaluation is not possible. For instance, data from camelids seems entirely missing, and for human-derived isolates it is mostly unclear whether they originate from adult or neonatal cases. Large-scale comparative genomics, especially studies dissecting genetic variation and adaptive genome dynamics, provides key insights into multi-host pathogens like GBS. But the continuous generation of sequencing data without adequately structured and curated metadata limits their utility. The lack of standardized and comprehensive metadata represents a major obstacle in modern genomics and personalized medicine, restricting both biological insight and meaningful medical application. Safeguarding the integrity of data and metadata is not optional, it is essential, and compromising on data quality ultimately compromises science itself.

References

1. Le Doare K, Heath PT. An overview of global GBS epidemiology. Vaccine. 2013 Aug 28;31:D7–12.

2. Jones N, Oliver K, Jones Y, Haines A, Crook D. Carriage of group B streptococcus in pregnant women from Oxford, UK. J Clin Pathol. 2006 Apr 1;59(4):363–6.

3. Bergeron MG, Ke D, Ménard C, François FJ, Gagnon M, Bernier M, et al. Rapid Detection of Group B Streptococci in Pregnant Women at Delivery. N Engl J Med. 2000 Jul 20;343(3):175–9.

4. Manning SD, Neighbors K, Tallman PA, Gillespie B, Marrs CF, Borchardt SM, et al. Prevalence of Group B Streptococcus Colonization and Potential for Transmission by Casual Contact in Healthy Young Men and Women. Clin Infect Dis. 2004 Aug 1;39(3):380–8.

5. Sendi P, Johansson L, Norrby-Teglund A. Invasive Group B Streptococcal Disease in Non-pregnant Adults. Infection. 2008 Apr 1;36(2):100–11.

6. Danmallam FA, Pimenov NV. Study on prevalence, clinical presentation, and associated bacterial pathogens of goat mastitis in Bauchi, Plateau, and Edo states, Nigeria. Vet World. 2019 May;12(5):638–45.

7. Shi H, Zhou M, Zhang Z, Hu Y, Song S, Hui R, et al. Molecular epidemiology, drug resistance, and virulence gene analysis of Streptococcus agalactiae isolates from dairy goats in backyard farms in China. Front Cell Infect Microbiol. 2022;12:1049167.

8. Ozavci V, Dolgun HTY, Kirkan S, Seferoglu Y, Semen Z, Parin U. Evaluation of Streptococcus species isolated from subclinical sheep mastitis by molecular methods and determination of virulence factors and antimicrobial resistance genes. Vet Med (Praha). 2023 Sep;68(9):359–67.

9. ΖΔΡΑΓΚΑΣ Α, ΤΣΑΚΟΣ Ρ, Kotzamanidis C, ΑΝΑΤΟΛΙΩΤΗΣ Κ, ΤΣΑΚΝΑΚΗΣ Ι. Outbreak of mastitis in ewes caused by Streptococcus agalactiae. J Hell Vet Med Soc. 2017 Nov 29;56:114.

10. Crestani C, Seligsohn D, Forde TL, Zadoks RN. How GBS Got Its Hump: Genomic Analysis of Group B Streptococcus from Camels Identifies Host Restriction as well as Mobile Genetic Elements Shared across Hosts and Pathogens. Pathogens. 2022 Sep;11(9):1025.

11. Fenske L, Jauneikaite E, Getino M, Wan Y, Goesmann A, Eisenberg T. Evidence of a novel sublineage of Streptococcus agalactiae in elephants from zoo populations in Germany [Internet]. bioRxiv; 2025 [cited 2025 Mar 28]. p. 2025.03.17.642359. Available from: https://www.biorxiv.org/content/10.1101/2025.03.17.642359v1

12. Shuster KA, Hish GA, Selles LA, Chowdhury MA, Wiggins RC, Dysko RC, et al. Naturally Occurring Disseminated Group B Streptococcus Infections in Postnatal Rats. Comp Med. 2013 Feb;63(1):55–61.

13. Delannoy CMJ, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, et al. Human Streptococcus agalactiae strains in aquatic mammals and fish. BMC Microbiol. 2013 Feb 18;13:41.

14. Numberger D, Siebert U, Fulde M, Valentin-Weigand P. Streptococcal Infections in Marine Mammals. Microorganisms. 2021 Feb 10;9(2):350.

15. Richards VP, Velsko IM, Alam MT, Zadoks RN, Manning SD, Pavinski Bitar PD, et al. Population Gene Introgression and High Genome Plasticity for the Zoonotic Pathogen Streptococcus agalactiae. Mol Biol Evol. 2019 Nov;36(11):2572–90.

16. Wilkinson HW, Moody MD. Serological relationships of type I antigens of group B streptococci. J Bacteriol. 1969 Feb;97(2):629–34.

17. Kong F, Gowan S, Martin D, James G, Gilbert GL. Serotype identification of group B streptococci by PCR and sequencing. J Clin Microbiol. 2002 Jan;40(1):216–26.

18. Bianchi-Jassir F, Paul P, To KN, Carreras-Abad C, Seale AC, Jauneikaite E, et al. Systematic review of Group B Streptococcal capsular types, sequence types and surface proteins as potential vaccine candidates. Vaccine. 2020 Oct 7;38(43):6682–94.

19. Teatero S, Ramoutar E, McGeer A, Li A, Melano RG, Wasserscheid J, et al. Clonal Complex 17 Group B Streptococcus strains causing invasive disease in neonates and adults originate from the same genetic pool. Sci Rep. 2016 Feb 4;6:20047.

20. Jamrozy D, Bijlsma MW, de Goffau MC, van de Beek D, Kuijpers TW, Parkhill J, et al. Increasing incidence of group B streptococcus neonatal infections in the Netherlands is associated with clonal expansion of CC17 and CC23. Sci Rep. 2020 Jun 12;10(1):9539.

21. Hsu JF, Tsai MH, Lin LC, Chu SM, Lai MY, Huang HR, et al. Genomic Characterization of Serotype III/ST-17 Group B Streptococcus Strains with Antimicrobial Resistance Using Whole Genome Sequencing. Biomedicines. 2021 Oct 15;9(10):1477.

22. Manning SD, Lewis MA, Springman AC, Lehotzky E, Whittam TS, Davies D. Genotypic Diversity and Serotype Distribution of Group B Streptococcus Isolated from Women Before and After Delivery. Clin Infect Dis. 2008 Jun 15;46(12):1829–37.

23. Teatero S, Ferrieri P, Martin I, Demczuk W, McGeer A, Fittipaldi N. Serotype Distribution, Population Structure, and Antimicrobial Resistance of Group B Streptococcus Strains Recovered from Colonized Pregnant Women. J Clin Microbiol. 2017 Feb;55(2):412–22.

24. Hayes K, O'Halloran F, Cotter L. A review of antibiotic resistance in Group B Streptococcus: the story so far. Crit Rev Microbiol. 2020 May 3;46(3):253–69.

25. Oliveira LMA, Simões LC, Costa NS, Zadoks RN, Pinto TCA. The landscape of antimicrobial resistance in the neonatal and multi-host pathogen group B Streptococcus: review from a One Health perspective. Front Microbiol [Internet]. 2022 Jul 28 [cited 2025 Apr 1];13. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.943413/full

26. Fenske L, Jelonek L, Goesmann A, Schwengers O. BakRep – a searchable large-scale web repository for bacterial genomes, characterizations and metadata. Microb Genomics. 2024;10(10):001305.

27. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. 2021 Nov;19(11):e3001421.

28. Hunt M, Lima L, Shen W, Lees J, Iqbal Z. AllTheBacteria - all bacterial genomes assembled, available and searchable [Internet]. bioRxiv; 2024 [cited 2025 Apr 30]. p. 2024.03.08.584059. Available from: https://www.biorxiv.org/content/10.1101/2024.03.08.584059v1

29. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018 Dec;19(1):307.

30. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019 Aug;63(11).

31. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genomics [Internet]. 2021 Nov 5 [cited 2022 Apr 5];7(11). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000685

32. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020 Jul 22;21(1):180.

33. Roder T, Pimentel G, Fuchsmann P, Stern MT, von Ah U, Vergères G, et al. Scoary2: rapid association of phenotypic multi-omics data with microbial pan-genomes. Genome Biol. 2024 Apr 11;25(1):93.

34. Iannelli F, Santoro F, Santagati M, Docquier JD, Lazzeri E, Pastore G, et al. Type M Resistance to Macrolides Is Due to a Two-Gene Efflux Transport System of the ATP-Binding Cassette (ABC) Superfamily. Front Microbiol [Internet]. 2018 Jul 31 [cited 2025 Sep 26];9. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.01670/full

35. Franken C, Haase G, Brandt C, Weber-Heynemann J, Martin S, Lämmler C, et al. Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing scpB and lmb. Mol Microbiol. 2001;41(4):925–35.

36. Crestani C. The role of the mobilome in the evolution and host-adaptation of Streptococcus agalactiae. 2021.

37. Marktgröße für klinische Studien, Anteil, Trends | Wachstumsbericht [2032] [Internet]. [cited 2025 Oct 14]. Available from: https://www.fortunebusinessinsights.com/clinical-trials-market-106930

38. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep;437(7057):376–80.

39. ENA [Internet]. [cited 2025 Sep 26]. Spatiotemporal Metadata Standards — ENA Documentation 1 documentation. Available from: https://ena-docs.readthedocs.io/en/latest/faq/spatiotemporal-metadata.html

40. ENA Browser [Internet]. [cited 2025 Sep 26]. Available from: https://www.ebi.ac.uk/ena/browser/view/ERC000028

41. Price E, Feyertag F, Evans T, Miskin J, Mitrophanous K, Dikicioglu D. What is the real value of omics data? Enhancing research outcomes and securing long-term data excellenc. Nucleic Acids Res. 2024 Nov 11;52(20):12130–40.

42. Wilson SL, Way GP, Bittremieux W, Armache JP, Haendel MA, Hoffman MM. Sharing biological data: why, when, and how. FEBS Lett. 2021 Apr;595(7):847–63.

43. Davies HD, Jones N, Whittam TS, Elsayed S, Bisharat N, Baker CJ. Multilocus Sequence Typing of Serotype III Group B Streptococcus and Correlation with Pathogenic Potential. J Infect Dis. 2004 Mar 15;189(6):1097–102.

44. Alhhazmi A, Hurteau D, Tyrrell GJ. Epidemiology of Invasive Group B Streptococcal Disease in Alberta, Canada, from 2003 to 2013. J Clin Microbiol. 2016 Jul;54(7):1774–81.

45. Farley MM, Strasbaugh LJ. Group B Streptococcal Disease in Nonpregnant Adults. Clin Infect Dis. 2001 Aug 15;33(4):556–61.

46. Slotved HC, Kong F, Lambertsen L, Sauer S, Gilbert GL. Serotype IX, a Proposed New Streptococcus agalactiae Serotype. J Clin Microbiol. 2007 Sep;45(9):2929–36.

47. Shabayek S, Spellerberg B. Group B Streptococcal Colonization, Molecular Characteristics, and Epidemiology. Front Microbiol. 2018 Mar 14;9:437.

48. Bellais S, Six A, Fouet A, Longo M, Dmytruk N, Glaser P, et al. Capsular Switching in Group B Streptococcus CC17 Hypervirulent Clone: A Future Challenge for Polysaccharide Vaccine Development. J Infect Dis. 2012 Dec 1;206(11):1745–52.

49. Meehan M, Cunney R, Cafferkey M. Molecular epidemiology of group B streptococci in Ireland reveals a diverse population with evidence of capsular switching. Eur J Clin Microbiol Infect Dis. 2014 Jul 1;33(7):1155–62.

50. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, et al. Multilocus sequence typing system for group B streptococcus. J Clin Microbiol. 2003 Jun;41(6):2530–6.

51. Springman AC, Lacher DW, Waymire EA, Wengert SL, Singh P, Zadoks RN, et al. Pilus distribution among lineages of group b streptococcus: an evolutionary and clinical perspective. BMC Microbiol. 2014 Jun 19;14(1):159.

52. Tong Z, Kong F, Wang B, Zeng X, Gilbert GL. A practical method for subtyping of Streptococcus agalactiae serotype III, of human origin, using rolling circle amplification. J Microbiol Methods. 2007 Jul;70(1):39–44.

53. Kao Y, Tsai MH, Lai MY, Chu SM, Huang HR, Chiang MC, et al. Emerging serotype III sequence type 17 group B streptococcus invasive infection in infants: the clinical characteristics and impacts on outcomes. BMC Infect Dis. 2019 Jun 19;19:538.

54. Tsai MH, Hsu JF, Lai MY, Lin LC, Chu SM, Huang HR, et al. Molecular Characteristics and Antimicrobial Resistance of Group B Streptococcus Strains Causing Invasive Disease in Neonates and Adults. Front Microbiol. 2019 Feb 18;10:264.

55. Kong F, Martin D, James G, Gilbert GL. Towards a genotyping system for Streptococcus agalactiae (group B streptococcus): use of mobile genetic elements in Australasian invasive isolates. J Med Microbiol. 2003 Apr 1;52(4):337–44.

56. Sun Y, Kong F, Zhao Z, Gilbert GL. Comparison of a 3-Set Genotyping System with Multilocus Sequence Typing for Streptococcus agalactiae (Group B Streptococcus). J Clin Microbiol. 2005 Sep;43(9):4704–7.

57. Cubria MB, Vega LA, Shropshire WC, Sanson MA, Shah BJ, Regmi S, et al. Population Genomics Reveals Distinct Temporal Association with the Emergence of ST1 Serotype V Group B Streptococcus and Macrolide Resistance in North America. Antimicrob Agents Chemother [Internet]. 2021 Oct 11 [cited 2025 Sep 30]; Available from: https://journals.asm.org/doi/10.1128/AAC.00714-21

58. Teatero S, Athey TBT, Van Caeseele P, Horsman G, Alexander DC, Melano RG, et al. Emergence of Serotype IV Group B Streptococcus Adult Invasive Disease in Manitoba and Saskatchewan, Canada, Is Driven by Clonal Sequence Type 459 Strains. J Clin Microbiol. 2015 Sep;53(9):2919–26.

59. Salloum M, Mee-Marquet N van der, Valentin-Domelier AS, Quentin R. Diversity of Prophage DNA Regions of Streptococcus agalactiae Clonal Lineages from Adults and Neonates with Invasive Infectious Disease. PLOS ONE. 2011 May 25;6(5):e20256.

60. Flores AR, Galloway-Peña J, Sahasrabhojane P, Saldaña M, Yao H, Su X, et al. Sequence type 1 group B Streptococcus, an emerging cause of invasive disease in adults, evolves by small genetic changes. Proc Natl Acad Sci U S A. 2015 May 19;112(20):6431–6.

61. Jamrozy D, Gopal Rao G, Feltwell T, Lamagni T, Khanna P, Efstratiou A, et al. Population genetics of group B Streptococcus from maternal carriage in an ethnically diverse community in London. Front Microbiol. 2023 May 18;14:1185753.

62. Ntozini B, Walaza S, Metcalf B, Hazelhurst S, de Gouveia L, Meiring S, et al. Molecular Epidemiology of Invasive Group B Streptococcus in South Africa, 2019–2020. J Infect Dis. 2025 Apr 15;231(4):e697–707.

63. Wadilo F, Hailemeskel E, Kedir K, El-Khatib Z, Asogba PC, Seyoum T, et al. Prevalence of Group B Streptococcus maternal colonization, serotype distribution, and antimicrobial resistance in Sub-Saharan Africa: A systematic review and meta-analysis. J Glob Antimicrob Resist. 2023 Mar 1;32:134–44.

64. Chukwu MO, Mavenyengwa RT, Monyama CM, Bolukaoto JY, Lebelo SL, Maloba MR, et al. Antigenic distribution of Streptococcus agalactiae isolates from pregnant women at Garankuwa hospital - South Africa. Germs. 2015 Dec;5(4):125–33.

65. Ip M, Ang I, Fung K, Liyanapathirana V, Luo MJ, Lai R. Hypervirulent Clone of Group B *Streptococcus* Serotype III Sequence Type 283, Hong Kong, 1993–2012. Emerg Infect Dis. 2016 Oct;22(10):1800–3.

66. Zhang L, Kang WJ, Zhu L, Xu LJ, Guo C, Zhang XH, et al. Emergence of Invasive Serotype Ib Sequence Type 10 Group B Streptococcus Disease in Chinese Infants Is Driven by a Tetracycline-Sensitive Clone. Front Cell Infect Microbiol. 2021 May 14;11:642455.

67. Silva MM, Silva ÉA, Oliveira CNT, Santos MLC, Souza CL, de Melo FF, et al. Distribution and Prevalence of Serotypes of Group B Streptococcus Isolated from Pregnant Women in 30 Countries: A Systematic Review. Matern-Fetal Med. 2023 Apr 25;05(02):97–103.

68. Khan UB, Portal EAR, Sands K, Lo S, Chalker VJ, Jauneikaite E, et al. Genomic Analysis Reveals New Integrative Conjugal Elements and Transposons in GBS Conferring Antimicrobial Resistance. Antibiot Basel Switz. 2023 Mar 9;12(3):544.

69. Bonofiglio L, Gagetti P, García Gabarrot G, Kaufman S, Mollerach M, Toresani I, et al. Susceptibility to β-lactams in β-hemolytic streptococci. Rev Argent Microbiol Argent J Microbiol. 2018 Oct 1;50(4):431–5.

70. Hsu CY, Moradkasani S, Suliman M, Uthirapathy S, Zwamel AH, Hjazi A, et al. Global patterns of antibiotic resistance in group B Streptococcus: a systematic review and meta-analysis. Front Microbiol [Internet]. 2025 Apr 16 [cited 2025 Oct 6];16. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2025.1541524/full

71. Gergova R, Boyanov V, Muhtarova A, Alexandrova A. A Review of the Impact of Streptococcal Infections and Antimicrobial Resistance on Human Health. Antibiotics. 2024 Apr 15;13(4):360.

72. Da Cunha V, Davies MR, Douarre PE, Rosinski-Chupin I, Margarit I, Spinali S, et al. Streptococcus agalactiae clones infecting humans were selected and fixed through the extensive use of tetracycline. Nat Commun. 2014 Aug 4;5:4544.

73. Pinto TCA, Costa NS, Corrêa AB de A, de Oliveira ICM, de Mattos MC, Rosado AS, et al. Conjugative transfer of resistance determinants among human and bovine Streptococcus agalactiae. Braz J Microbiol. 2014 Oct 9;45(3):785–9.

74. Six A, Bellais S, Bouaboud A, Fouet A, Gabriel C, Tazi A, et al. Srr2, a multifaceted adhesin expressed by ST-17 hypervirulent Group B Streptococcus involved in binding to both fibrinogen and plasminogen. Mol Microbiol. 2015;97(6):1209–22.

7.2. Co-authorships in peer-reviewed publications

Metabolism of the Genus *Guyparkeria* Revealed by Pangenome Analysis
Maggie C. Y. Lau Vetter, Baowei Huang, **Linda Fenske**, Jochen Blom
Microorganism (2022), DOI: 10.3390/microorganisms10040724

Four new members of the family *Cytophagaceae*: *Chryseosolibacter histidini* gen. nov., sp. nov., *Chryseosolibacter indicus* gen. nov., sp. nov., *Dawidia cretensis*, gen. nov., sp. nov., and *Dawidia soli*, gen. nov., sp. nov. isolated from diverse habitat
Senlie Octaviana, Stefan Lorenczyk, Frederike Ackert, **Linda Fenske**, Joachim Wink
Antonie van Leeuwenhoek (2022), DOI: 10.1007/s10482-022-01756-2

Development and validation of a triplex real-time qPCR for sensitive detection and quantification of major rat bite fever pathogen *Streptobacillus moniliformis*
Ahmad Fawzy, Ann-Sophie Giel, **Linda Fenske**, Alexa Bach, Christiane Herden, Katharina Engel, Elisa Heuser, Marc Boelhauve, Rainer G. Ulrich, Klaus Vogel, Katja Schmidt, Tobias Eisenberg
Journal of Microbiological Methods (2022), DOI: 10.1016/j.mimet.2022.106525

Genome-Based Retrospective Analysis of a *Providencia stuartii* Outbreak in Rome, Italy: Broad Spectrum IncC Plasmids Spread the NDM Carbapenemase within the Hospital

Valerio Capitani, Gabriele Arcari, Alessandra Oliva, Federica Sacco, Gaia Menichincheri, **Linda Fenske**, Riccardo Polani, Giammarco Raponi, Guido Antonelli, Alessandra Carattoli

Antibiotics (2023), DOI: 10.3390/antibiotics12050943

Race-specific genotypes of *Pseudomonas syringae pv. tomato* are defined by the presence of mobile DNA elements within the genome

Benedetta Orfei, Joël F. Porthier, **Linda Fenske**, Jochen Blom, Chiaraluce Moretti, Roberto Buonaurio, Theo H. M. Smits

Frontiers in Plan Science (2023), DOI: 10.3389/fpls.2023.1197706

Comprehensive genomic analysis of *Bacillus paralicheniformis* strain BP9, pan-genomic and genetic basis of biocontrol mechanism

Muhammad Asif, Zhang Li-Qun, Qingchao Zeng, Muhammad Atiq, Khalil Ahmad, Aqil Tariq, Nadhir Al-Ansari, Jochen Blom, **Linda Fenske**, Hissah Abdulrahman Alodaini, Ashraf Atef Hatamleh

Computational and Structural Biotechnology (2023),
DOI: 10.1016/j.csbj.2023.09.043

Whole-genome draft assemblies of *Paracoccus pantotrophus* DSM 11073 and *Paracoccus* sp. AS002: Phylogenetics entails classification as *Paracoccus versutus* AS002

Upasana Pan, Denise Bachmann, **Linda Fenske**, Lars Mathias Blank, Till Tiso

Elsevier Journal of Bioscience and Bioengineering (2025),
DOI: 10.1016/j.jbiosc.2025.09.003

Ecological ubiquity and phylogeny drive nestedness in phages–bacteria networks and shape the bacterial defensome
Chloé Feltin, Sylvain Piry, Benoit Moury, Lola Chateau, Karine Berthier, Jonathan M. Jacobs, Jules Butchacas, **Linda Fenske**, Lillian Ebeling-Koning, Theo H. M. Smits, Cindy E. Morris, Clara Torres-Barceló
PLOS Pathogens (2025), DOI: 10.1371/journal.ppat.1013428

Novel bacterium *Enterocloster sp.* M3 promotes colorectal tumorigenesis via the production of the carcinogen styrene
Yao Zeng, Yao Huang, Silin Ye, Effie Yin Tung Lau, Man Chun Chiu, **Linda Fenske**, Yuting Sun, Liting Jiang, Jiangying Chen, Yanqing Huang, Tingyu Zhou, Jiawei Lu, Jie Zhou, Shu Zheng, Francis Ka Leung Chan, Jessie Qiaoyi Liang
Gut Microbes (2026), DOI: 10.1080/19490976.2026.2630481

