Sequence-Function Relationships in Phage-Encoded Bacterial Cell Wall Lytic Enzymes and Their Implications for Phage-Derived Product Design

Roberto Vázquez,[a,b] Ernesto García,[a,b] Pedro García[a,b]

aDepartamento de Biotecnología Microbiana y de Plantas, Centro de Investigaciones Biológicas Margarita Salas, Madrid, Spain
bCentro de Investigación Biomédica en Red de Enfermedades Respiratorias, Madrid, Spain

ABSTRACT Phage (endo)lysins are thought to be a viable alternative to usual antibiotic chemotherapy to fight resistant bacterial infections. However, a comprehensive view of lysins' structure and properties regarding their function, with an applied focus, is somewhat lacking. Current literature suggests that specific features typical of lysins from phages infecting Gram-negative bacteria (G−) (higher net charge and amphipathic helices) are responsible for improved interaction with the G− envelope. Such antimicrobial peptide (AMP)-like elements are also of interest for antimicrobial molecule design. Thus, this study aims to provide an updated view on the primary structural landscape of phage lysins to clarify the evolutionary importance of several sequence-predicted properties, particularly for the interaction with the G− surface. A database of 2,182 lysin sequences was compiled, containing relevant information such as domain architectures, data on the phages' host bacteria, and sequence-predicted physicochemical properties. Based on such classifiers, an investigation of the differential appearance of certain features was conducted. This analysis revealed different lysin architectural variants that are preferably found in phages infecting certain bacterial hosts. In particular, some physicochemical properties (higher net charge, hydrophobicity, hydrophobic moment, and aliphatic index) were associated with G− phage lysins, appearing specifically at their C-terminal end. Information on the remarkable genetic specialization of lysins regarding the features of the bacterial hosts is provided, specifically supporting the nowadays-common hypothesis that lysins from G− usually contain AMP-like regions.

IMPORTANCE Phage-encoded lytic enzymes, also called lysins, are one of the most promising alternatives to common antibiotics. The potential of lysins as novel antimicrobials to tackle antibiotic-resistant bacteria not only arises from features such as a lower chance to provoke resistance but also from their versatility as synthetic biology parts. Functional modules derived from lysins are currently being used for the design of novel antimicrobials with desired properties. This study provides a view of the lysin diversity landscape by examining a set of phage lysin genes. We have uncovered the fundamental differences between the lysins from phages that infect bacteria with different superficial architectures and, thus, the reach of their specialization regarding cell wall structures. These results provide clarity and evidence to sustain some of the common hypotheses in current literature, as well as making available an updated and characterized database of lysins sequences for further developments.

KEYWORDS endolysins, bacteriophages, bacteriophage therapy, genomics, bioinformatics, antimicrobial agents

Since the antibiotic pipeline started drying up, a worrying increase in the antibiotic-resistant fraction of bacterial populations has been reported (1, 2), and high

percentages of antibiotic-resistant populations have been sustained (3, 4). If this situation continues, the cost, both economic and in human lives, will be enormous due to the lack of effective treatments (5, 6). This has prompted the interest in novel antimicrobial development by many public health actors, such as international oversight organizations (7), public health and disease control agencies (3, 4), governments (8), researchers, and several companies (9). Some of the current efforts to gather a new antimicrobial armamentarium have led science toward bacteriophages (10, 11).

To allow the dissemination of their progeny, double-stranded DNA phages provoke lysis of the bacterial host, which is fundamentally accomplished by degradation of the peptidoglycan. This polymer is an essential constituent of the bacterial cell wall, and the breakage of specific bonds within its three-dimensional mesh leads to bacterial death, largely by osmotic shock. The main phage molecule responsible for peptidoglycan degradation is the lysin (also referred to as endolysin). Lysins are released toward their polymeric target, usually with the assistance of another kind of protein, the holin, which create pores in the plasma membrane and thus allow lysin leakage to the periplasm (12). Other phage products collaborate in hampering the cell wall in some of its particular settings; for example, lysins B detach the arabino-mycolyl outer layer of mycobacteria and their relatives (e.g., *Rhodococcus* and *Corynebacterium*) (13). In addition, in some Gram-negative bacteria (G−), effective lysis also needs the concurrent participation of additional phage products named spanins (14). This reveals the significant amount of genetic resources allocated by phages to overcome the barriers that the bacterial cell walls represent.

