MICROBIAL GENOMICS

RESEARCH ARTICLE
Kovacec *et al*., *Microbial Genomics* 2024;10:001297
DOI 10.1099/mgen.0.001297

Revisiting typing systems for group B *Streptococcus* prophages: an application in prophage detection and classification in group B *Streptococcus* isolates from Argentina

Veronica Kovacec[1], Sabrina Di Gregorio[1,2], Mario Pajon[1], Chiara Crestani[3], Tomás Poklepovich[4], Josefina Campos[4,†], Uzma Basit Khan[5], Stephen D. Bentley[5], Dorota Jamrozy[5], Marta Mollerach[1,2] and Laura Bonofiglio[1,2,*]

Abstract

Group B *Streptococcus* (GBS) causes severe infections in neonates and adults with comorbidities. Prophages have been reported to contribute to GBS evolution and pathogenicity. However, no studies are available to date on the presence and diversity of prophages in GBS isolates from humans in South America. This study provides insights into the prophage content of 365 GBS isolates collected from clinical samples in the context of an Argentinean multicentric study. Using whole-genome sequence data, we implemented two previously proposed methods for prophage typing: a PCR approach (carried out *in silico*) coupled with a BLASTX-based method to classify prophages based on their prophage group and integrase type, respectively. We manually searched the genomes and identified 325 prophages. However, only 80% of prophages could be accurately categorized with the previous approaches. Integration of phylogenetic analysis, prophage group and integrase type allowed for all to be classified into 19 prophage types, which correlated with GBS clonal complex grouping. The revised prophage typing approach was additionally improved by using a BLASTN search after enriching the database with ten new genes for prophage group classification combined with the existing integrase typing method. This modified and integrated typing system was applied to the analysis of 615 GBS genomes (365 GBS from Argentina and 250 from public databases), which revealed 29 prophage types, including two novel integrase subtypes. Their characterization and comparative analysis revealed major differences in the lysogeny and replication modules. Genes related to bacterial fitness, virulence or adaptation to stressful environments were detected in all prophage types. Considering prophage prevalence, distribution and their association with bacterial virulence, it is important to study their role in GBS epidemiology. In this context, we propose the use of an improved and integrated prophage typing system suitable for rapid phage detection and classification with little computational processing.

Received 29 May 2024; Accepted 30 August 2024; Published 17 October 2024

Author affiliations: [1]Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Instituto de Bacteriología y Virología Molecular, Buenos Aires, Argentina; [2]CONICET, Buenos Aires, Argentina; [3]Global Health Department, Institut Pasteur, Paris, France; [4]Unidad Operativa Centro Nacional de Genómica y Bioinformática, ANLIS Dr. Carlos G. Malbrán, Buenos Aires, Argentina; [5]Parasites and Microbes Programme, Wellcome Sanger Institute, Cambridgeshire, UK.

***Correspondence:** Laura Bonofiglio, lbonofi@ffyb.uba.ar

Keywords: GBS; genomic analysis; group B *Streptococcus*; phage typing; prophage; *Streptococcus agalactiae*.

Abbreviations: ARG, antibiotic resistance gene; CC, clonal complexes; GBS, group B *Streptococcus*; MATE, multidrug and toxic compound extrusion; ND, not determined.

†Present address: International Pathogen Genomic Surveillance Network (IPSN), WHO Berlin Hub, Berlin, Germany.

The datasets generated for this study can be found in the ENA repository, under study accession PRJEB34470 (https://www.ebi.ac.uk/ena/browser/view/PRJEB34470) and PRJEB78912 (https://www.ebi.ac.uk/ena/browser/view/PRJEB78912) and the individual run accessions are detailed in Table S6 in Supplementary Material 2. Sequences of the 29 reference prophages can be found in the FigShare repository: https://doi.org/10.6084/m9.figshare.27068074. Sequences of the 325 prophages obtained from Argentinian GBS can be found at the FigShare repository: https://doi.org/10.6084/m9.figshare.26521381. All supporting data and protocols have been provided within the article or through the supplementary material.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary material files. Five supplementary figures and six supplementary tables are available with the online version of this article.

Impact Statement

It has been proposed that prophage acquisition played a role in the emergence of *Streptococcus agalactiae* [Group B *Streptococcus* (GBS)] as a human pathogen in European countries. Further study and characterization of prophages of GBS from around the world and their role in GBS epidemiology are necessary. In this work, we propose a new typing system that allows for the rapid detection and classification of GBS prophages based on their phylogenetic lineage and integration site. This methodology, applied to the genomes of 615 GBS globally distributed, revealed a much larger diversity of GBS prophages than those detected by existing tools for GBS-prophage screening and classification. Furthermore, our research increases the current knowledge about GBS prophages by the characterization of each prophage type, the search for genes potentially beneficial for GBS and the analysis of prophage distribution according to the GBS clonal complex. This study also provides insights into GBS-prophage evolution by the comparative analysis of their modules and the study of their phylogenetic relation with prophages of other streptococcal species. The groundwork hereby presented will contribute to future projects exploring the prophage-mediated mechanisms underlying GBS evolution and epidemiology.

DATA SUMMARY

The supplementary details on the methodology and supplementary figures and data can be found in Supplementary Material 1. Supplementary tables can be found in Supplementary Material 2. The datasets generated for this study can be found in the ENA repository, under study accession numbers PRJEB34470 (https://www.ebi.ac.uk/ena/browser/view/PRJEB34470) and PRJEB78912 (https://www.ebi.ac.uk/ena/browser/view/PRJEB78912), and the individual run accessions are detailed in Table S6. Sequences of the 29 reference prophages can be found in the FigShare repository: https://doi.org/10.6084/m9.figshare.27068074. Sequences of the 325 prophages obtained from Argentinian GBS can be found at the FigShare repository: https://doi.org/10.6084/m9.figshare.26521381. All datasets analysed in this study are detailed in the Supplementary Materials. The links of the microreact projects created are:

https://microreact.org/project/philogeny-argentinean-gbs-prophages

https://microreact.org/project/philogeny-of-modular-genes

https://microreact.org/project/gbs-prophages-in-a-global-context.

INTRODUCTION

Streptococcus agalactiae [group B *Streptococcus* (GBS)] is a commensal bacterium that colonizes the human intestinal and genitourinary tracts. GBS is a major cause of neonatal sepsis and other perinatal infections, such as meningitis and pneumonia, globally [1]. In recent decades, invasive infections caused by GBS in non-pregnant adults have been on the rise, especially in elderly people and those suffering from underlying medical conditions [2–4].

Prophages are important vehicles for horizontal gene transfer [5] and can constitute up to 20% of a bacterial genome. Prophages play a significant role in bacterial evolution by introducing genes that enhance bacterial fitness and virulence [6–8]. Furthermore, pathogenic strains tend to carry more phage-related genes than non-pathogenic strains [9–11], which was also observed for GBS [12].

GBS temperate bacteriophages (lysogenic prophages) were first described in 1969 in strains of bovine origin [13]. Recent studies on human GBS isolates revealed an association between certain prophages (some of possible animal origin) and the emergence of specific pathogenic GBS clones among isolates recovered from neonates and adults in Europe [12, 14–16]. Little is known about the epidemiology of GBS prophages and their impact on pathogenicity in other geographical areas. To date, there are no reports of prophages in GBS isolates from South America [17].

Two approaches have been previously developed for the screening and classification of prophages in GBS genomes, one based on full-prophage sequence diversity [12] and the other based on integrase typing [17]. However, both have limitations and may underestimate or overestimate prophage presence.

This study aims to analyse the prophage content in GBS genomes from Argentina by integrating and improving the existing methods for the detection and typing of GBS prophages, providing a novel strategy for global surveillance of GBS prophage epidemiology and diversity.

METHODS

Isolate collection

We collected 450 GBS isolates from the maternal carriage (100/450) as well as from invasive (162/450) and urinary tract (188/450) infections as part of a national multicentric study that involved 40 health centres in 12 provinces of Argentina between 2014 and 2015. The invasive infections were defined by the isolation of GBS from a normally sterile body site. All isolates had been characterized phenotypically (antibiotic susceptibility and serotyping), and invasive isolates had also been characterized genotypically (PFGE) [4, 18, 19].