Completion of the *Rhizobium favelukesii* LPU83 genome reveals a highly plastic symbiotic plasmid unable to facilitate efficient nitrogen fixation
Abril Luchetti, Catalina D'Addona, Lucas Gabriel Castellani, María Delfina Cabrera, Daniel Wibberg, Carolina Vacca, **Linda Fenske**, Jochen Blom, Anika Winkler, Tobias Busche, Christian Rückert, Jörn Kalinowski, Andreas Schlüter, Alfred Pühler, Karsten Niehaus, Antonio Lagares, María Florencia Del Papa, Mariano Pistorio, Gonzalo Torres Tejerizo
Agronomy (2026), DOI: 10.3390/agronomy16050523

7.3. Further contributions

Scientific conferences

EDGAR User Meeting 2021, online:
Comparative genomic analysis of *Streptococcus uberis* in dairy cows with clinical and subclinical mastitis (Oral presentation)

40. AVID-Tagung "Bakteriologie" 2022, Kloster Banz:
A dominant clonal lineage of *Streptococcus uberis* in cattle in Germany
(Oral presentation)

VAAM 2023, Göttingen:
Evaluation of the zoonotic potential of *Streotococcus agalactiae* (Poster presentation)

BakRep: A searchable web repository for bacterial genomes and standardized characterizations (Poster presentation)

7. Joint Conference of the DHM & VAAM 2024, Würzburg:
BakRep - A searchable large-scale web repository for bacterial genomes, characterizations and metadata (Poster presentation)

42. AVID-Tagung "Bakteriologie" 2024, Kloster Banz:
Streptococcus agalactiae in Elefanten (Oral presentation)

BakRep - A searchable large-scale web repository for bacterial genomes, characterizations and metadata (Oral presentation)

VAAM 2025, Bochum:
Streptococcus agalactiae in elephants - a comparative genomic analysis
(Poster presentation)

BakRep v2 - Database update (Poster presentation)

Supervised student projects

B.Sc. Moritz Hobein, 2021:
Ausarbeitung eines Skripts zum Erstellen von CVL-Plots

M.Sc. Naoul Eldjouher, 2025:
Design and Development of an automated and scalable pipeline for bacterial analysis using Nextflow

Hochdurchsatzdatenanalyse Teil 2, 2022-2025 (anually):
Bioinformatik im klinischen Alltag

A. List of commands for bioinformatic analyses

A.1. BakRep workflow

The nextflow-workflow used for data characterization of all genomes currently contained in the web repository is available at: `github.com/ag-computational-bio/bakrep`.

A.2. Processing of the elephant-derived GBS

List of commands for the processing of the raw-reads of the elephant isolates.

```
# Trimming and quality-control with fastp
fastp --in1 sample_forward.fastq-gz --in2 sample_reverse.fastq.gz
--out1 sample_R1.fastq.gz --out2 sample_R2.fastq.gz
--unpaired1 sample_SE.fastq.gz --unpaired2 sample_SE.fastq.gz
--detect_adapter_for_pe --trim_poly_g --cut_front --cut_tail
--length_required 21 --low_complexity_filter --correction


# Quality control with fastqc and multiqc
for file in fastqs/; do fastqc $file --outdir fastqc/; done
multiqc fastqc/ --outdir multiqc/
```

```
# Assembly with Unicycler and Trycycler

## short-reads
unicycler --short1 sample_R1.fastq.gz --short2 sample_R2.fastq.gz
--unpaired sample_SE.fastq.gz --out sample.fasta

## hybrid-reads
unicycler --short1 sample_R1.fastq.gz --short2 sample_R2.fastq.gz
--long sample_long.fastq.gz --out sample.fasta
```

Long-read assembly and polishing was performed following the guides for bacterial genome assembly by Ryan Wick [171, 172].

```
# Annotation of the assemblies with Bakta
bakta --db baktadb/ --keep-contig-headers --threads 8 --prefix sample
--genus Streptococcus --species agalactiae --strain sample

# Estimation of completeness and contamination with CheckM2
checkm2 predict --threads 8 --input fastas/elephants/
--output_directory checkm2/elephants/

# Taxonomic classification with GTDB-tk
gtdbtk classify_wf --genome_dir fastas/elephants/
--out_dir gtdbtk/elephants/ --prefix elephants
--extension fasta --cpus 1 --mash_db /gtdbtk

# Subtyping using mlst
for file in /fastas/elephants/*.fasta;
do mlst $file --csv --outfile sample.csv;
done
```

```
# Target-free split k-mer analysis with SKA
for file in /fastas/elephants/*.fasta;
do ska fasta -o /ska/elephants/sample $file;
done

ska summary *.skf > ska_summary.txt
ska distance -s 10 -i 0.9 -o /ska/elephants/ *.skf

# AMR detection with AMRFinderPlus
amrfinder -n /fastas/elephants/sample.fasta -o sample_amrfinder.tsv

# Searching for plasmids with MOB-suite
mob_recon --infile fastas/elephants/sample.fasta
--outdir /plasmids/elephants/

# Searching for capsular-serotypes with GBS-SBG, srst2 and KMA
srst2 --input_pe sample_R1.fastq.gz sample_R2.fastq.gz
--output sample --log --gene_db GBS-SBG/GBS-SBG.fasta

srst2 --prev_output *results.txt --output samples

/kma/kma -ipe sample_R1.fasta sample_R2.fasta -o /kma_out/elephants/

# GWAS with panaroo and scoary2
panaroo -i sample.gff3 --remove-invalid-genes --clean-mode sensitive
--refind-mode off --threads 12 -o /panaroo/elephants/

scoary2 --genes /panaroo/elephants/gene_presence_absence.csv
--gene-data-type 'gene-list:,' --traits trait.txt
--outdir /scoary2/elephants/
```

A.3. Processing of all GBS genomes in BakRep

List of commands for processing all GBS genomes contained in BakRep.

```
# Searching for capsular-serotypes with KMA
for fasta in bakrep_gbs/fasta/*.fasta;
do kma/kma/ -i $fasta -o sample_kma
-t_db /kma/database/GBS-SBG

# AMR detection with AMRFinderPlus
find bakrep_gbs/fasta/ -name *.fna
| xargs -n1 -P 500 -I bash -c 'file=¨¨;
base=$(basename $̈file¨ .ffn);
amrfinder -n $̈file¨ -o ̈/amrfinder/bakrep_gbs/$base.tsv¨'"

# GWAS with panaroo and scoary2
panaroo -i bakrep_gbs/gff3/batch1/*.gff3 --remove-invalid-genes
--clean-mode strict --refind-mode off --threads 12
-o /panaroo/elephants/batch1
(...)

panaroo-merge -d /panaroo/bakrep_split/batch1
/panaroo/bakrep_split/batch2 (...)
-o panaroo/bakrep_split/merge
```

```
## Filter panaroo output to keep genes present in
more than 0.25% and less than 95% of genomes

awk -F',' 'NR==1print; nextp=0; for(i=4;i<=NF;i++) if($i!="") p++;
f=p/(NF-3); if(f>=0.0025 && f<=0.95) print'
gene_presence_absence.csv > gene_presence_absence_filtered00025095.csv

scoary2 --genes gene_presence_absence_filtered00025095.csv
 -gene-data-type 'gene-list:,' --traits trait.txt
 --outdir scoary2/bakrep_gbs/
```

References

[1] American Academy of Microbiology. The Microbial World: Foundation of the Biosphere. 1997 (cited on page 1).

[2] Wu, D., Seshadri, R., Kyrpides, N. C., and Ivanova, N. N.
"A metagenomic perspective on the microbial prokaryotic genome census".
Science Advances 11.3 (2025). DOI: 10.1126/sciadv.adq2166 (cited on page 1).

[3] Fraser, C. M., Eisen, J. A., and Salzberg, S. L. "Microbial genome sequencing".
Nature 406.6797 (2000). DOI: 10.1038/35021244 (cited on pages 1, 56).

[4] Zhang, Z., Wang, J., Wang, J., Wang, J., and Li, Y.
"Estimate of the sequenced proportion of the global prokaryotic genome".
Microbiome 8.1 (2020). DOI: 10.1186/s40168-020-00903-z (cited on page 1).

[5] Ng, D. "Culturing the Unculturable: Working with Difficult Bacteria" (2025)
(cited on page 1).

[6] Vartoukian, S. R., Palmer, R. M., and Wade, W. G.
"Strategies for culture of 'unculturable' bacteria". *FEMS Microbiology Letters* 309.1 (2010).
DOI: 10.1111/j.1574-6968.2010.02000.x (cited on page 1).

[7] Deurenberg, R. H., Bathoorn, E., Chlebowicz, M. A., Couto, N., Ferdous, M.,
García-Cobos, S., Kooistra-Smid, A. M. D., Raangs, E. C., Rosema, S., Veloo, A. C. M.,
et al.
"Application of next generation sequencing in clinical microbiology and infection prevention".
Journal of Biotechnology 243 (2017). DOI: 10.1016/j.jbiotec.2016.12.022
(cited on pages 1, 14).

[8] Venkova, T., Yeo, C. C., and Espinosa, M.
"Editorial: The Good, The Bad, and The Ugly: Multiple Roles of Bacteria in Human Life".
Frontiers in Microbiology 9 (2018). DOI: 10.3389/fmicb.2018.01702 (cited on page 1).

[9] Hunt, M., Lima, L., Anderson, D., Bouras, G., Hall, M., Hawkey, J., Schwengers, O.,
Shen, W., Lees, J. A., and Iqbal, Z.
"AllTheBacteria – all bacterial genomes assembled, available, and searchable".
bioRxiv (2024). DOI: 10.1101/2024.03.08.584059 (cited on pages 2, 30, 38).

[10] Bartlett, A., Padfield, D., Lear, L., Bendall, R., and Vos, M.
"A comprehensive list of bacterial pathogens infecting humans". *Microbiology* 168.12 (2022).
DOI: 10.1099/mic.0.001269 (cited on page 2).

[11] Soni, J., Sinha, S., and Pandey, R.
"Understanding bacterial pathogenicity: a closer look at the journey of harmful microbes".
Frontiers in Microbiology 15 (2024). DOI: 10.3389/fmicb.2024.1370818 (cited on page 2).

[12] Krzyściak, W., Pluskwa, K. K., Jurczak, A., and Kościelniak, D.
"The pathogenicity of the Streptococcus genus".
European Journal of Clinical Microbiology & Infectious Diseases 32.11 (2013).
DOI: 10.1007/s10096-013-1914-9 (cited on page 2).

[13] Dijk, E. L. van, Auger, H., Jaszczyszyn, Y., and Thermes, C.
"Ten years of next-generation sequencing technology". *Trends in Genetics* 30.9 (2014). DOI: 10.1016/j.tig.2014.07.001 (cited on page 2).

[14] Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N.
"A brief history of bioinformatics". *Briefings in Bioinformatics* 20.6 (2019). DOI: 10.1093/bib/bby063 (cited on page 2).

[15] Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms". *Proceedings of the National Academy of Sciences of the United States of America* 95.6 (1998). DOI: 10.1073/pnas.95.6.3140 (cited on page 2).

[16] Urwin, R. and Maiden, M. C. J.
"Multi-locus sequence typing: a tool for global epidemiology".
Trends in Microbiology 11.10 (2003). DOI: 10.1016/j.tim.2003.08.006 (cited on page 2).

[17] Maiden, M. C. "Multilocus Sequence Typing of Bacteria".
Annual Review of Microbiology 60.1 (2006).
DOI: 10.1146/annurev.micro.59.030804.121325 (cited on page 2).

[18] Baker, S., Thomson, N., Weill, F.-X., and Holt, K. E. "Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens".
Science (New York, N.y.) 360.6390 (2017). DOI: 10.1126/science.aar3777
(cited on page 2).