In addition to using whole phage particles as therapeutic agents against bacterial infections (so-called phage therapy), current efforts also point out to artificially repurposing certain phage products, such as lysins, as antimicrobials (enzybiotics) (11, 15, 16). The concept is rather simple: the external addition of purified lysins to a susceptible bacterium would cause bacterial lysis whenever the lysin degrades the peptidoglycan. This process is straightforward in the case of Gram-positive bacteria (G+), and the therapeutic effect of enzybiotics on G+ has been fully confirmed experimentally (15). The most important characteristics that make enzybiotics amenable to be developed as therapeutics are (i) a certain specificity for the original bacterial host and some closely related bacteria, which would prevent normal microbiota from being harmed (16, 17), or, conversely, the possibility of broad-range lysins, if needed (18); (ii) a lower chance of provoking the appearance of resistant bacteria, which is speculated to be due to the essential nature of the highly conserved peptidoglycan (that is, changes in its structure lead to a decreased fitness and/or virulence) (19); (iii) the expectation that neither adverse immune responses nor production of *in vivo* neutralizing antibodies will occur, possibly due to the presence of phages—and their products— among the normal microbiota in humans, and the rapid mode of action that is thought to avoid antibody neutralization *in vivo* (20). Moreover, lysins are amenable to protein engineering strategies (18, 21–24). Typically, the architectural organization of lysins comprises one (or more) enzymatically active domain (EAD) together with a cell wall-binding domain (CWBD) or, sometimes, just an EAD (Fig. 1). Therefore, synthetic biology strategies, such as the construction of completely new lysins made up of different modules as building blocks, have been shown to be achievable. Such strategies enable the design and production of tailor-made antimicrobials, based on the conjunction of diverse functions of interest into a single protein. Functions of interest may include, besides a catalytic activity against the peptidoglycan network (i.e., an antimicrobial activity), increased stability in complex media (25), or, more typically, a certain tropism toward a specific element on the bacterial surface (26) or some other macromolecules, like cellulose (27). The engineering approaches mentioned above have circumvented the alleged inability of lysins to cross the outer membrane (OM) of G− (28, 29). Different kinds of synthetic lysins have been devised to that end. Among them are the so-called artilysins, which are lysins fused to different kinds of membrane-permeabilizing peptides (30), the lysocins, which are lysins fused to elements from bacteriocins that enable bacterial surface recognition and import
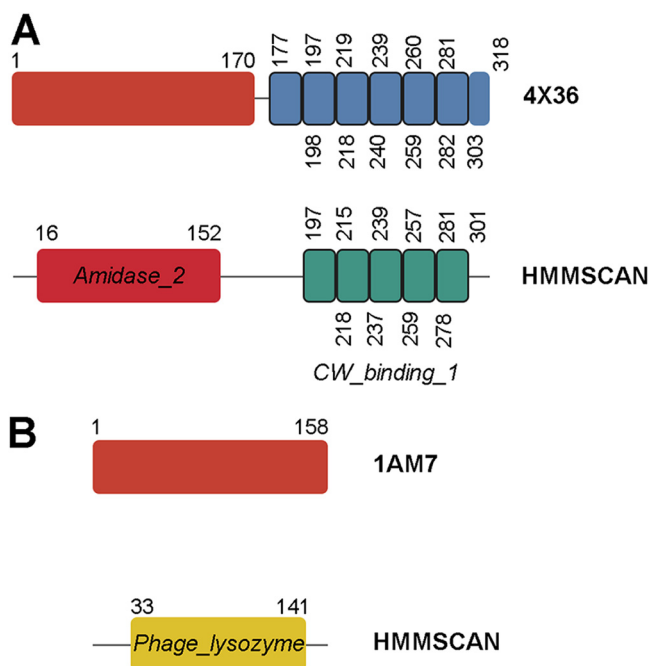
FIG 1 Schematic examples of the architecture of peptidoglycan hydrolases. The top row in each panel corresponds to the domain structure based on the three-dimensional folding according to the corresponding PDB entry (red boxes are EADs; blue boxes are CWBD repeats). In the bottom rows, the domain structure predicted using HMMSCAN is shown. Numbers indicate the amino acid coordinates. (A) LytA autolysin from *S. pneumoniae*. Note that only five choline-binding repeats (*CW_binding_1*) of the six present in the three-dimensional structure are predicted. (B) Lysozyme from lambda coliphage.

into the periplasm (22), and the innolysins, lysins fused to phage receptor-binding proteins (31).

However, some lysins also exhibit intrinsic bactericidal activity on G− (32–34). This activity was first noticed for the T4 phage lysozyme (35) and several *Pseudomonas aeruginosa* phage lysins (36). This unexpected property was attributed to nonenzymatic mechanisms, previously observed in partially denatured hen egg white lysozyme (37), and relies on the presence of antimicrobial peptide (AMP)-like subdomains within such lysins, usually at a C-terminal position (32, 38). Recently, it has been suggested that such AMP-like elements are widespread among lysins from phages infecting G− and that they might cooperate in host lysis by providing an additional affinity for the OM, because of their high net charge in comparison with the negatively charged lipopolysaccharides, and might play a role in lysis itself by exerting a membrane-destabilizing effect (28, 33, 39–41). However, how widespread this trait is has not been yet properly examined, and therefore, its true functional and evolutionary implications are largely unknown. Of note, such AMP-like elements have been successfully used to design AMPs that are active on their own (36, 39, 42).