Whole-genome sequencing and data processing

Genomic DNA of 450 GBS isolates was extracted using a QIACube HT protocol and sequenced at the Wellcome Sanger Institute on the Illumina NovaSeq 6000 platform (as part of our collaboration with the Juno consortium, https://www.gbsgen.net/). For ten GBS isolates, genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega) and sequenced at the Malbrán Institute on Illumina MiSeq. The quality of the reads was assessed with FastQC v0.11.7 [20] and Kraken v0.10.6 [21]. *De novo* assemblies were obtained with SPAdes v3.12.0 [22] and quality checked with Quast v5.0.0 [23]. The 365/450 assemblies that passed the quality controls were annotated with Prokka v1.12 [24]. Multilocus sequence types (MLSTs) were determined with the software mlst v2.22.1 [25] and assigned to clonal complexes (CC) using the PubMLST website [26, 27] (https://pubmlst.org/organisms/streptococcus-agalactiae).

Prophage detection and typing

The methodology followed for the detection and typing of the GBS prophages is summarized in this section and Fig. 1. For detailed information, see 'Detailed Materials and Methods, Section M3' in Supplementary Material 1.

In the first instance, prophage sequences were detected and classified using the previous screening methods for the prophage groups [12] (here performed *in silico*) and integrase type [17]. The results of the two methods were combined, and the putative prophages were preliminarily classified into prophage types according to the prophage group and integrase type (Fig. 1).

Prophage sequences within the assembled genomes were manually searched and extracted. Prophages fragmented across contigs were reconstructed by *de novo* assembly against reference prophages. All prophage sequences were annotated (Fig. 1).

The extracted prophage sequences were aligned, and a phylogenetic tree was reconstructed. Prophages that could not be assigned to a prophage group during the initial screening stage were classified with the same group as the prophages in their phylogenetic cluster. Classification by the prophage type was updated accordingly (Fig. 1).

Improvement of the prophage typing system

The methods followed for the improvement of the prophage typing system are summarized in this section and Fig. 1. For detailed information, see 'Detailed Materials and Methods, Section M4' in Supplementary Material 1.

In order to avoid false-positive and false-negative results obtained in the initial screening, we improved the detection of prophage groups by performing a BLASTN search against a curated database of prophage group-specific genes (Table S1, available in the online Supplementary Material 2). The methodology was tested on the 22 prophages detected by van der Mee-Marquet *et al.* [12], all prophages from Argentinean GBS genomes and 615 GBS complete genomes from Argentina and public databases (Table S2). A result was considered positive when at least one of the genes for the prophage group was detected with a minimum of 75% identity and coverage. Prophage types were then defined combining these results with those of integrase types.

Prophage characterization

The steps followed for prophage characterization are summarized in this section and Fig. 2. For detailed information, see 'Detailed Materials and Methods, Section M5' in Supplementary Material 1 and 'Improvement of the screening and typing system Section' in Results.

Prophage sequences were searched for genetic determinants of virulence and antimicrobial resistance, as well as any genes potentially beneficial for the host bacteria. Genes coding for integrase, helicase, terminase large subunit, major capsid protein and lysin were used for the phylogenetic analysis of each prophage module.

One phage of each prophage type (*n* = 29, see Results section) was selected for further characterization. The morphology of the prophages was determined by the recognition of their head-neck-tail modules. The function of the genes annotated as encoding hypothetical proteins was predicted based on their conserved domains. The catalytic domains of the putative integrases and lysins were analyzed. Comparative sequence analysis was performed to study the genetic differences between prophages of the same prophage group but different integrase types and vice versa.

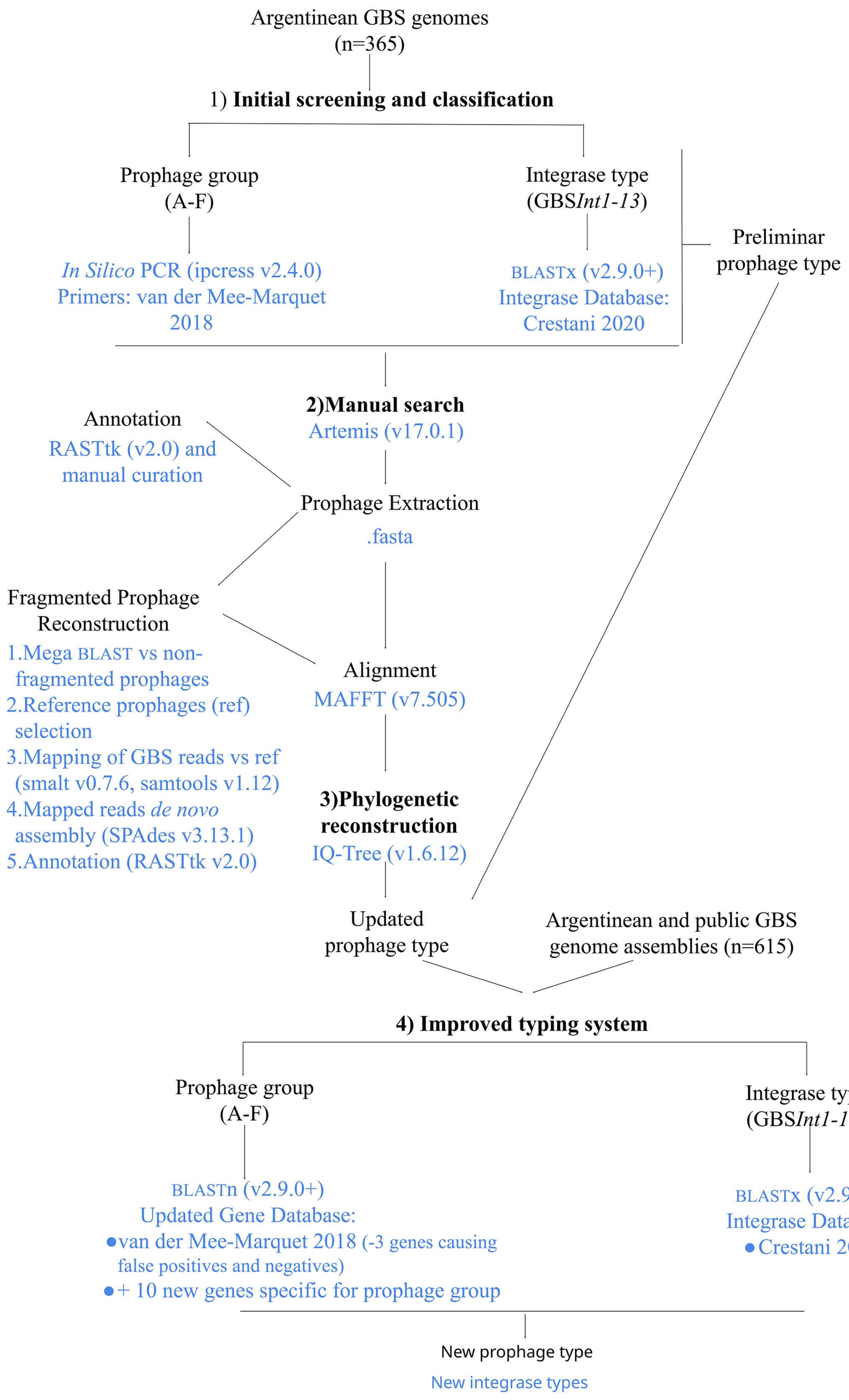
Argentinean GBS genomes
(n=365)
1) Initial screening and classification
Prophage group
(A-F)
Integrase type
(GBSInt1-13)
Preliminar prophage type
In Silico PCR (ipcress v2.4.0)
Primers: van der Mee-Marquet 2018
BLASTx (v2.9.0+)
Integrase Database: Crestani 2020
2)Manual search
Artemis (v17.0.1)
Annotation
RASTtk (v2.0) and manual curation
Prophage Extraction
.fasta
Fragmented Prophage Reconstruction
1.Mega BLAST vs non-fragmented prophages
2.Reference prophages (ref) selection
3.Mapping of GBS reads vs ref (smalt v0.7.6, samtools v1.12)
4.Mapped reads de novo assembly (SPAdes v3.13.1)
5.Annotation (RASTtk v2.0)
Alignment
MAFFT (v7.505)
3)Phylogenetic reconstruction
IQ-Tree (v1.6.12)
Updated prophage type
Argentinean and public GBS genome assemblies (n=615)
4) Improved typing system
Prophage group
(A-F)
Integrase type
(GBSInt1-13)
BLASTn (v2.9.0+)
Updated Gene Database:
●van der Mee-Marquet 2018 (-3 genes causing false positives and negatives)
●+ 10 new genes specific for prophage group
BLASTx (v2.9.0+)
Integrase Database:
●Crestani 2020
New prophage type
New integrase types

Fig. 1. Summarized methodology used for GBS-prophage detection, typing and improvement of the prophage typing system.