[19] Blackwell, G. A., Hunt, M., Malone, K. M., Lima, L., Horesh, G., Alako, B. T. F., Thomson, N. R., and Iqbal, Z. "Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences". *PLoS biology* 19.11 (2021).
DOI: 10.1371/journal.pbio.3001421 (cited on pages 2 sq., 14–16, 19, 29, 31).

[20] Albright, S. and Louca, S. "Trait biases in microbial reference genomes".
Scientific Data 10.1 (2023). DOI: 10.1038/s41597-023-01994-7 (cited on page 2).

[21] Lagkouvardos, I., Overmann, J., and Clavel, T.
"Cultured microbes represent a substantial fraction of the human and mouse gut microbiota".
Gut Microbes 8.5 (2017). DOI: 10.1080/19490976.2017.1320468 (cited on page 2).

[22] Jensen, P. A.
"Ten species comprise half of the bacteriology literature, leaving most species unstudied" (2025). DOI: 10.1101/2025.01.04.631297 (cited on pages 3, 56).

[23] Alhhazmi, A., Hurteau, D., and Tyrrell, G. J. "Epidemiology of Invasive Group B Streptococcal Disease in Alberta, Canada, from 2003 to 2013".
Journal of Clinical Microbiology 54.7 (2016). DOI: 10.1128/JCM.00355-16
(cited on pages 3, 11).

[24] Kumar, G. C., Chaudhary, J., Meena, L. K., Meena, A. L., and Kumar, A.
"Function-driven microbial genomics for ecofriendly agriculture".
Microbes in Land Use Change Management. Elsevier, 2021.
DOI: 10.1016/B978-0-12-824448-7.00021-8 (cited on page 3).

[25] Gergova, R., Boyanov, V., Muhtarova, A., and Alexandrova, A. "A Review of the Impact of Streptococcal Infections and Antimicrobial Resistance on Human Health".
Antibiotics 13.4 (2024). DOI: 10.3390/antibiotics13040360 (cited on pages 5, 51).

[26] RKI. Inzidenzanstieg von Gruppe-A-Streptokokken-Infektionen. URL: https://www.rki.de/DE/Themen/Infektionskrankheiten/Infektionskrankheiten-A-Z/S/Scharlach/invasive_Gruppe-A-Streptokokken-Infektionen.html (visited on 10/16/2025) (cited on page 5).

[27] Pneumokokken-Infektionen in Deutschland. Statista. URL: https://de.statista.com/statistik/daten/studie/1471992/umfrage/pneumokokken-infektionen-in-deutschland/ (visited on 12/16/2025) (cited on page 5).

[28] Terukina, A. Rare tissue-damaging bacteria spreads in Japan. The Japan Times. 2024. URL: https://www.japantimes.co.jp/news/2024/06/15/japan/science-health/stss-japan-spread/ (visited on 01/26/2026) (cited on page 5).

[29] Bradley, A. J. "Bovine Mastitis: An Evolving Disease". *The Veterinary Journal* 164.2 (2002). DOI: 10.1053/tvjl.2002.0724 (cited on page 5).

[30] Kabelitz, T., Kashongwe, O. B., Doherr, M., Nübel, U., Ammon, C., Boloña, P. S., Keane, O., Amon, T., and Amon, B. "Occurrence, treatment and pathogens involved in mastitis on a commercial German dairy farm: A retrospective study from 2012 to 2021". *Journal of Advanced Veterinary and Animal Research* 11.4 (2024). DOI: 10.5455/javar.2024.k837 (cited on page 5).

[31] Whiley, R. A. and Hardie, J. M. "*Streptococcus*". *Bergey's Manual of Systematics of Archaea and Bacteria*. Chichester, UK: John Wiley & Sons, Ltd, 2015. DOI: 10.1002/9781118960608.gbm00612 (cited on pages 5–8).

[32] Abranches, J., Zeng, L., Kajfasz, J. K., Palmer, S. R., Chakraborty, B., Wen, Z. T., Richards, V. P., Brady, L. J., and Lemos, J. A. "Biology of Oral Streptococci". *Microbiology Spectrum* 6.5 (2018). DOI: 10.1128/microbiolspec.gpp3-0042-2018 (cited on pages 5, 7).

[33] Cole, J. N., Henningham, A., Gillen, C. M., Ramachandran, V., and Walker, M. J. "Human pathogenic streptococcal proteomics and vaccine development". *Proteomics – Clinical Applications* 2.3 (2008). DOI: 10.1002/prca.200780048 (cited on pages 5 sq.).

[34] LPSN. Genus Streptococcus. LPSN.dsmz.de. 2026. URL: https://lpsn.dsmz.de/genus/streptococcus (cited on page 6).

[35] Fulde, M. "Virulenzmechanismen zoonotischer Streptokokken". 2015 (cited on page 6).

[36] Fulde, M. and Valentin-Weigand, P. "Epidemiology and Pathogenicity of Zoonotic Streptococci". *Host-Pathogen Interactions in Streptococcal Diseases* 368 (2012). DOI: 10.1007/82_2012_277 (cited on page 6).

[37] Spellerberg, B. and Brandt, C. "Streptococcus". *Manual of Clinical Microbiology*. John Wiley & Sons, Ltd, 2015. DOI: 10.1128/9781555817381.ch22 (cited on pages 6, 9).

[38] Maeda, T., Tsuyuki, Y., Fujita, T., Fukushima, Y., Goto, M., Yoshida, H., and Takahashi, T. "Comparison of *Streptococcus agalactiae* Isolates from Humans and Companion Animals Reveals Genotypic and Phenotypic Differences". *Japanese Journal of Infectious Diseases* 73.4 (2020). DOI: 10.7883/yoken.JJID.2019.441 (cited on page 6).

[39] Glajzner, P., Szewczyk, E. M., and Szemraj, M.
"Phenotypic and Genotypic Characterization of Antimicrobial Resistance in Streptococci Isolated from Human and Animal Clinical Specimens". *Current Microbiology* 80.7 (2023). DOI: 10.1007/s00284-023-03337-6 (cited on page 6).

[40] Crestani, C., Forde, T. L., Bell, J., Lycett, S. J., Oliveira, L. M. A., Pinto, T. C. A., Cobo-Ángel, C. G., Ceballos-Márquez, A., Phuoc, N. N., Sirimanapong, W., et al.
"Genomic and functional determinants of host spectrum in Group B Streptococcus". *PLOS Pathogens* 20.8 (2024). DOI: 10.1371/journal.ppat.1012400 (cited on pages 6, 44).

[41] Stoneham, S., Peters, J., and Price, J. "Staphylococcal and streptococcal infections". *Medicine*. Infections Part 3 of 3 49.12 (2021). DOI: 10.1016/j.mpmed.2021.09.001 (cited on page 6).

[42] Lancefield, R. C.
"A Serological Differentiation of Human and Other Groups of Hemolytic Streptococci". *The Journal of Experimental Medicine* 57.4 (1933) (cited on page 6).

[43] J. H. Jorgensen, K. C. Carroll, G. Funke, M. A. Pfaller, M. L. Landry, S. S. Richter, and D. W. Warnock, editors. *Manual of Clinical Microbiology*.
Washington, DC, USA: ASM Press, 2015. DOI: 10.1128/9781555817381 (cited on page 6).

[44] Menon, T. "Understanding the viridians group streptococci: Are we there yet?" *Indian Journal of Medical Microbiology* 34.4 (2016). DOI: 10.4103/0255-0857.195371 (cited on page 6).

[45] Richards, V. P., Palmer, S. R., Pavinski Bitar, P. D., Qin, X., Weinstock, G. M., Highlander, S. K., Town, C. D., Burne, R. A., and Stanhope, M. J.
"Phylogenomics and the Dynamic Genome Evolution of the Genus Streptococcus". *Genome Biology and Evolution* 6.4 (2014). DOI: 10.1093/gbe/evu048 (cited on page 7).

[46] Hogeveen, H., Huijps, K., and Lam, T. "Economic aspects of mastitis: New developments". *New Zealand Veterinary Journal* 59.1 (2011). DOI: 10.1080/00480169.2011.547165 (cited on page 8).

[47] Zadoks, R. N., Gillespie, B. E., Barkema, H. W., Sampimon, O. C., Oliver, S. P., and Schukken, Y. H. "Clinical, epidemiological and molecular characteristics of *Streptococcus uberis* infections in dairy herds". *Epidemiology and Infection* 130.2 (2003). DOI: 10.1017/S0950268802008221 (cited on page 8).

[48] Nolte, O. "Streptococcus agalactiae". *Lexikon der Infektionskrankheiten des Menschen: Erreger, Symptome, Diagnose, Therapie und Prophylaxe*. Berlin, Heidelberg: Springer, 2009. DOI: 10.1007/978-3-540-39026-8_1050 (cited on page 8).

[49] Monterrosa-Castro, Á., Rosales-Becerra, A., Monterrosa-Blanco, A., Monterrosa-Castro, Á., Rosales-Becerra, A., and Monterrosa-Blanco, A.
"Streptococcus agalactiae and genital ulcers in a heterosexual male". *Iberoamerican Journal of Medicine* 3.3 (2021). DOI: 10.5281/zenodo.4721400 (cited on page 8).

[50] Crestani, C., Forde, T. L., Lycett, S. J., Holmes, M. A., Fasth, C., Persson-Waller, K., and Zadoks, R. N. "The fall and rise of group B Streptococcus in dairy cattle: reintroduction due to human-to-cattle host jumps?" *Microbial Genomics* 7.9 (2021). DOI: 10.1099/mgen.0.000648 (cited on pages 8–10, 12).

[51] Le Doare, K. and Heath, P. T. "An overview of global GBS epidemiology". *Vaccine*. Prevention of Perinatal Group B Streptococcal Disease through Maternal Immunization 31 (2013). DOI: 10.1016/j.vaccine.2013.01.009 (cited on page 9).

[52] Jones, N., Oliver, K., Jones, Y., Haines, A., and Crook, D. "Carriage of group B streptococcus in pregnant women from Oxford, UK". *Journal of Clinical Pathology* 59.4 (2006). DOI: 10.1136/jcp.2005.029058 (cited on page 9).

[53] High, K. P., Edwards, M. S., and Baker, C. J. "Group B Streptococcal Infections in Elderly Adults". *Clinical Infectious Diseases* 41.6 (2005). DOI: 10.1086/432804 (cited on page 9).

[54] Sendi, P., Johansson, L., and Norrby-Teglund, A. "Invasive Group B Streptococcal Disease in Non-pregnant Adults". *Infection* 36.2 (2008). DOI: 10.1007/s15010-007-7251-0 (cited on page 9).

[55] Farley, M. M. and Strasbaugh, L. J. "Group B Streptococcal Disease in Nonpregnant Adults". *Clinical Infectious Diseases* 33.4 (2001). DOI: 10.1086/322696 (cited on pages 9, 11).

[56] Kusumadewi, Y. P., Febiyanti, A. M. G., Tazkiya, I., Allatief, G. R., Somaningtyas, A., Astuti, C. W., Puspitasari, I., Triyana, K., Wibawa, T., and Nuryastuti, T. "Streptococcus agalactiae is resistant to β-lactam antibiotics in a diabetic patient with foot infection: a case report". *Journal of Clinical Microbiology and Infectious Diseases* 2.1 (2022). DOI: 10.51559/jcmid.v2i1.13 (cited on page 9).

[57] Matheson, E. M., Bragg, S. W., and Blackwelder, R. S. "Diabetes-Related Foot Infections: Diagnosis and Treatment". *American Family Physician* 104.4 (2021) (cited on page 9).

[58] Reinscheid, F. "A new proposal for the causative agent of the sporadic form of Alzheimer‘s disease". *Medical Hypotheses* 146 (2021). DOI: 10.1016/j.mehy.2020.110453 (cited on page 9).

[59] Ozavci, V., Dolgun, H. T. Y., Kirkan, S., Seferoglu, Y., Semen, Z., and Parin, U. "Evaluation of Streptococcus species isolated from subclinical sheep mastitis by molecular methods and determination of virulence factors and antimicrobial resistance genes". *Veterinarni Medicina* 68.9 (2023). DOI: 10.17221/42/2023-VETMED (cited on page 10).

[60] Shi, H., Zhou, M., Zhang, Z., Hu, Y., Song, S., Hui, R., Wang, L., Li, G., and Yao, L. "Molecular epidemiology, drug resistance, and virulence gene analysis of Streptococcus agalactiae isolates from dairy goats in backyard farms in China". *Frontiers in Cellular and Infection Microbiology* 12 (2022). DOI: 10.3389/fcimb.2022.1049167 (cited on page 10).

[61] Shuster, K. A., Hish, G. A., Selles, L. A., Chowdhury, M. A., Wiggins, R. C., Dysko, R. C., and Bergin, I. L. "Naturally Occurring Disseminated Group B Streptococcus Infections in Postnatal Rats". *Comparative Medicine* 63.1 (2013) (cited on page 10).

[62] Bishop, E. J., Shilton, C., Benedict, S., Kong, F., Gilbert, G. L., Gal, D., Godoy, D., Spratt, B. G., and Currie, B. J. "Necrotizing fasciitis in captive juvenile Crocodylus porosus caused by Streptococcus agalactiae: an outbreak and review of the animal and human literature". *Epidemiology and Infection* 135.8 (2007). DOI: 10.1017/S0950268807008515 (cited on page 10).

[63] Delannoy, C. M. J., Crumlish, M., Fontaine, M. C., Pollock, J., Foster, G., Dagleish, M. P., Turnbull, J. F., and Zadoks, R. N.
"Human Streptococcus agalactiae strains in aquatic mammals and fish".
BMC microbiology 13 (2013). DOI: 10.1186/1471-2180-13-41 (cited on pages 10, 51).

[64] Evans, J. J., Pasnik, D. J., Klesius, P. H., and Al-Ablani, S. "First Report of Streptococcus Agalactiae and Lactococcus Garvieae from a Wild Bottlenose Dolphin (tursiops Truncatus)".
Journal of Wildlife Diseases 42.3 (2006). DOI: 10.7589/0090-3558-42.3.561
(cited on page 10).

[65] Eisenberg, T., Rau, J., Westerhüs, U., Knauf-Witzens, T., Fawzy, A., Schlez, K., Zschöck, M., Prenger-Berninghoff, E., Heydel, C., Sting, R., et al.
"Streptococcus agalactiae in elephants - A comparative study with isolates from human and zoo animal and livestock origin". *Veterinary Microbiology* 204 (2017).
DOI: 10.1016/j.vetmic.2017.04.018 (cited on pages 10, 22, 40).

[66] Richards, V. P., Velsko, I. M., Alam, M. T., Zadoks, R. N., Manning, S. D., Pavinski Bitar, P. D., Hassler, H. B., Crestani, C., Springer, G. H., Probert, B. M., et al.
"Population Gene Introgression and High Genome Plasticity for the Zoonotic Pathogen Streptococcus agalactiae". *Molecular Biology and Evolution* 36.11 (2019).
DOI: 10.1093/molbev/msz169 (cited on pages 10, 12).

[67] Crestani, C.
"The role of the mobilome in the evolution and host-adaptation of Streptococcus agalactiae". PhD thesis. 2021 (cited on pages 10, 43, 50, 53, 56).

[68] Kalimuddin, S., Chen, S. L., Lim, C. T. K., Koh, T. H., Tan, T. Y., Kam, M., Wong, C. W., Mehershahi, K. S., Chau, M. L., Ng, L. C., et al.
"2015 Epidemic of Severe Streptococcus agalactiae Sequence Type 283 Infections in Singapore Associated With the Consumption of Raw Freshwater Fish: A Detailed Analysis of Clinical, Epidemiological, and Bacterial Sequencing Data".
Clinical Infectious Diseases 64 (suppl_2 2017). DOI: 10.1093/cid/cix021
(cited on page 10).

[69] Goldfain, A., Cowell, L., and Smith, B. "Clonal Complexes in Biomedical Ontologies".
Nature Precedings (2009). DOI: 10.1038/npre.2009.3476.1 (cited on page 10).

[70] Teatero, S., Ramoutar, E., McGeer, A., Li, A., Melano, R. G., Wasserscheid, J., Dewar, K., and Fittipaldi, N. "Clonal Complex 17 Group B Streptococcus strains causing invasive disease in neonates and adults originate from the same genetic pool".
Scientific Reports 6 (2016). DOI: 10.1038/srep20047 (cited on page 10).

[71] Tsai, M.-H., Hsu, J.-F., Lai, M.-Y., Lin, L.-C., Chu, S.-M., Huang, H.-R., Chiang, M.-C., Fu, R.-H., and Lu, J.-J. "Molecular Characteristics and Antimicrobial Resistance of Group B Streptococcus Strains Causing Invasive Disease in Neonates and Adults".
Frontiers in Microbiology 10 (2019). DOI: 10.3389/fmicb.2019.00264
(cited on pages 10 sq.).

[72] Almeida, A., Alves-Barroco, C., Sauvage, E., Bexiga, R., Albuquerque, P., Tavares, F., Santos-Sanches, I., and Glaser, P. "Persistence of a dominant bovine lineage of group B Streptococcus reveals genomic signatures of host adaptation".
Environmental Microbiology 18.11 (2016). DOI: 10.1111/1462-2920.13550
(cited on page 10).

[73] Lyhs, U., Kulkas, L., Katholm, J., Waller, K. P., Saha, K., Tomusk, R. J., and Zadoks, R. N. "Streptococcus agalactiae Serotype IV in Humans and Cattle, Northern Europe1". *Emerging Infectious Diseases* 22.12 (2016). DOI: 10.3201/eid2212.151447 (cited on page 10).

[74] Manning, S. D., Springman, A. C., Million, A. D., Milton, N. R., McNamara, S. E., Somsel, P. A., Bartlett, P., and Davies, H. D. "Association of Group B Streptococcus Colonization and Bovine Exposure: A Prospective Cross-Sectional Cohort Study". *PLoS ONE* 5.1 (2010). DOI: 10.1371/journal.pone.0008795 (cited on page 10).

[75] Shabayek, S. and Spellerberg, B. "Group B Streptococcal Colonization, Molecular Characteristics, and Epidemiology". *Frontiers in Microbiology* 9 (2018). DOI: 10.3389/fmicb.2018.00437 (cited on pages 10, 51).

[76] Cieslewicz, M. J., Chaffin, D., Glusman, G., Kasper, D., Madan, A., Rodrigues, S., Fahey, J., Wessels, M. R., and Rubens, C. E. "Structural and Genetic Diversity of Group B Streptococcus Capsular Polysaccharides". *Infection and Immunity* 73.5 (2005). DOI: 10.1128/IAI.73.5.3096-3103.2005 (cited on page 10).

[77] Kapatai, G., Patel, D., Efstratiou, A., and Chalker, V. J. "Comparison of molecular serotyping approaches of Streptococcus agalactiae from genomic sequences". *BMC Genomics* 18 (2017). DOI: 10.1186/s12864-017-3820-5 (cited on page 10).

[78] Kong, F., Gowan, S., Martin, D., James, G., and Gilbert, G. L. "Serotype identification of group B streptococci by PCR and sequencing". *Journal of Clinical Microbiology* 40.1 (2002). DOI: 10.1128/JCM.40.1.216-226.2002 (cited on page 10).

[79] Bianchi-Jassir, F., Paul, P., To, K.-N., Carreras-Abad, C., Seale, A. C., Jauneikaite, E., Madhi, S. A., Russell, N. J., Hall, J., Madrid, L., et al. "Systematic review of Group B Streptococcal capsular types, sequence types and surface proteins as potential vaccine candidates". *Vaccine* 38.43 (2020). DOI: 10.1016/j.vaccine.2020.08.052 (cited on page 11).

[80] Hsu, J.-F., Tsai, M.-H., Lin, L.-C., Chu, S.-M., Lai, M.-Y., Huang, H.-R., Chiang, M.-C., Yang, P.-H., and Lu, J.-J. "Genomic Characterization of Serotype III/ST-17 Group B Streptococcus Strains with Antimicrobial Resistance Using Whole Genome Sequencing". *Biomedicines* 9.10 (2021). DOI: 10.3390/biomedicines9101477 (cited on page 11).

[81] Davies, H. D., Jones, N., Whittam, T. S., Elsayed, S., Bisharat, N., and Baker, C. J. "Multilocus Sequence Typing of Serotype III Group B Streptococcus and Correlation with Pathogenic Potential". *The Journal of Infectious Diseases* 189.6 (2004). DOI: 10.1086/382087 (cited on pages 11, 51).

[82] Hernandez, L., Bottini, E., Cadona, J., Cacciato, C., Monteavaro, C., Bustamante, A., and Sanso, A. M. "Multidrug Resistance and Molecular Characterization of Streptococcus agalactiae Isolates From Dairy Cattle With Mastitis". *Frontiers in Cellular and Infection Microbiology* 11 (2021). DOI: 10.3389/fcimb.2021.647324 (cited on page 11).

[83] Dogan, B., Schukken, Y. H., Santisteban, C., and Boor, K. J. "Distribution of serotypes and antimicrobial resistance genes among Streptococcus agalactiae isolates from bovine and human hosts". *Journal of Clinical Microbiology* 43.12 (2005). DOI: 10.1128/JCM.43.12.5899-5906.2005 (cited on page 11).

[84] Duarte, R. S., Miranda, O. P., Bellei, B. C., Brito, M. A. V. P., and Teixeira, L. M. "Phenotypic and molecular characteristics of Streptococcus agalactiae isolates recovered from milk of dairy cows in Brazil". *Journal of Clinical Microbiology* 42.9 (2004). DOI: 10.1128/JCM.42.9.4214-4222.2004 (cited on page 11).

[85] HajiAhmadi, P., Momtaz, H., and Tajbakhsh, E. "Capsular Typing and Molecular Characterization of Streptococcus agalactiae Strains Isolated From Bovine Mastitis in Iran". *Veterinary Medicine and Science* 11.2 (2025). DOI: 10.1002/vms3.70275 (cited on page 11).

[86] Yang, Y., Liu, Y., Ding, Y., Yi, L., Ma, Z., Fan, H., and Lu, C. "Molecular Characterization of Streptococcus agalactiae Isolated from Bovine Mastitis in Eastern China". *PLOS ONE* 8.7 (2013). DOI: 10.1371/journal.pone.0067755 (cited on page 11).

[87] Pang, M., Sun, L., He, T., Bao, H., Zhang, L., Zhou, Y., Zhang, H., Wei, R., Liu, Y., and Wang, R. "Molecular and virulence characterization of highly prevalent Streptococcus agalactiae circulated in bovine dairy herds". *Veterinary Research* 48.1 (2017). DOI: 10.1186/s13567-017-0461-2 (cited on page 11).

[88] Radtke, A., Bruheim, T., Afset, J. E., and Bergh, K. "Multiple-locus variant-repeat assay (MLVA) is a useful tool for molecular epidemiologic analysis of *Streptococcus agalactiae* strains causing bovine mastitis". *Veterinary Microbiology* 157.3 (2012). DOI: 10.1016/j.vetmic.2011.12.034 (cited on page 11).

[89] Rosini, R. and Margarit, I. "Biofilm formation by Streptococcus agalactiae: influence of environmental conditions and implicated virulence factors". *Frontiers in Cellular and Infection Microbiology* 5 (2015). DOI: 10.3389/fcimb.2015.00006 (cited on page 11).