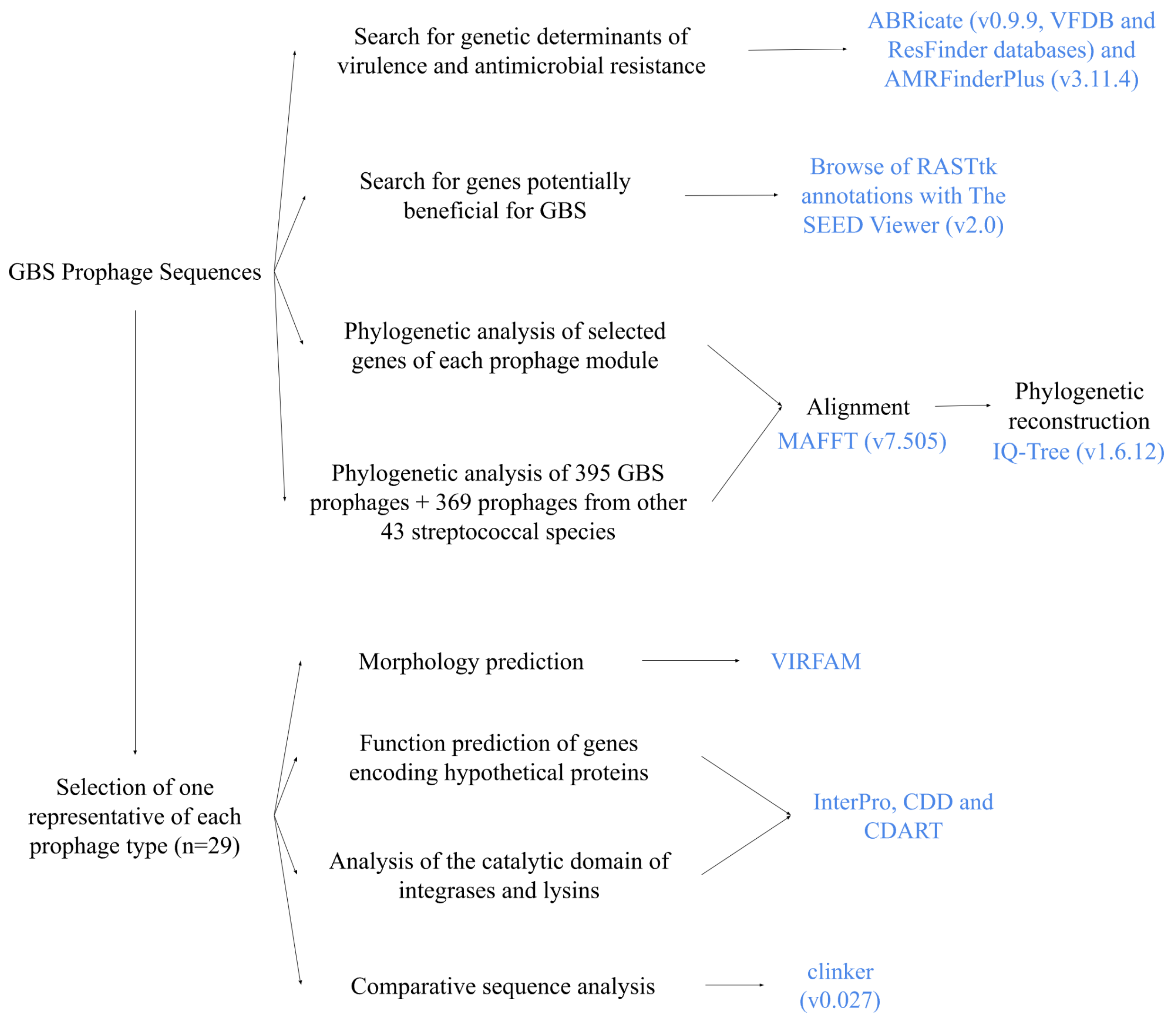
GBS Prophage Sequences
Search for genetic determinants of virulence and antimicrobial resistance
ABRicate (v0.9.9, VFDB and ResFinder databases) and AMRFinderPlus (v3.11.4)
Search for genes potentially beneficial for GBS
Browse of RASTtk annotations with The SEED Viewer (v2.0)
Phylogenetic analysis of selected genes of each prophage module
Phylogenetic analysis of 395 GBS prophages + 369 prophages from other 43 streptococcal species
Alignment
MAFFT (v7.505)
Phylogenetic reconstruction
IQ-Tree (v1.6.12)
Selection of one representative of each prophage type (n=29)
Morphology prediction
VIRFAM
Function prediction of genes encoding hypothetical proteins
Analysis of the catalytic domain of integrases and lysins
InterPro, CDD and CDART
Comparative sequence analysis
clinker (v0.027)

Fig. 2. Summarized methodology used for GBS-prophage characterization.

To provide a broader context to the prophages from Argentinian GBS, a phylogenetic analysis of 764 prophages from GBS (isolated in Argentina and worldwide) and other 43 streptococcal species was performed (Table S3).

Integration of information

The Microreact application [28] was used for an integral visualization of the collected information.

Statistical analysis

Fisher's exact test (two-tailed) was used to evaluate the association between prophage presence and GBS CC. A *P*-value of ≤0.05 was considered to be significant.

RESULTS

Prophage detection and typing in GBS genomes from Argentina

Prophage screening based on the prophage phylogenetic group and integrase type [12, 17] detected 383 putative prophages in the 365 GBS genomes. A total of 200/383 (52%) prophages were grouped into ten prophage types (prophage group + integrase type). In 60/383 (16%) prophages, only the integrase type was determined [the prophage group was not determined (ND)]. A total of 123/383 prophages (32%) belonged to prophage group A, known to lack the integrase gene [8, 17].

In contrast, a manual search revealed 325 prophages among the 365 GBS genomes. Forty-two of the 325 prophages (13%) were fragmented in two or more contigs but were successfully assembled with reference-based read mapping. The additional prophage sequences detected based on the screening approach (*n* = 58) represented mostly groups A (36/58 lacked the lysis module so were considered prophage remnants) and F (in 19/58 only the integrase gene was present), both cases being considered false positives. In 3/58 cases, the screened prophages were disregarded due to the high level of fragmentation.

Only the 325 manually detected prophage sequences were analysed further. Phylogenetic analysis revealed clustering by the prophage group (Figs 3 and S1A). Prophages that could not be typed by group with the initial screening (ND prophages) were assigned a prophage group corresponding to their phylogenetic cluster. Prophage groups E and F were further divided into two subclusters each (100% bootstrap support). In both cases, prophages of one subcluster had not been detected by the *in silico* PCR. Those that were detected by the *in silico* PCR were reclassified as subgroups E1 and F1, respectively, while those not detected were reclassified as subgroups E2 and F2, respectively (Fig. S1B). The reclassification of group F prophages into the subgroups F1 and F2 coincides with van der Mee-Marquet's classification of prophages with insertion sites F1 and F2, respectively [12].

As a result of this integrated analysis, all prophages (*n* = 325) were reclassified into 19 prophage types (Fig. 3). The majority of GBS genomes (54%) carried a single prophage. Multiple integrase types/subtypes were found within prophage groups and vice versa (Fig. 3), as previously described [17], indicating that screening of GBS prophages by previous methods alone is insufficient to accurately classify the prophages.

Improvement of the screening and typing system

In order to improve the screening and typing system for GBS prophages, we developed a database of phylogroup-specific prophage genes (Table S1, Data S1). The genes were selected based on 12 PCR-amplified fragments using previously described primers [12] and refined to avoid false-positive results (detection of prophage remnants or isolated prophage genes). Detection of group A prophage is based on gene *hha*I or *clp*P combined with the presence of either a holin or a lysin gene (Table S1). The F1 prophage integrase gene (*hin*) was replaced by a gene coding for a terminase large subunit (Table S1). The hypothetical prophage gene representative of group D prophages was replaced by two new genes (Table S1). Finally, gene sequences for the detection of groups E2 and F2 prophages were also added to the database (Table S1). These changes markedly improved the specificity and accuracy of prophage detection and typing in the 365 GBS genomes from Argentina (Fig. 4).

More importantly, our improved method detected ten additional prophage types when tested on a collection of 615 globally derived GBS genomes (including the 365 from Argentina), giving a total of 29 distinct prophage types. The ten prophage types not found among the GBS genomes from Argentina included two new integrase subtypes, GBS*Int*6.3 and GBS*Int*8.2, and their sequences were added to the integrase database (Data S2).

A graphical summary of the proposed improved screening and typing method for GBS prophages is shown in Fig. 5.