[90] Bellais, S., Six, A., Fouet, A., Longo, M., Dmytruk, N., Glaser, P., Trieu-Cuot, P., and Poyart, C. "Capsular Switching in Group B Streptococcus CC17 Hypervirulent Clone: A Future Challenge for Polysaccharide Vaccine Development". *The Journal of Infectious Diseases* 206.11 (2012). DOI: 10.1093/infdis/jis605 (cited on page 11).

[91] Meehan, M., Cunney, R., and Cafferkey, M. "Molecular epidemiology of group B streptococci in Ireland reveals a diverse population with evidence of capsular switching". *European Journal of Clinical Microbiology & Infectious Diseases* 33.7 (2014). DOI: 10.1007/s10096-014-2055-5 (cited on pages 11, 51).

[92] Flores, A. R., Galloway-Peña, J., Sahasrabhojane, P., Saldaña, M., Yao, H., Su, X., Ajami, N. J., Holder, M. E., Petrosino, J. F., Thompson, E., et al. "Sequence type 1 group B Streptococcus, an emerging cause of invasive disease in adults, evolves by small genetic changes". *Proceedings of the National Academy of Sciences of the United States of America* 112.20 (2015). DOI: 10.1073/pnas.1504725112 (cited on page 11).

[93] Mee-Marquet, N. van der, Domelier, A.-S., Salloum, M., Violette, J., Arnault, L., Gaillard, N., Bind, J.-L., Lartigue, M.-F., and Quentin, R. "Molecular Characterization of Temporally and Geographically Matched Streptococcus agalactiae Strains Isolated from Food Products and Bloodstream Infections". *Foodborne Pathogens and Disease* 6.10 (2009). DOI: 10.1089/fpd.2009.0287 (cited on page 11).

[94] Oliveira, L. M. A. d., Simões, L. C., Crestani, C., Costa, N. S., Pantoja, J. C. d. F., Rabello, R. F., Teixeira, L. M., Khan, U. B., Bentley, S., Jamrozy, D., et al. "Long-Term Co-Circulation of Host-Specialist and Host-Generalist Lineages of Group B Streptococcus in Brazilian Dairy Cattle with Heterogeneous Antimicrobial Resistance Profiles". *Antibiotics* 13.5 (2024). DOI: 10.3390/antibiotics13050389 (cited on pages 11, 44).

[95] Campisi, E., Rosini, R., Ji, W., Guidotti, S., Rojas-López, M., Geng, G., Deng, Q., Zhong, H., Wang, W., Liu, H., et al. "Genomic Analysis Reveals Multi-Drug Resistance Clusters in Group B Streptococcus CC17 Hypervirulent Isolates Causing Neonatal Invasive Disease in Southern Mainland China". *Frontiers in Microbiology* 7 (2016). DOI: 10.3389/fmicb.2016.01265 (cited on page 11).

[96] Hayes, K., O'Halloran, F., and Cotter, L. "A review of antibiotic resistance in Group B Streptococcus: the story so far". *Critical Reviews in Microbiology* 46.3 (2020). DOI: 10.1080/1040841X.2020.1758626 (cited on pages 11 sq.).

[97] Da Cunha, V., Davies, M. R., Douarre, P.-E., Rosinski-Chupin, I., Margarit, I., Spinali, S., Perkins, T., Lechat, P., Dmytruk, N., Sauvage, E., et al. "Streptococcus agalactiae clones infecting humans were selected and fixed through the extensive use of tetracycline". *Nature communications* 5 (2014). DOI: 10.1038/ncomms5544 (cited on page 11).

[98] Oliveira, I. C. M., Mattos, M. C. de, Pinto, T. A., Ferreira-Carvalho, B. T., Benchetrit, L. C., Whiting, A. A., Bohnsack, J. F., and Figueiredo, A. M. S. "Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B Streptococcus type V cluster displaying an identical pulsed-field gel electrophoresis pattern". *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 12.9 (2006). DOI: 10.1111/j.1469-0691.2006.01508.x (cited on page 12).

[99] Lamagni, T. L., Keshishian, C., Efstratiou, A., Guy, R., Henderson, K. L., Broughton, K., and Sheridan, E. "Emerging Trends in the Epidemiology of Invasive Group B Streptococcal Disease in England and Wales, 1991–2010". *Clinical Infectious Diseases* 57.5 (2013). DOI: 10.1093/cid/cit337 (cited on page 12).

[100] Kimura, K., Suzuki, S., Wachino, J.-i., Kurokawa, H., Yamane, K., Shibata, N., Nagano, N., Kato, H., Shibayama, K., and Arakawa, Y. "First Molecular Characterization of Group B Streptococci with Reduced Penicillin Susceptibility". *Antimicrobial Agents and Chemotherapy* 52.8 (2008). DOI: 10.1128/AAC.00185-08 (cited on page 12).

[101] Oliveira, L. M. A., Simões, L. C., Costa, N. S., Zadoks, R. N., and Pinto, T. C. A. "The landscape of antimicrobial resistance in the neonatal and multi-host pathogen group B Streptococcus: review from a One Health perspective". *Frontiers in Microbiology* 13 (2022). DOI: 10.3389/fmicb.2022.943413 (cited on page 12).

[102] Lohrmann, F., Hufnagel, M., Kunze, M., Afshar, B., Creti, R., Detcheva, A., Kozakova, J., Rodriguez-Granger, J., Sørensen, U. B. S., Margarit, I., et al. "Neonatal invasive disease caused by Streptococcus agalactiae in Europe: the DEVANI multi-center study". *Infection* 51.4 (2023). DOI: 10.1007/s15010-022-01965-x (cited on page 12).

[103] Pinto, T. C. A., Costa, N. S., Corrêa, A. B. d. A., Oliveira, I. C. M. de, Mattos, M. C. de, Rosado, A. S., and Benchetrit, L. C. "Conjugative transfer of resistance determinants among human and bovine Streptococcus agalactiae". *Brazilian Journal of Microbiology* 45.3 (2014) (cited on page 12).

[104] Parkhill, J. and Wren, B. W.
"Bacterial epidemiology and biology - lessons from genome sequencing".
Genome Biology 12.10 (2011). DOI: 10.1186/gb-2011-12-10-230 (cited on page 14).

[105] Rasche, H. "Bioinformagic: Towards a book of spells to make every analysis magical".
PhD thesis. Rotterdam, 2025 (cited on page 14).

[106] Hendriksen, R. S., Bortolaia, V., Tate, H., Tyson, G. H., Aarestrup, F. M., and McDermott, P. F. "Using Genomics to Track Global Antimicrobial Resistance".
Frontiers in Public Health 7 (2019). DOI: 10.3389/fpubh.2019.00242 (cited on page 14).

[107] Blaxter, M., Danchin, A., Savakis, B., Fukami-Kobayashi, K., Kurokawa, K., Sugano, S., Roberts, R. J., Salzberg, S. L., and Wu, C.-I. "Reminder to deposit DNA sequences".
Science 352.6287 (2016). DOI: 10.1126/science.aaf7672 (cited on page 14).

[108] Karsch-Mizrachi, I., Arita, M., Burdett, T., Cochrane, G., Nakamura, Y., Pruitt, K. D., Schneider, V. A., and International Nucleotide Sequence Database Collaboration, o. b. of the. "The international nucleotide sequence database collaboration (INSDC): enhancing global participation". *Nucleic Acids Research* 53 (D1 2024). DOI: 10.1093/nar/gkae1058 (cited on page 14).

[109] Staff, N. GenBank Release 269.0. NCBI Insights. 2025.
URL: https://ncbiinsights.ncbi.nlm.nih.gov/2025/12/22/genbank-release-269-0/ (visited on 02/24/2026) (cited on page 14).

[110] Beura, A., Manjunath, G. K., Mahalingam, S., Rajguru, M. S., Dakal, T. C., and Kumar, A. "Next generation DNA sequencing data analysis and its application in clinical genomics".
Pathology - Research and Practice 276 (2025). DOI: 10.1016/j.prp.2025.156280 (cited on page 15).

[111] Schwengers, O.
"Novel scalable approaches for the computational analysis of bacterial genomes".
Dissertation (2021) (cited on page 15).

[112] Maljkovic Berry, I., Melendrez, M. C., Bishop-Lilly, K. A., Rutvisuttinunt, W., Pollett, S., Talundzic, E., Morton, L., and Jarman, R. G. "Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity".
The Journal of Infectious Diseases 221 (Supplement_3 2020).
DOI: 10.1093/infdis/jiz286 (cited on page 15).

[113] Index of /genomes/ASSEMBLY_REPORTS. NCBI.
URL: https://ftp.ncbi.nih.gov/genomes/ASSEMBLY_REPORTS/ (visited on 01/15/2026) (cited on page 16).

[114] GenBank and WGS Statistics.
URL: https://www.ncbi.nlm.nih.gov/genbank/statistics/ (visited on 01/15/2026) (cited on page 16).

[115] Bingmann, T., Bradley, P., Gauger, F., and Iqbal, Z.
"COBS: A Compact Bit-Sliced Signature Index".
String Processing and Information Retrieval (2019). DOI: 10.1007/978-3-030-32686-9_21 (cited on page 16).

[116] Pierce, N. T., Irber, L., Reiter, T., Brooks, P., and Brown, C. T.
"Large-scale sequence comparisons with sourmash". *F1000Research* 8 (2019).
DOI: 10.12688/f1000research.19675.1 (cited on page 16).

[117] Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., Corander, J., Bentley, S. D., and Croucher, N. J.
"Fast and flexible bacterial genomic epidemiology with PopPUNK".
Genome Research 29.2 (2019). DOI: 10.1101/gr.241455.118 (cited on page 16).

[118] Aupperle, H., Reischauer, A., Bach, F., Hildebrandt, T., Göritz, F., Jäger, K., Scheller, R., Klaue, H.-J., and Schoon, H.-A.
"Chronic Endometritis in an Asian Elephant (Elephas maximus)".
Journal of Zoo and Wildlife Medicine 39.1 (2008). DOI: 10.1638/2006-0045.1
(cited on page 22).

[119] Keet, D. F., Grobler, D. G., Raath, J. P., Gouws, J., Carstens, J., and Nesbit, J. W.
"Ulcerative pododermatitis in free-ranging African elephant (Loxodonta africana) in the Kruger National Park" (1997) (cited on page 22).

[120] Lewis, K. D., Shepherdson, D. J., Owens, T. M., and Keele, M.
"A survey of elephant husbandry and foot health in North American zoos".
Zoo Biology 29.2 (2010). DOI: 10.1002/zoo.20291 (cited on page 22).

[121] Gori, A., Harrison, O. B., Mlia, E., Nishihara, Y., Chan, J. M., Msefula, J., Mallewa, M., Dube, Q., Swarthout, T. D., Nobbs, A. H., et al. "Pan-GWAS of Streptococcus agalactiae Highlights Lineage-Specific Genes Associated with Virulence and Niche Adaptation".
mBio 11.3 (2020). DOI: 10.1128/mbio.00728-20 (cited on page 23).

[122] Chklovski, A., Parks, D. H., Woodcroft, B. J., and Tyson, G. W. "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning".
Nature Methods 20.8 (2023). DOI: 10.1038/s41592-023-01940-w (cited on page 29).

[123] Petit, R. A. assembly-scan. GitHub. 2025.
URL: https://github.com/rpetit3/assembly-scan (visited on 10/21/2025)
(cited on page 29).

[124] Seemann, T. mlst. GitHub. 2025.
URL: https://github.com/tseemann/mlst (visited on 10/21/2025) (cited on page 29).

[125] Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H.
"GTDB-Tk v2: memory friendly classification with the genome taxonomy database".
Bioinformatics 38.23 (2022). DOI: 10.1093/bioinformatics/btac672 (cited on page 29).

[126] Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". *Nature Biotechnology* 36.10 (2018).
DOI: 10.1038/nbt.4229 (cited on page 30).

[127] Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., and Goesmann, A.
"Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification". *Microbial Genomics* 7.11 (2021). DOI: 10.1099/mgen.0.000685
(cited on page 30).

[128] AllTheBacteria/AllTheBacteria. 2026.
URL: https://github.com/AllTheBacteria/AllTheBacteria (visited on 02/24/2026)
(cited on pages 30, 37).

[129] Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P.
"GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy".
Nucleic Acids Research 50 (D1 2022). DOI: 10.1093/nar/gkab776 (cited on page 37).

[130] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". *Nucleic Acids Research* 44 (D1 2016). DOI: 10.1093/nar/gkv1189 (cited on page 37).

[131] Dyer, S. C., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Barrera-Enriquez, V. P., Becker, A., Bennett, R., Beracochea, M., Berry, A., et al. "Ensembl 2025". *Nucleic Acids Research* 53 (D1 2025). DOI: 10.1093/nar/gkae1071 (cited on page 37).

[132] Fullam, A., Letunic, I., Maistrenko, O. M., Castro, A. A., Coelho, L. P., Grekova, A., Schudoma, C., Khedkar, S., Robbani, M., Kuhn, M., et al. "proGenomes4: providing 2 million accurately and consistently annotated high-quality prokaryotic genomes". *Nucleic Acids Research* 54 (D1 2026). DOI: 10.1093/nar/gkaf1208 (cited on page 37).

[133] Aruna overview - Aruna Documentation. URL: https://docs.aruna-engine.org/latest/ (visited on 04/29/2025) (cited on page 38).

[134] Patel, J. K. and Elangovan, R. *MetaMiner: Streamlined GUI Tool for Retrieving, Normalizing and Exploring Metadata.* 2025. DOI: 10.1101/2025.08.20.666107 (cited on pages 39, 50).

[135] Schaefer, M., Peneder, P., Malzl, D., Lombardo, S. D., Peycheva, M., Burton, J., Hakobyan, A., Sharma, V., Krausgruber, T., Sin, C., et al. "Multimodal learning enables chat-based exploration of single-cell data". *Nature Biotechnology* (2025). DOI: 10.1038/s41587-025-02857-9 (cited on page 39).

[136] Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., et al. "Sequence modeling and design from molecular to genome scale with Evo". *Science* 386.6723 (2024). DOI: 10.1126/science.ado9336 (cited on page 39).

[137] Sun, Y., Kong, F., Zhao, Z., and Gilbert, G. L. "Comparison of a 3-Set Genotyping System with Multilocus Sequence Typing for Streptococcus agalactiae (Group B Streptococcus)". *Journal of Clinical Microbiology* 43.9 (2005). DOI: 10.1128/jcm.43.9.4704-4707.2005 (cited on page 39).

[138] mBioWorks AI Assistants – mBioWorks Copenhagen. URL: https://mbioworks.com/?p=9586 (visited on 01/23/2026) (cited on page 39).

[139] Blom, J. "Comparative genomics on gene and single nucleotide level." PhD thesis (cited on page 43).

[140] Sukhnanand, S., Dogan, B., Ayodele, M. O., Zadoks, R. N., Craver, M. P. J., Dumas, N. B., Schukken, Y. H., Boor, K. J., and Wiedmann, M. "Molecular subtyping and characterization of bovine and human Streptococcus agalactiae isolates". *Journal of Clinical Microbiology* 43.3 (2005). DOI: 10.1128/JCM.43.3.1177-1186.2005 (cited on page 43).

[141] Bowater, R. O., Forbes-Faulkner, J., Anderson, I. G., Condon, K., Robinson, B., Kong, F., Gilbert, G. L., Reynolds, A., Hyland, S., McPherson, G., et al. "Natural outbreak of Streptococcus agalactiae (GBS) infection in wild giant Queensland grouper, Epinephelus lanceolatus (Bloch), and other wild fish in northern Queensland, Australia". *Journal of Fish Diseases* 35.3 (2012). DOI: 10.1111/j.1365-2761.2011.01332.x (cited on page 43).

[142] Sørensen, U. B. S., Poulsen, K., Ghezzo, C., Margarit, I., and Kilian, M. "Emergence and Global Dissemination of Host-Specific Streptococcus agalactiae Clones". *mBio* 1.3 (2010). DOI: 10.1128/mBio.00178-10 (cited on pages 43, 46).

[143] Franken, C., Haase, G., Brandt, C., Weber-Heynemann, J., Martin, S., Lämmler, C., Podbielski, A., Lütticken, R., and Spellerberg, B. "Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing scpB and lmb". *Molecular Microbiology* 41.4 (2001). DOI: 10.1046/j.1365-2958.2001.02563.x (cited on page 43).

[144] Franken, C., Brandt, C., Bröker, G., and Spellerberg, B. "IS*Sag1* in streptococcal strains of human and animal origin". *International Journal of Medical Microbiology* 294.4 (2004). DOI: 10.1016/j.ijmm.2004.04.002 (cited on page 43).

[145] Roder, T. Scoary2. GitHub. URL: https://github.com/MrTomRod/scoary-2/wiki/Home (visited on 03/05/2026) (cited on page 45).

[146] Crestani, C., Seligsohn, D., Forde, T. L., and Zadoks, R. N. "How GBS Got Its Hump: Genomic Analysis of Group B Streptococcus from Camels Identifies Host Restriction as well as Mobile Genetic Elements Shared across Hosts and Pathogens". *Pathogens* 11.9 (2022). DOI: 10.3390/pathogens11091025 (cited on pages 45 sq.).

[147] Rothen, J., Pothier, J. F., Foucault, F., Blom, J., Nanayakkara, D., Li, C., Ip, M., Tanner, M., Vogel, G., Pflüger, V., et al. "Subspecies Typing of Streptococcus agalactiae Based on Ribosomal Subunit Protein Mass Variation by MALDI-TOF MS". *Frontiers in Microbiology* 10 (2019). DOI: 10.3389/fmicb.2019.00471 (cited on page 45).

[148] Creti, R., Imperi, M., Pataracchia, M., Alfarone, G., Recchia, S., and Baldassarri, L. "Identification and molecular characterization of a S. agalactiae strain lacking the capsular locus". *European Journal of Clinical Microbiology & Infectious Diseases* 31.3 (2012). DOI: 10.1007/s10096-011-1298-7 (cited on page 45).

[149] Brousseau, R., Hill, J. E., Préfontaine, G., Goh, S.-H., Harel, J., and Hemmingsen, S. M. "Streptococcus suis Serotypes Characterized by Analysis of Chaperonin 60 Gene Sequences". *Applied and Environmental Microbiology* 67.10 (2001). DOI: 10.1128/AEM.67.10.4828-4833.2001 (cited on page 46).

[150] Bedeley, E., Gori, A., Yeboah-Manu, D., and Diallo, K. "Control of Streptococcal Infections: Is a Common Vaccine Target Achievable Against Streptococcus agalactiae and Streptococcus pneumoniae". *Frontiers in Microbiology* 12 (2021). DOI: 10.3389/fmicb.2021.658824 (cited on page 46).

[151] Bisharat, N., Crook, D. W., Leigh, J., Harding, R. M., Ward, P. N., Coffey, T. J., Maiden, M. C., Peto, T., and Jones, N. "Hyperinvasive neonatal group B streptococcus has arisen from a bovine ancestor". *Journal of Clinical Microbiology* 42.5 (2004). DOI: 10.1128/JCM.42.5.2161-2167.2004 (cited on page 46).

[152] Ernst, C. M., Staubitz, P., Mishra, N. N., Yang, S.-J., Hornig, G., Kalbacher, H., Bayer, A. S., Kraus, D., and Peschel, A. "The Bacterial Defensin Resistance Protein MprF Consists of Separable Domains for Lipid Lysinylation and Antimicrobial Peptide Repulsion". *PLOS Pathogens* 5.11 (2009). DOI: 10.1371/journal.ppat.1000660 (cited on page 48).

[153] Cobo-Angel, C. G., Jaramillo-Jaramillo, A. S., Palacio-Aguilera, M., Jurado-Vargas, L., Calvo-Villegas, E. A., Ospina-Loaiza, D. A., Rodriguez-Lecompte, J. C., Sanchez, J., Zadoks, R., and Ceballos-Marquez, A. "Potential group B Streptococcus interspecies transmission between cattle and people in Colombian dairy farms". *Scientific Reports* 9.1 (2019). DOI: 10.1038/s41598-019-50225-w (cited on page 48).

[154] Sørensen, U. B. S., Klaas, I. C., Boes, J., and Farre, M. "The distribution of clones of Streptococcus agalactiae (group B streptococci) among herdspersons and dairy cows demonstrates lack of host specificity for some lineages". *Veterinary Microbiology* 235 (2019). DOI: 10.1016/j.vetmic.2019.06.008 (cited on page 48).

[155] Caliskan, A., Dangwal, S., and Dandekar, T. "Metadata integrity in bioinformatics: Bridging the gap between data and knowledge". *Computational and Structural Biotechnology Journal* 21 (2023). DOI: 10.1016/j.csbj.2023.10.006 (cited on page 50).

[156] ENA Browser. URL: https://www.ebi.ac.uk/ena/browser/view/ERC000028 (visited on 09/26/2025) (cited on page 50).

[157] Tong, Z., Kong, F., Wang, B., Zeng, X., and Gilbert, G. L. "A practical method for subtyping of Streptococcus agalactiae serotype III, of human origin, using rolling circle amplification". *Journal of Microbiological Methods* 70.1 (2007). DOI: 10.1016/j.mimet.2007.03.010 (cited on page 51).

[158] Kao, Y., Tsai, M.-H., Lai, M.-Y., Chu, S.-M., Huang, H.-R., Chiang, M.-C., Fu, R.-H., Lu, J.-J., and Hsu, J.-F. "Emerging serotype III sequence type 17 group B streptococcus invasive infection in infants: the clinical characteristics and impacts on outcomes". *BMC Infectious Diseases* 19 (2019). DOI: 10.1186/s12879-019-4177-y (cited on page 51).

[159] Cubria, M. B., Vega, L. A., Shropshire, W. C., Sanson, M. A., Shah, B. J., Regmi, S., Rench, M., Baker, C. J., and Flores, A. R. "Population Genomics Reveals Distinct Temporal Association with the Emergence of ST1 Serotype V Group B Streptococcus and Macrolide Resistance in North America". *Antimicrobial Agents and Chemotherapy* (2021). DOI: 10.1128/AAC.00714-21 (cited on page 51).

[160] Salloum, M., Mee-Marquet, N. v. d., Valentin-Domelier, A.-S., and Quentin, R. "Diversity of Prophage DNA Regions of Streptococcus agalactiae Clonal Lineages from Adults and Neonates with Invasive Infectious Disease". *PLOS ONE* 6.5 (2011). DOI: 10.1371/journal.pone.0020256 (cited on page 51).

[161] Teatero, S., Athey, T. B. T., Van Caeseele, P., Horsman, G., Alexander, D. C., Melano, R. G., Li, A., Flores, A. R., Shelburne, S. A., McGeer, A., et al. "Emergence of Serotype IV Group B Streptococcus Adult Invasive Disease in Manitoba and Saskatchewan, Canada, Is Driven by Clonal Sequence Type 459 Strains". *Journal of Clinical Microbiology* 53.9 (2015). DOI: 10.1128/JCM.01128-15 (cited on page 51).