Characterization of GBS prophages from Argentina

The most prevalent prophage types in GBS isolates from Argentina were A (87/325, 27%), E1/GBS*Int*3 (85/325, 26%), E2/GBS*Int*3 (42/325, 13%) and F1/GBS*int*11.2 (28/325, 9%), (Fig. 4). Significant associations were found between most prophage types and CC assignments ($P < 0.05$, Fig. S2). Prophages of type A were associated with CC23 and CC1; B/GBS*Int*9.2 with CC17; C/GBS*Int*4 with CC17 and CC26; D/GBS*Int*1, D/GBS*Int*2.2 and E2/GBS*Int*3 with CC19; D/GBS*Int*8.1 with CC103; E1/GBS*Int*3 with CC23 and CC452; E1/GBS*Int*8.1 with CC23; and F1/GBS*Int*11.2 with CC12.

The phylogenetic analysis (Fig. 3) revealed monophyletic clusters for most prophages within a group or subgroup with the exception of group D prophages, which were more diverse.

The insertion site and *att* sequences matched those described by Crestani *et al.* [16] for each integrase type, with the exception of some prophages with minor changes in their *att* sequences (Table S4). One gene of each prophage module (integrase, helicase, terminase large subunit, major capsid protein and lysin) was selected for phylogenetic analysis of their nucleotide sequences to determine whether similar modules could be found in different prophage types. These analyses showed, in general, similar clustering between phylogenies based on individual prophage module genes and that observed based on the alignment of full prophage sequences (Fig. S3).

Several genes potentially beneficial for the host bacteria were found exclusively in B/GBS*Int*9.2 prophages, including genes involved in carbohydrate or RNA metabolism, which were present in all prophages of this type. A single B/GBS*Int*9.2 prophage additionally carried genes coding for phosphoenolpyruvate synthase, a multidrug resistance efflux pump, genes involved in DNA metabolism and several genes encoding different types of permeases (Table S5).

Phylogenetic analysis of 764 *Streptococcus* spp. prophages revealed that GBS phages from Argentina were related to other GBS prophages and with bacteriophages from other streptococcal species (Fig. 6). In particular, group B prophages were closely related to *S. pyogenes*, *S. iniae*, *S. oralis* and *S. pneumoniae* phages, while groups A and F prophages were related to phages from more than ten different streptococcal species. However, prophages from groups C, D and E were more closely related to each other in this phylogeny

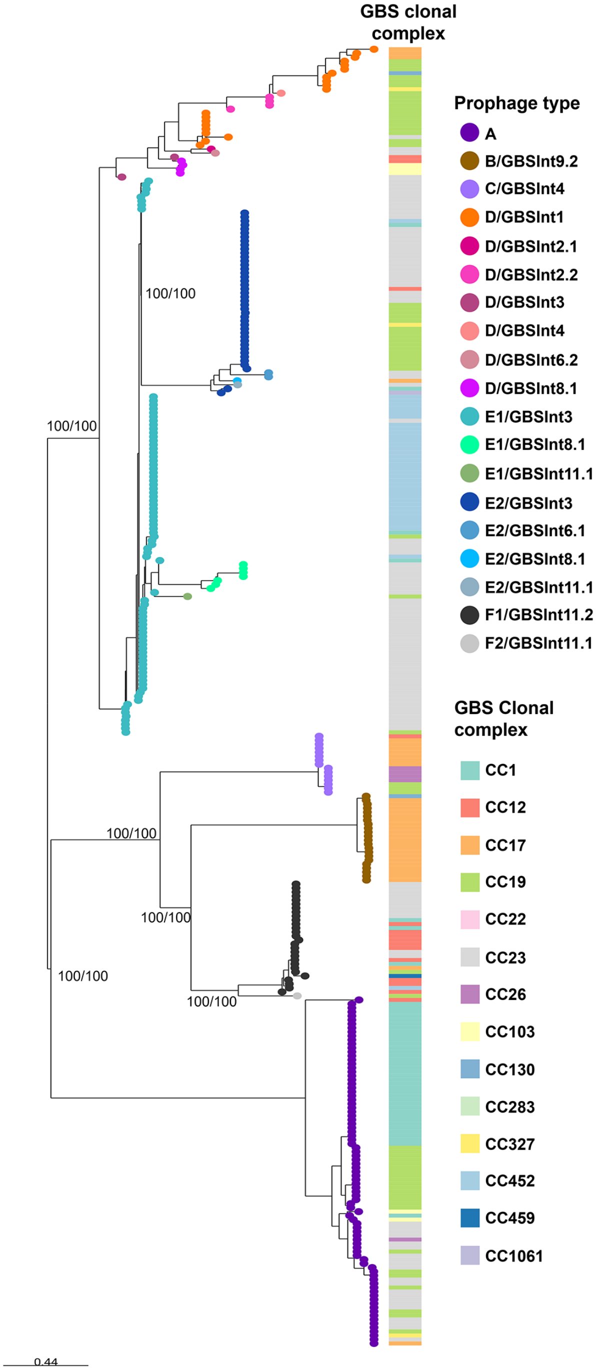

Fig. 3. Phylogeny of 325 prophages found in 365 Argentinian GBS genomes. Maximum-likelihood phylogenetic tree, midpoint rooted, with nodes coloured by prophage type (determined based on the combination of prophage group and integrase type). Support values (SH-aLRT/ Ultrafast Bootstrap) are shown as labels for selected nodes. Host GBS CC is shown as coloured blocks. The scale bar represents the number of SNPs per variable site: https://microreact.org/project/philogeny-argentinean-gbs-prophages.

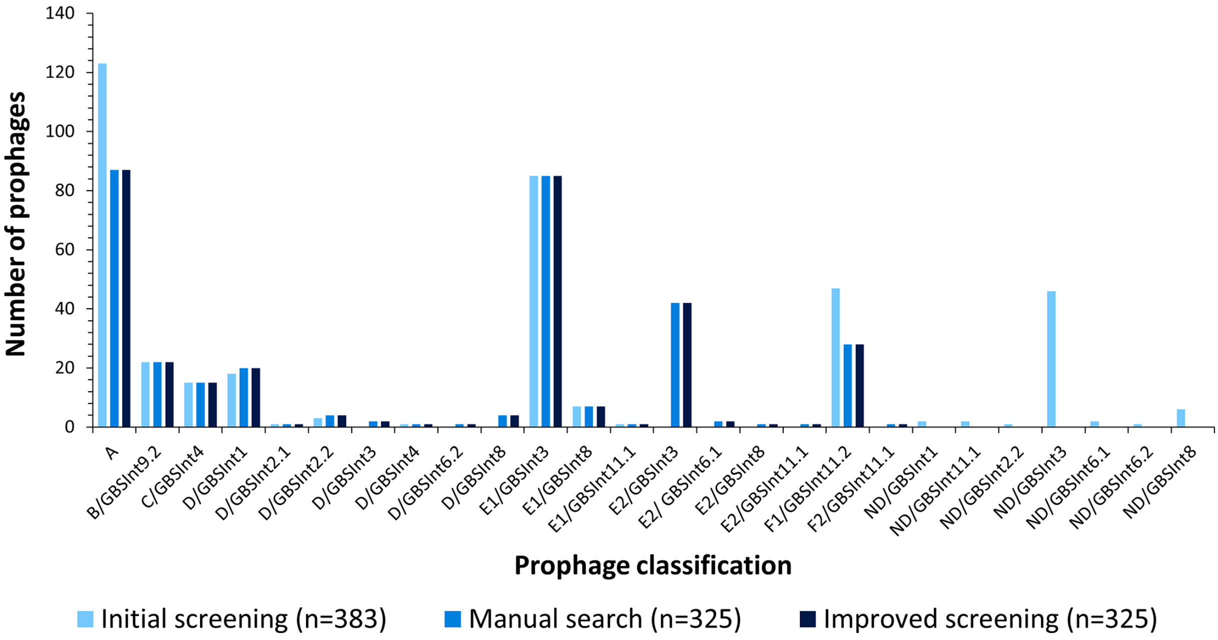

Fig. 4. Prophage detection and typing in 365 Argentinian GBS genomes. Three methods are compared: the initial integrated screening of the prophage group by *in silico* PCR (with primers designed by van der Mee-Marquet *et al.*) and integrase type by BLASTX search against a GBS-prophage integrase database (built by Crestani *et al.*); the manual search in the GBS genomes of the screened prophages and their classification according to their phylogeny; and the improved screening and typing method by prophage group: BLASTN search against a database of the representative genes for each group proposed by van der Mee-Marquet *et al.*, and new genes proposed in this work, integrated with the integrase typing designed by Crestani *et al.*

(100% bootstrap support) and seem to be mainly restricted to GBS, with an occasional spread to *S. dysgalactiae*, *S. equi*, *S. iniae* and *S. pyogenes* (clustering with group C) and one single *S. uberis* prophage (clustering with group E1).

Comparative analysis of the 29 distinct prophage types

More than one integrase type or subtype was found in all prophage groups, with the exception of group A (prophages without integrase). Integrase types GBS*Int*1, GBS*Int*3, GBS*Int*4, GBS*Int*8.1 and GBS*Int*11.1 were found in more than one prophage group, mostly D and E. Also, D and E2 prophages had the most diversity in integrase types (7 and 6 types, respectively). Group B and F prophages had two subtypes of the same integrase type (GBS*Int*9 and GBS*Int*11, respectively).

One prophage sequence for each of the 29 types was selected for further characterization. Based on their predicted morphology, all phages were classified as siphovirus (former *Siphoviridae* family). Visual inspection of the annotated prophage sequences confirmed the modular organization following a specific order based on their function in the phage life cycle: lysogeny, replication, packaging, morphogenesis and host lysis (Fig. 7). Genome sizes ranged from 32.6 to 47.9 Kb, with group B prophages being the smallest (likely due to shorter genes and a shorter morphogenesis module) and the prophages of groups D and E1 the largest (likely due to higher gene content in the lysogeny and replication modules). Manual inspection of the sequences revealed that all prophage types contained ORFs encoding putative proteins related to bacterial fitness, defence mechanisms and/or virulence (Table S4). ORFs coding for hypothetical proteins where no known conserved domain was found constituted from 29 to 59% of the coding sequences of each prophage. These genes were found throughout the entire prophage genome and not organized in a single module, although they were more frequent in the lysogeny and replication modules.

The amino acid sequences of each integrase type, lysins found in each of the 29 reference prophages and lysins coded by genes from each phylogenetic cluster (Fig. S3 (E)) were analysed to determine their catalytic domains (Table S4). All integrase types had a tyrosine recombinase domain, except for GBS*Int*11.1 and GBS*Int*11.2, which had a serine recombinase domain. The putative lysins were classified based on their cleavage site into three of the five major endolysin classes [29]: *N*-acetyl-β-D-glucosaminidase, *N*-acetyl-β-D-muramidase and *N*-acetylmuramoyl-L-alanine amidase. Interestingly, lysins from all prophages contained the domain *N*-acetylmuramoyl-L-alanine amidase and, in most cases, the lysins were bifunctional, as they also carried a second catalytic domain with a different cleavage site, either *N*-acetyl-β-D-glucosaminidase or *N*-acetyl-β-D-muramidase. All prophages from the same type had lysins of the same class, except for types A, C/GBS*Int*4, D/GBS*Int*3, E1/GBS*Int*3, E2/GBS*Int*4, E2/GBS*Int*11.1 and F1/GBS*Int*11.2, in which some lysins had only the *N*-acetylmuramoyl-L-alanine amidase domain, while others were bifunctional.

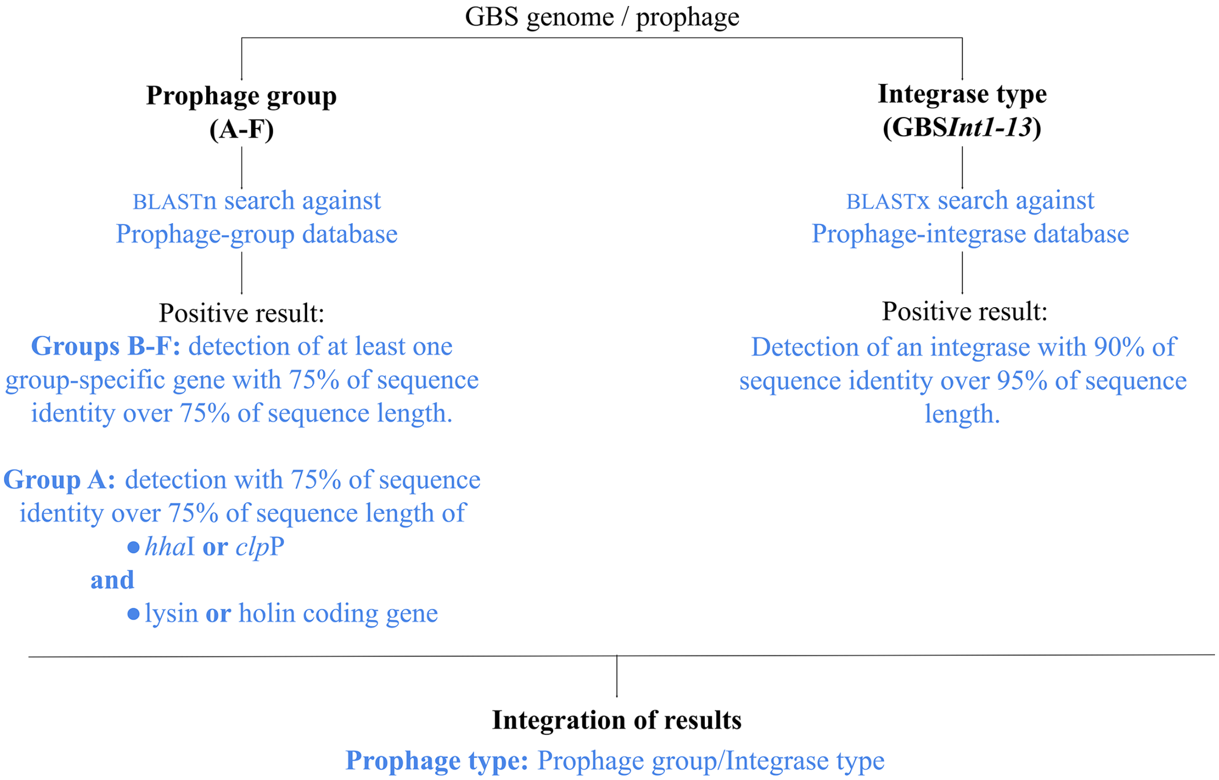
GBS genome / prophage
Prophage group (A-F)
BLASTn search against Prophage-group database
Positive result:
Groups B-F: detection of at least one group-specific gene with 75% of sequence identity over 75% of sequence length.
Group A: detection with 75% of sequence identity over 75% of sequence length of
● hhaI or clpP
and
● lysin or holin coding gene
Integrase type (GBSInt1-13)
BLASTx search against Prophage-integrase database
Positive result:
Detection of an integrase with 90% of sequence identity over 95% of sequence length.
Integration of results
Prophage type: Prophage group/Integrase type

Fig. 5. Methodology summary for the improved screening and typing system for GBS prophages.

Comparative sequence analysis of the 29 prophage types (Fig. 7) showed more than 60% identity between the morphogenesis modules encoding most tail proteins of the groups D, E and F and around 45% identity between most genes in the replication modules of the type A, D/GBS*Int*4 and D/GBS*Int*2.2. The group F prophages shared more similarities on a modular level with groups D and E than phylogenetically closer groups B and C, which did not show identity on a modular level with other prophage types.

Comparative analysis of prophages belonging to the same prophage group but carrying different integrase types or subtypes (Fig. S4) revealed that the major differences between the prophages within groups B, C, D and E and subgroup F2 resided in the lysogeny and replication modules. The F1 prophage showed less than 60% identity with subgroup F2 between all their modules except for the morphogenesis. In addition, prophages with the same integrase type but belonging to different prophage groups (Fig. S5) did not reveal any similarity other than the integrase gene itself. The exceptions were prophages with integrase type GBS*Int*4, where more than 90% of identity was found between the lysogeny module and the genes following the host lysis module (Fig. S5 (C)).

DISCUSSION

The presence of prophages has been reported to impact GBS epidemiology in the collections from Europe [12, 14–16]. In particular, the acquisition of certain prophages has been linked with the emergence of GBS infections in neonates and adults in France [12], and it has been postulated that the presence of prophages encoding virulence factors was responsible for the increased incidence of severe neonatal infections both in France [8, 14] and the Netherlands [15].

This is the first report on the diversity of prophages in human GBS isolates from South America. Employing two methods for GBS prophage typing [12, 17] and subsequent manual inspection, we were able to detect 325 prophages among 365 GBS isolates collected in an Argentinian multicentric study.