[162] Ntozini, B., Walaza, S., Metcalf, B., Hazelhurst, S., Gouveia, L. de, Meiring, S., Mogale, D., Mtshali, S., Ismail, A., Ndlangisa, K., et al. "Molecular Epidemiology of Invasive Group B Streptococcus in South Africa, 2019–2020". *The Journal of Infectious Diseases* 231.4 (2025). DOI: 10.1093/infdis/jiae633 (cited on page 51).

[163] Wadilo, F., Hailemeskel, E., Kedir, K., El-Khatib, Z., Asogba, P. C., Seyoum, T., Landis, F. C., Howe, R., and Boltena, M. T.
"Prevalence of Group B Streptococcus maternal colonization, serotype distribution, and antimicrobial resistance in Sub-Saharan Africa: A systematic review and meta-analysis".
Journal of Global Antimicrobial Resistance 32 (2023). DOI: 10.1016/j.jgar.2023.02.004 (cited on page 51).

[164] Chukwu, M. O., Mavenyengwa, R. T., Monyama, C. M., Bolukaoto, J. Y., Lebelo, S. L., Maloba, M. R., Nchabeleng, M., and Moyo, S. R. "Antigenic distribution of Streptococcus agalactiae isolates from pregnant women at Garankuwa hospital - South Africa".
Germs 5.4 (2015). DOI: 10.11599/germs.2015.1080 (cited on page 51).

[165] Ip, M., Ang, I., Fung, K., Liyanapathirana, V., Luo, M. J., and Lai, R. "Hypervirulent Clone of Group B *Streptococcus* Serotype III Sequence Type 283, Hong Kong, 1993–2012".
Emerging Infectious Diseases 22.10 (2016). DOI: 10.3201/eid2210.151436 (cited on page 51).

[166] Khan, U. B., Portal, E. A. R., Sands, K., Lo, S., Chalker, V. J., Jauneikaite, E., and Spiller, O. B. "Genomic Analysis Reveals New Integrative Conjugal Elements and Transposons in GBS Conferring Antimicrobial Resistance".
Antibiotics (Basel, Switzerland) 12.3 (2023). DOI: 10.3390/antibiotics12030544 (cited on page 51).

[167] Six, A., Bellais, S., Bouaboud, A., Fouet, A., Gabriel, C., Tazi, A., Dramsi, S., Trieu-Cuot, P., and Poyart, C.
"Srr2, a multifaceted adhesin expressed by ST-17 hypervirulent Group B Streptococcus involved in binding to both fibrinogen and plasminogen".
Molecular Microbiology 97.6 (2015). DOI: 10.1111/mmi.13097 (cited on page 52).

[168] Cohan, F. M. "What are Bacterial Species?" *Annual Review of Microbiology* 56.1 (2002). DOI: 10.1146/annurev.micro.56.012302.160634 (cited on page 56).

[169] Locey, K. J. and Lennon, J. T. "Scaling laws predict global microbial diversity".
Proceedings of the National Academy of Sciences 113.21 (2016).
DOI: 10.1073/pnas.1521291113 (cited on page 56).

[170] Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., Küsel, K., Rillig, M. C., Rivett, D. W., Salles, J. F., et al.
"Where less may be more: how the rare biosphere pulls ecosystems strings".
The ISME Journal 11.4 (2017). DOI: 10.1038/ismej.2016.174 (cited on page 56).

[171] Wick, R. R. "Choose-your-own-adventure guide to bacterial genome assembly" (2021).
DOI: 10.5281/zenodo.7471199 (cited on page 142).

[172] Wick, R. R. Generating assemblies. GitHub.
URL: https://github.com/rrwick/Trycycler/wiki/Generating-assemblies (visited on 02/26/2026) (cited on page 142).

Declaration

Ich erkläre: Ich habe die vorgelegte Dissertation selbstständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Ich stimme einer evtl. Überprüfung meiner Dissertation durch eine Antiplagiat-Software zu. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der "Satzung der Justus-Liebig-Universität Giessen zur Sicherung guter wissenschaftlicher Praxis" niedergelegt sind, eingehalten.

Angaben zu auf künstlicher Intelligenz (KI) basierender Hilfen wie ChatGPT oder SchulKI von OpenAI oder Gemini von Google zur Erstellung meiner Dissertation (Zutreffendes angekreuzt):

- ☐ Ich habe bei der Erstellung dieses Textes kein KI-Tool verwendet.
- ■ Ich habe ein KI-Tool in den folgenden Bereichen eingesetzt:
  - ☐ Ideen finden, meine Kreativität anregen.
  - ■ Verstehen von Konzepten, Recherche von Fakten und Definitionen
  - ■ Optimierung eines von mir verfassten Textes
  - ☐ Erstellen ganzer Textpassagen nach meinen Vorgaben

Folgende KI-Tools habe ich verwendet: ChatGPT, perplexity, grammarly, JLU kiChat

....

Linda Fenske Datum

Acknowledgements

Zuallererst möchte ich mich bei Prof. Dr. Alexander Goesmann bedanken. Danke, **Alex**, dass ich direkt nach meiner Masterthesis in deiner Arbeitsgruppe anfangen durfte und du mir so viel Freiheit bei der Wahl meiner Projekte gelassen hast. Vielen Dank, für deine Unterstützung und dafür, dass du mir die nötigen Ressourcen und den Entscheidungsfreiraum gegeben hast, um diese Thesis zu vollenden.

Außerdem möchte ich mich bei Prof. Dr. Tobias Eisenberg bedanken. Danke **Tobias**, dass du mich nicht nur seit meiner Masterthesis so sehr unterstützt, sondern mir auch ein Thema für meine Dissertation bereitgestellt hast. Ich erinnere mich noch so gut daran wie du während meines Masterstudiums, deinen Vortrag in der Ringvorlesung gehalten hast, um jemanden für ein Praktikum beim Landeslabor zu finden. Ich wusste damals direkt, dass das der Bereich ist, in dem ich arbeiten möchte, und ich bin so froh, dass du mich auch in all den Jahren danach noch so unterstützt hast. Das ist wirklich nicht selbstverständlich!

An dieser Stelle auch ein Dankeschön an Prof. Dr. Stefan Janssen und Dr. Oliver Roßbach. Danke **Stefan und Oli**, dass ihr euch so schnell bereit erklärt habt Teil meiner Prüfungskommission zu werden.

Ein ganz großer Dank geht außerdem an Dr. Jochen Blom. Danke, **Jochen**, dass du mich in die Arbeitsgruppe geholt hast und bei allen Problemen immer mit Rat und Tat zur Seite standest. Ich weiß noch, wie unsicher ich mir nach meinem Bachelor war, ob die Bioinformatik wirklich das Richtige für mich ist und deine Antwort auf die E-Mail, die ich dir damals geschrieben hatte, war: *„Unsere Abbrecherquote ist eigentlich sehr gering.“* Obwohl ich während des Masters ein paar Mal übers Abbrechen nachgedacht habe, wollte ich diese Quote dann doch nicht erhöhen :D

Ich bin sehr froh, dass ich letztendlich hier gelandet bin und du mir die Verantwortung über die EDGAR Projekte überlassen hast, wodurch ich so viele tolle Menschen und Kooperationspartner kennenlernen durfte. Ich war zu Beginn meines PhD tatsächlich etwas besorgt, wie es denn weitergeht, wenn ich nicht an der EDGAR Backend Entwicklung mitwirken möchte. Doch auch von deiner Seite habe ich jegliche Freiheit für die Wahl meiner Projekte bekommen. Vielen Dank dafür!

Ebenfalls bedanken möchte ich mich bei Dr. Oliver Schwengers. Danke, **Oli**, dass dir nicht nur die Idee für BakRep gekommen ist, sondern auch, dass du mir dieses Thema anvertraut hast. Trotz meinem fehlenden Informatik-Backgrounds hast du daran geglaubt, dass ich das schaffen kann, und dein wissenschaftlicher Input hat mir vieles erleichtert.

Darüber hinaus möchte ich mich natürlich auch bei der gesamten **AG Goesmann**, sowie natürlich auch der **AG Janssen** bedanken. Danke, Lukas für deine stetigen Entwicklungstätigkeiten und deine Hilfe bei allen möglichen Nextflow-, Programmierungs- und mittlerweile sogar Boulder-Problemen. Danke, Frank, für deine Unterstützung bei Cloud und Cluster. Da wurden so manche Probleme sogar mal um 4 Uhr nachts gelöst. Danke, Burkhard, dass du meine schlechten Informatikkenntnisse schon so lange tolerierst und auch die zehnte aus Versehen gelöschte Datei noch aus dem Backup fischst. Danke, ihr alle, für Film-, Spiele- und Käse-Abende; Feuerzangenbowle und Karaoke; Glühwein und Eierlikör; wissenschaftlichen und alltäglichen Wahnsinn.

Ganz großer Dank auch an meine aktuellen und ehemaligen Bürokollegen. **Julian, Sonja, Andi**, danke, dass ihr mich aufgenommen habt und wir so viele unzählige lustige wie auch verzweifelte, produktive und absolut unproduktive Stunden verbringen konnten. Ob wir gelacht, geweint, geschrien; Energiegetränke, Kuchen oder auch nur Hühnchen mit Reis zu uns genommen haben: Ohne euch wäre das Ganze hier nicht mal halb so gut machbar gewesen.

An meine **Familie und Freunde**, es sprengt leider den Rahmen euch hier alle zu erwähnen, aber ich bin einfach froh so viele Menschen an meiner Seite zu haben, die mein Leben bereichern und mich immer unterstützen. Ob nun Konzerte, Festivals (vor allem selbst organisierte) oder Spieleabende, Urlaube oder Poolpartys, Dienstage, Freitage oder Wochenenden, ohne diesen Teil meines Lebens, wäre ich nicht so weit gekommen. Also, an euch alle und passend zum Titel dieser Thesis: *„To Infinity and Beyond!“*

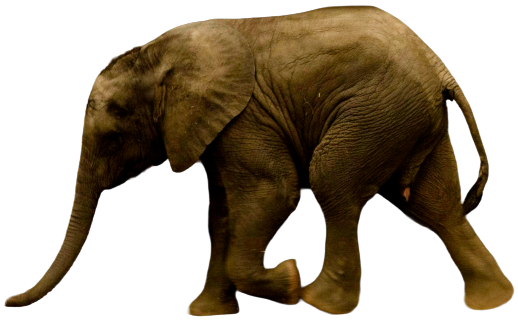
Danke auch an meine Therapeutin, die mir klargemacht hat, dass ich das hier schaffen kann. Und wenn nicht? Dann versuche ich's halt einfach nochmal!

Zuletzt möchte ich mich aber noch beim wichtigsten Menschen in meinem Leben bedanken. Danke, **Aniko**, dass du seit über 14 Jahren an meiner Seite bist und mich unterstützt, egal was passiert. Ohne dich hätte ich schon so viele Male aufgegeben, ohne dich wäre ich nicht hier!

Opa, auch wenn du diese Arbeit nicht mehr lesen wirst, bist du auch heute immer noch der Grund, warum ich mich für die Biologie entschieden habe. In meinen letzten Arbeiten habe ich immer geschrieben: *„Ich hoffe, du wärst stolz auf mich.“*
Heute bin ich mir sicher: Du bist es!
Alles in allem ist das hier für dich.

In Gießen ist der Elefant sehr froh, denn dort gibt es ein Elefantenklo. Zwar stört ihn dessen Baustruktur, doch sagt er sich: «Hauptsache Kultur».

– Prof. Dr. Volkmar Wolters, JLU Gießen;
(Aus dem Zyklus der Elefantengedichte)



"Kaja", Opelzoo (2025); *(Photo: Linda Fenske)*