New prophage typing system

The use of standard prophage detection softwares, such as PHASTER [30], Prophinder [31] or PhiSpy [32], had not proven accurate for prophage detection in GBS genomes in our initial search for prophages in a small number of genomes (data not shown). In some cases, prophage presence was predicted in a region where all the genes were bacterial or along 2 contigs ordered consecutively by size but which were not consecutive in the real genome; in others, regions with the presence of prophages were not detected (and were later detected manually). When a region with an actual prophage was identified, it generally did not accurately indicate the beginning or

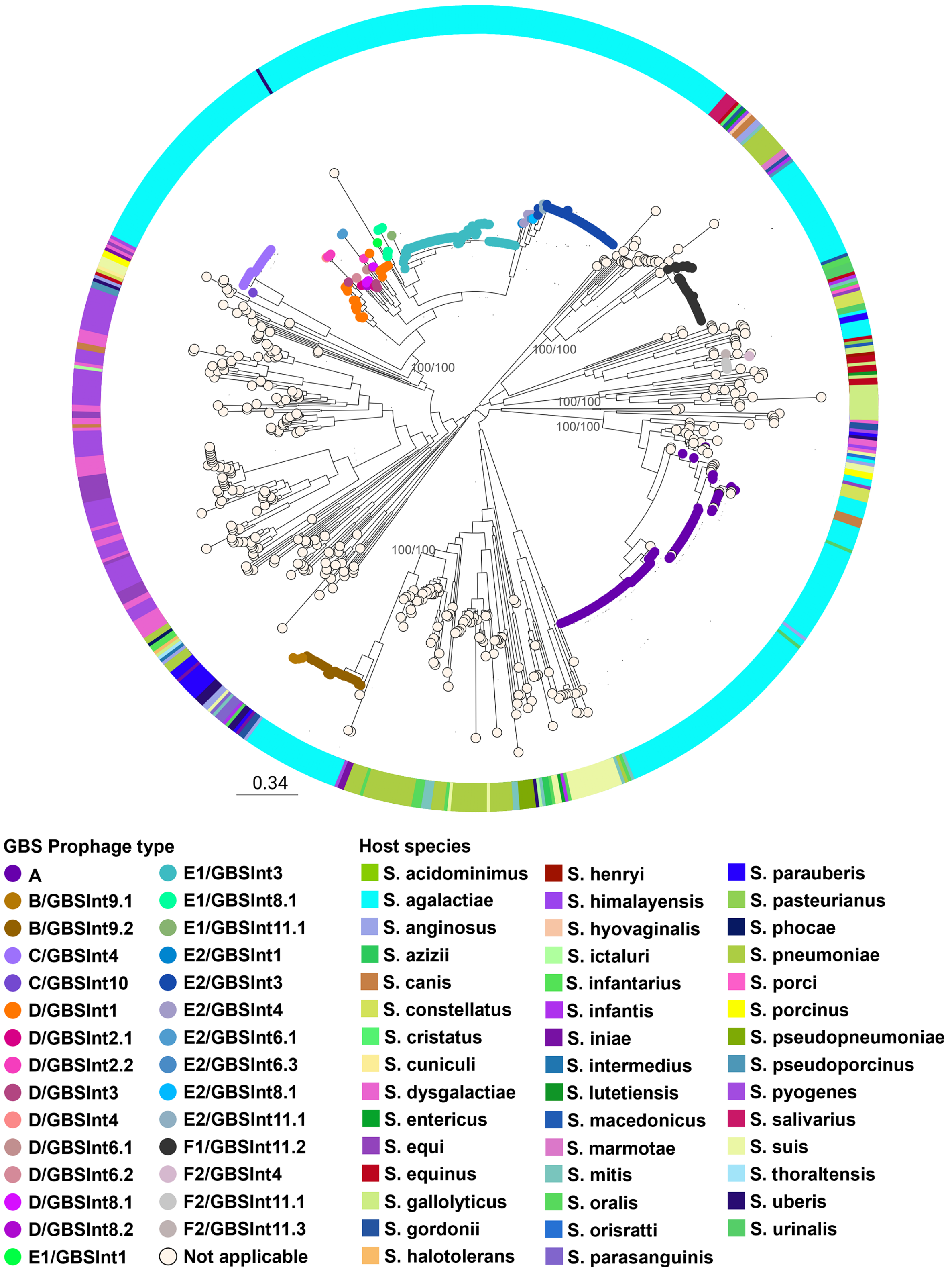

Fig. 6. Phylogenetic tree of prophages from streptococcal species. Maximum-likelihood phylogenetic tree of 764 prophages found in 580 genomes from 44 streptococcal species, including GBS. Nodes are coloured by the GBS prophage type, where applicable. Host species are shown in the coloured ring. The tree was rooted at the midpoint. Support values (SH-aLRT/ Ultrafast Bootstrap) are shown as labels for selected nodes. The scale bar represents the number of SNPs per variable site: https://microreact.org/project/gbs-prophages-in-a-global-context.

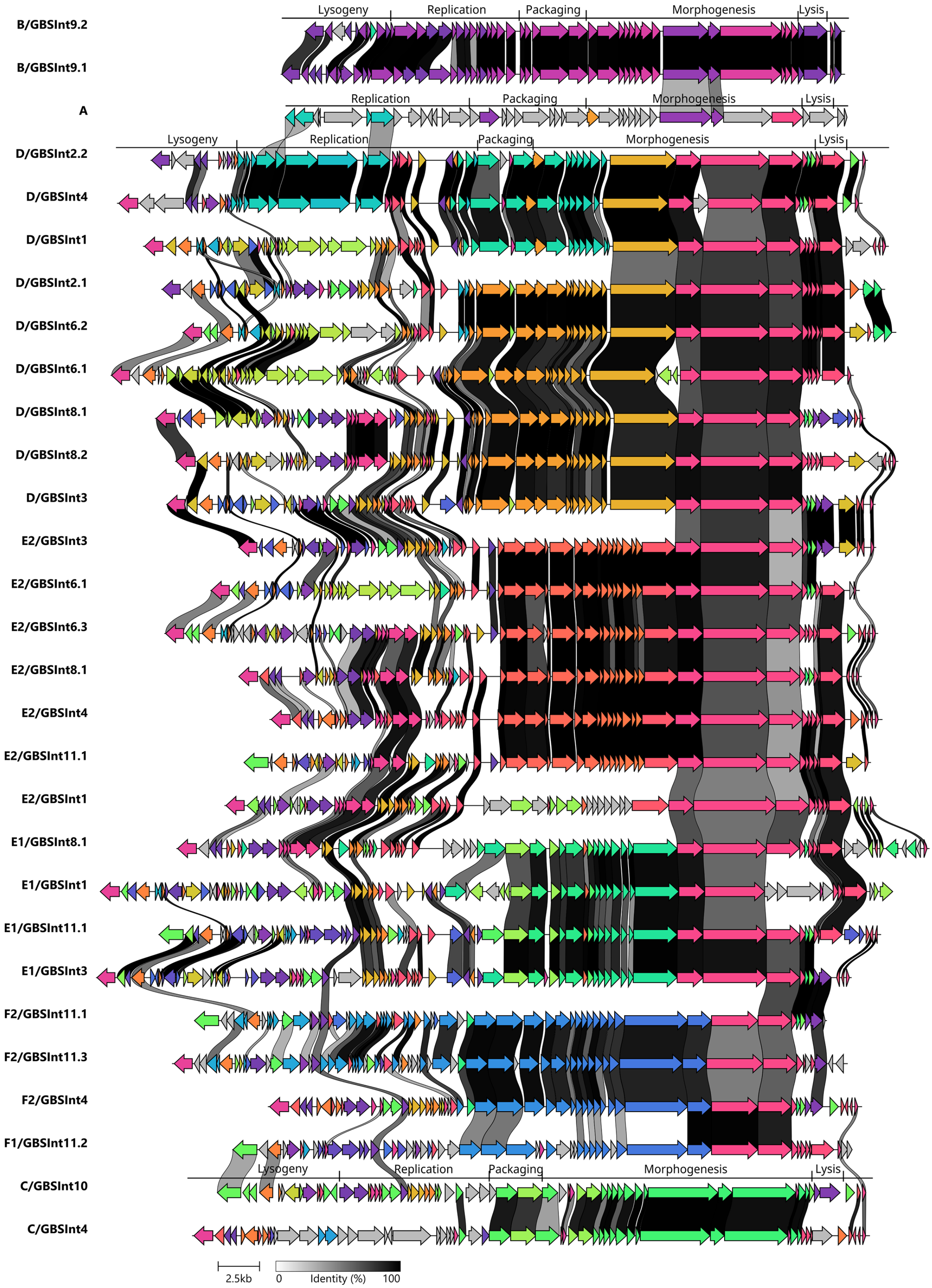

Fig. 7. Comparative analysis of representative sequences of 29 distinct prophage types found in the analysed GBS. The colours represent groups of homologous genes. Genes with more than 40% identity are linked with grey-black strokes, as shown in the scale.

ending of the prophage. In the case of PHASTER specifically, the prophages att sites were never accurately identified and, in general, there was no correlation between the classification according to phage integrity and what was actually observed in the sequences. For example, it classified coding regions with only bacterial genes as complete prophages, or it classified prophages with all phage modules as incomplete. For all these reasons we did not keep on using these methods for phage detection in our whole collection of GBS genomes and resorted to the screening methods available in literature that were designed for prophage detection in GBS, allowed for their classification according to phylogenetic group [12] or integrase type [17], which do not often correlate [17, this work], and were more practical for screening of a large volume of genomes.

In this work, we propose an enhanced method that combines and simplifies some screening steps of previously developed GBS-prophage typing methods. Our results show that using each method individually can lead to false-negative or false-positive results. Furthermore, it can be insufficient to classify GBS prophages accurately, as determining the prophage group provides information about the genomic characteristics of the phage but not its insertion site. The latter can be determined based on integrase type, which in contrast does not offer insights into the composition of the full prophage sequence.

The new method increases the detection of full prophage sequences, as well as prophages that are fragmented into multiple contigs. Additionally, it allows the classification of prophages based on both their phylogenetic lineage and integration site within the bacterial genome. Using this approach, we were able to identify a total of 29 distinct prophage types, including 19 prophage types in GBS from Argentina and 10 additional types in prophages from GBS collected in other countries.

Our results demonstrate that this improved integrated method is less likely to detect prophage remnants, allows the identification of novel prophage integrases and offers fast detection of GBS prophages in a large genomic dataset with little computational processing.

Evolution of GBS prophages

Genome mosaicism was observed in all prophage types, in accordance with the proposed modular evolution of prophages [33–35]. Genes belonging to the packaging module, those encoding capsid proteins and a few coding for tail proteins appear to have been acquired as a block, independently of the rest of the prophage genes. This is especially evident in groups with a greater number of prophage types, D and E, since the clustering of phages into phylogenetic subclades (Fig. 3) correlates with their grouping according to the homology in the aforementioned gene region (Fig. 7). The presence of homology between several genes encoding tail proteins in prophages D, E and F suggests that this group of genes share a common ancestor and that they might have been acquired in an independent recombination event from the rest of the structural genes, which did not present homology between the different phage groups (Fig. 7). In general, prophages within the same group presented the same classes of lysins (same catalytic domains, Table S4), even if they did not share high homology in their lysis modules. However, there was no lysin class exclusive to one prophage group, which could indicate an independent recombination of the lysin-coding gene from the rest of the lysis module.

The divergence between prophages, even belonging to the same phylogenetic group, was typically observed in the lysogeny and replication modules, where the majority of the genes encoding hypothetical proteins and genes potentially beneficial for GBS were located. This suggests that these regions would be the most prone to suffer recombination events. This is also supported by the lack of similarities in the lysogeny module of prophages with the same integrase type but different prophage group, which implies that the prophage integration site in the host genome is dependent only on the integrase gene and *att* sequence. The general absence of homology between prophages of groups A, B and C with other prophage types would suggest that these prophage groups have little propensity for horizontal gene exchange with GBS prophages from other groups.

Most prophage groups and subgroups seem to be highly clonal, whereas group D showed more divergence, demonstrating evidence of microevolution (Figs 3, 6 and S1A). The reason for such variability is not clear yet, but future studies are needed to understand if it might be advantageous to the host.

The level of conservation of genes encoding tail proteins among various prophage groups (Fig. 7) could mean that they are involved in the specific recognition between the phage and the bacterial receptor, a process that has not yet been studied in GBS. If so, these phages could have a similar host range. Interestingly, tail protein genes are also conserved in prophage types from other streptococcal species (*S. pyogenes* and *S. pneumoniae*) [36, 37]. Further analysis of the tail protein sequence conservation can reveal new insights into the mechanisms of prophage sharing between *Streptococcus* spp.

Prophage presence in the context of GBS epidemiology

The prevalence of GBS isolates carrying at least one prophage and the average number of prophages per isolate found in our dataset correlate with previous reports on the prophage content in GBS [12, 14, 16, 38, 39]. Our results also aligned with the integrase type and CC associations reported previously by Crestani *et al.* [17]. The reason for such association should be explored but could be related to the presence of specific restriction-modification systems, which may limit the recombination and horizontal transfer of mobile genetic elements between GBS lineages, as seen in other bacterial species [40–42].

In line with previous reports, GBS prophages identified in our dataset carried genes potentially involved in bacterial fitness, host adaptation to stressful environments and virulence [8, 12, 14, 16, 43]. However, most of these genes were not detected by the online prophage genome-screening tools, which stresses how manual curation remains an important part of the prophage genome investigations. A single B/GBS*Int*9.2 prophage was found to carry a gene encoding a multidrug and toxic compound extrusion (MATE)-like protein, similar to MepA from *Staphylococcus aureus* [44]. It has been reported that the expression of MATE proteins in bacteria confers resistance to toxic dyes and multiple antibiotics [45, 46]. While carriage of antibiotic resistance genes (ARGs) is not common in prophages [47], there have been reports of ARG carriage in conjugative elements within prophages of streptococcal species [48, 49]. The fact that these elements were found in only a single GBS prophage analysed in this study highlights the overall low prevalence of ARGs in GBS prophages. However, it is important to note that, on average, 50% of GBS prophage genes code for hypothetical proteins lacking recognizable conserved domains. Hence, it is important to continue the study of these prophages to better understand their biology and impact on the host cells, particularly in the context of bacteriophage therapy development [50].

The GBS carriage of prophages associated with those from other streptococcal species causing infections in humans and animals has been extensively documented in the literature [12, 16, 38, 51, 52]. Our results further show that prophages detected in GBS isolates from Argentina are globally distributed and suggest that prophages belonging to groups A, B and F might have evolved from prophages horizontally transferred between different species of streptococci (Fig. 6). In contrast, prophage groups C, D and E, which were found to share a common ancestor in this global phylogeny, seem to be mainly restricted to GBS (Fig. 6).

Further experimental work is needed in order to confirm the prophage transfer between streptococcal species, receptor specificity (tail proteins) and the restriction of C, D and E prophages to only GBS. It is also crucial to investigate if the insertion sites of these prophages favour the mobilization of genetic material and the potential for horizontal transfer of genes present in defective prophages or phage remnants.

Study limitations

Limitations of this study include using GBS genome assemblies from short-read data to detect and classify prophages. All prophage sequences identified after manual curation remain theoretical, even when found in the same contig, due to the possibility of assembly errors, particularly in prophage modules with high sequence similarity between different phage types (e.g. structural modules). This could be addressed by performing long-read sequencing followed by a hybrid assembly.

CONCLUSIONS

This study performed a comprehensive comparative analysis of prophage genomes in GBS isolates from Argentina and is the first report on GBS prophage diversity in Latin America. We propose the use of an improved and integrated prophage typing system suitable for rapid phage detection in GBS genomes and their classification with little computational processing. The presence of prophages in most GBS isolates analysed here, the association between prophage groups and GBS lineages and carriage of genes beneficial for the host bacteria reinforce the hypothesis that the acquisition of prophages confers an evolutionary advantage to GBS and may play an important role in its epidemiology. The diversity of prophage types found in GBS isolates from Argentina along with the observed lysin diversity is a promising finding, which can be explored further to identify novel lysins with activity against GBS as an alternative therapy for GBS infections. In light of the global challenge posed by antimicrobial-resistant bacteria, it is imperative to advance our current knowledge of bacteriophage biology and the applicability of lysins as an alternative treatment against bacterial infections to promote and expedite the approval and regulation of such therapies.

Funding information

This study was funded by grants from Universidad de Buenos Aires (UBACYT 20020190100189BA and UBACYT 20020170200303BA) to LB, from CONICET (11220150100692CO) to MM, and The Bill and Melinda Gates Foundation (INV-010426) to SB.

Acknowledgements

We want to thank everyone involved in the genome sequencing of our GBS collection at the sequence facilities of the Wellcome Sanger Institute and Instituto Malbrán.

Author contributions

V.K., S.D.G., M.M. and L.B. conceived this study and L.B. supervised it. T.P., J.C., U.B.K., D.J. and S.D.B. conducted and coordinated the whole genome sequencing. V.K. and S.D.G. designed the bioinformatic analyses. V.K., M.P. and S.D.G. performed the acquisition of the data. V.K. and S.D.G. conducted the data analysis. S.D.G. guided and supervised the bioinformatic analysis. C.C. participated in the discussion of the results and gave critical insights for their analysis. V.K. wrote the original draft of the manuscript. All authors reviewed, edited and approved the final version of the manuscript.

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethical statement

Ethical approval for the 'Argentinian multicentric study on infections due to *Streptococcus agalactiae*' was provided by the Ethics Committee of the Faculty of Pharmacy and Biochemistry, University of Buenos Aires, Res [D] RESCD-2022-400-E-UBA-DCT.

References

1. Alotaibi NM, Alroqi S, Alharbi A, Almutiri B, Alshehry M, *et al*. Clinical characteristics and treatment strategies for group B *Streptococcus* (GBS) infection in pediatrics: a systematic review. *Medicina* 2023;59:1279.
2. Björnsdóttir ES, Martins ER, Erlendsdóttir H, Haraldsson G, Melo-Cristino J, *et al*. Changing epidemiology of group B streptococcal infections among adults in Iceland: 1975-2014. *Clin Microbiol Infect* 2016;22:379..
3. Navarro-Torné A, Curcio D, Moïsi JC, Jodar L. Burden of invasive group B *Streptococcus* disease in non-pregnant adults: a systematic review and meta-analysis. *PLoS One* 2021;16:e0258030.
4. Arias B, Kovacec V, Vigliarolo L, Suárez M, Tersigni C, *et al*. Epidemiology of invasive infections caused by *Streptococcus agalactiae* in Argentina. *Microb Drug Resist* 2022;28:322–329.
5. Juhas M. Horizontal gene transfer in human pathogens. *Crit Rev Microbiol* 2015;41:101–108.
6. Davies EV, Winstanley C, Fothergill JL, James CE. The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol Lett* 2016;363:fnw015.
7. Khan A, Wahl LM. Quantifying the forces that maintain prophages in bacterial genomes. *Theor Popul Biol* 2020;133:168–179.
8. Renard A, Diene SM, Courtier-Martinez L, Gaillard JB, Gbaguidi-Haore H, *et al*. 12/111phiA prophage domestication is associated with autoaggregation and increased ability to produce biofilm in *Streptococcus agalactiae Microorganisms* 2021;9:1112.
9. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, *et al*. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 2001;8:11–22.
10. Winstanley C, Langille MGI, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, *et al*. Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool epidemic strain of *Pseudomonas aeruginosa*. *Genome Res* 2009;19:12–23.
11. Busby B, Kristensen DM, Koonin EV. Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environ Microbiol* 2013;15:307–312.
12. van der Mee-Marquet N, Diene SM, Barbera L, Courtier-Martinez L, Lafont L, *et al*. Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of Group B *Streptococcus* species. *Clin Microbiol Infect* 2018;24:514–521.
13. Russell H, Norcross NL, Kahn DE. Isolation and characterization of *Streptococcus agalactiae* bacteriophage. *J Gen Virol* 1969;5:315–317.
14. Renard A, Barbera L, Courtier-Martinez L, Dos Santos S, Valentin A-S, *et al*. phiD12-Like livestock-associated prophages are associated with novel subpopulations of *Streptococcus agalactiae* infecting neonates. *Front Cell Infect Microbiol* 2019;9:166.
15. Jamrozy D, Bijlsma MW, de Goffau MC, van de Beek D, Kuijpers TW, *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;10:9539.
16. 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.
17. Crestani C, Forde TL, Zadoks RN. Development and application of a prophage integrase typing scheme for group B *Streptococcus*. *Front Microbiol* 1993;11.
18. Arias B, Kovacec V, Vigliarolo L, Suárez M, Tersigni C, *et al*. Fluoroquinolone-resistant *Streptococcus agalactiae* invasive isolates recovered in Argentina. *Microb Drug Resist* 2019;25:739–743.
19. Vigliarolo L, Arias B, Suárez M, Van Haute E, Kovacec V, *et al*. Argentinian multicenter study on urinary tract infections due to *Streptococcus agalactiae* in adult patients. *J Infect Dev Ctries* 2019;13:77–82.
20. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
21. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* 2013;63:41–49.
22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
23. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;29:1072–1075.
24. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
25. Seemann T. mlst Github; 2016. https://github.com/tseemann/mlst
26. 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.
27. 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.
28. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, *et al*. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016;2.
29. Love MJ, Abeysekera GS, Muscroft-Taylor AC, Billington C, Dobson RCJ. On the catalytic mechanism of bacteriophage endolysins: opportunities for engineering. *Biochim Biophys Acta Proteins Proteom* 2020;1868:140302.
30. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–21.
31. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 2008;24:863–865.
32. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* 2012;40:e126.
33. Botstein D. A theory of modular evolution for bacteriophages. *Ann N Y Acad Sci* 1980;354:484–490.
34. Lima-Mendez G, Toussaint A, Leplae R. A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res Microbiol* 2011;162:737–746.
35. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* 2020;18:125–138.
36. Canchaya C, Proux C, Fournous G, Bruttin A, Brüssow H. Prophage genomics. *Microbiol Mol Biol Rev* 2003;67:238–276.
37. Garriss G, Henriques-Normark B. Lysogeny in *Streptococcus pneumoniae*. *Microorganisms* 2020;8:1546.
38. 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.
39. Sirimanapong W, Phước NN, Crestani C, Chen S, Zadoks RN. Geographical, temporal and host-species distribution of potentially human-pathogenic group B *Streptococcus* in aquaculture species in southeast Asia. *Pathogens* 2023;12:525.

40. McCarthy AJ, Witney AA, Lindsay JA. *Staphylococcus aureus* temperate bacteriophage: carriage and horizontal gene transfer is lineage associated. *Front Cell Infect Microbiol* 2012;2:6.
41. Oliveira PH, Touchon M, Rocha EPC. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 2016;113:5658–5663.
42. DebRoy S, Shropshire WC, Tran CN, Hao H, Gohel M, *et al*. Characterization of the type I restriction modification system broadly conserved among group A *Streptococci*. *mSphere* 2021;6:e0079921.
43. Furfaro LL, Payne MS, Chang BJ. Host range, morphological and genomic characterisation of bacteriophages with activity against clinical *Streptococcus agalactiae* isolates. *PLoS One* 2020;15:e0235002.
44. McAleese F, Petersen P, Ruzin A, Dunman PM, Murphy E, *et al*. A novel MATE family efflux pump contributes to the reduced susceptibility of laboratory-derived *Staphylococcus aureus* mutants to tigecycline. *Antimicrob Agents Chemother* 2005;49:1865–1871.
45. Claxton DP, Jagessar KL, Mchaourab HS. Principles of alternating access in Multidrug and Toxin Extrusion (MATE) transporters. *J Mol Biol* 2021;433:166959.
46. Huang H, Wan P, Luo X, Lu Y, Li X, *et al*. Tigecycline resistance-associated mutations in the MepA efflux pump in *Staphylococcus aureus*. *Microbiol Spectr* 2023;11:e0063423.
47. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, *et al*. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* 2017;11:237–247.
48. Dai X, Sun J, Zhu B, Lv M, Chen L, *et al*. Various mobile genetic elements involved in the dissemination of the phenicol-oxazolidinone resistance gene *optrA* in the zoonotic pathogen *Streptococcus suis*: a nonignorable risk to public health. *Microbiol Spectr* 2023;11:e0487522.
49. Santoro F, Pastore G, Fox V, Petit M-A, Iannelli F, *et al*. *Streptococcus pyogenes* Φ1207.3 is a temperate bacteriophage carrying the macrolide resistance gene pair *mef*(A)-*msr*(D) and capable of lysogenizing different Streptococci. *Microbiol Spectr* 2023;11:e0421122.
50. Monteiro R, Pires DP, Costa AR, Azeredo J. Phage therapy: going temperate? *Trends Microbiol* 2019;27:368–378.
51. Bai Q, Zhang W, Yang Y, Tang F, Nguyen X, *et al*. Characterization and genome sequencing of a novel bacteriophage infecting *Streptococcus agalactiae* with high similarity to a phage from *Streptococcus pyogenes*. *Arch Virol* 2013;158:1733–1741.
52. Rezaei Javan R, Ramos-Sevillano E, Akter A, Brown J, Brueggemann AB. Prophages and satellite prophages are widespread in *Streptococcus* and may play a role in pneumococcal pathogenesis. *Nat Commun* 2019;10:4852.