To uncover the actual evolutionary relevance of these AMP-like elements, as well as other lysin features, such as their domain architecture, in this work, a bioinformatic approach examining a wide collection of lysins was proposed. There are several precedents for the application of homology-based analysis of putative lysin sequences that have paved the way to the systemic comprehension of the coevolution of phage lysins and their hosts (13, 43). The present study aimed to update the picture with the latest available information, as well as to provide answers to the recent questions brought forward by the literature regarding lysin engineering. Therefore, based on current knowledge and available genomic data, we have constructed and curated a comprehensive database of phage lysin sequences. Subsequent analyses on the data included (i) an initial exploration of the database composition, (ii) a cross-reference of

information added to the database to check for differential distribution of distinct domain families and their architectural combinations across different bacterial groups, and (iii) an overview of easily computable physicochemical properties (net charge, hydrophobicity, etc.) along amino acid (aa) sequences to explore widespread, relevant differences between groups. The conclusions presented here will strengthen our understanding of lysins' specificity and variability and help in future drug design efforts based on phage products.

RESULTS AND DISCUSSION

Outline. A total of 9,539 genomes were prospectively obtained from the National Center for Biotechnology Information (NCBI) database (retrieved on April 2020). After a careful curation process (for details, see Materials and Methods), the final database contained 2,182 proteins and a total of 3,303 Pfam (PF) hits (Table 1). Each of these sequences was associated with a bacterial genus corresponding to its described host, for which data on its Gram group and peptidoglycan chemotype were added (Table 1). In total, our database comprised phage lysins from 47 bacterial genera, accounting for up to a total of 2,179 sequences, plus three lysin sequences from PRD1-like phages that infect several enterobacteria. Taking into account all of the identical sequences, the 2,182 different sequences of our data set correspond to 36,365 entries in the NCBI Reference Sequence database (RefSeq; release 202) (44). It should be mentioned that the screening method is biased by the annotation of the source genomes and by current knowledge about the functional domains present in lysins (i.e., by the present set of Pfam-identified lysin functional domains). Therefore, while the final curated data set will be useful to produce a broad view of our current understanding of phage lysins, generalizations should be interpreted with care, since future works may clarify some of the still-unknown aspects of lysin's composition and function.

General differences among lysins. For 1,512 of 2,182 sequences (69.3%), only one significant PF hit could be predicted (Fig. 2A). This was especially relevant for lysins from phages infecting G−, given that 90.6% of these proteins were predicted to contain a single functional domain. Nearly 60% of the lysins from phages infecting G+ (for the sake of this work, mycobacteria and their relatives like *Rhodococcus* or *Corynebacterium* were included among G+) harbored only one predicted functional domain. Few lysins appear to contain ≥4 PF hits (Fig. 2A). However, these figures should be considered with caution, since they do not correspond to the number of real functional modules within the protein but to a relatively high number (up to 5) of individual repeats that, together, make up a single functional module. For example, the 37 sequences with 6 PF hits correspond to streptococcal phage lysins having the typical structure [EAD]5×[CW_binding_1], with the EAD being either *Amidase_2* (31 hits), *Glyco_hydro_25* (3 hits), or *CHAP* (3 hits) domains. Likewise, not all sequences with a single PF hit should be assumed to contain only a single domain, since many of them might contain other, yet-undefined domains. Also, some repeats (or even full domains) might not be appropriately predicted if there is enough evolutionary sequence divergence. As an example, the domain structure based upon the three-dimensional folding of pneumococcal major autolysin LytA (45) does not concur with the domains predicted by a homology search, since this method is unable to uncover the latest CWBD repeat (Fig. 1A).

As a whole, however, the differential relative amount of single and multiple PF hit sequences between G− and G+ phage lysins (Fig. 2A and B) can be taken into account, in accordance with the usual proposal that G− lysins are typically monomodular, while G+ ones are multimodular (46). Figure 2B shows that some "structural" domains were detected among the lysins from G−-infecting phages. These four entries, bearing domains such as *Gp5_C* or *Gp5_OB*, are probably virion-associated lysins that were kept through the curation process. The mainly monomodular nature of G− lysins is further supported by the evident difference in protein length distributions (Fig. 2C), where G+ phage lysins tend to be larger (median = 317 aa residues) than G− ones (median = 164 aa residues), and also by the differential distribution of sequence lengths before and

TABLE 1 Classifiers assigned to host genera and yield of the curation process

Host genus or group	Gram group	Outer architecture	Chemotype	No. of:			
				Retrieved phage genomes	Genes detected by annotation	Genes after curation (final)	PF hits
Achromobacter	−	Diderm (lipopolysaccharide)	A1γ	36	32	10	13
Acidovorax	−	Diderm (lipopolysaccharide)	A1γ	3	4	2	2
Acinetobacter	−	Diderm (lipopolysaccharide)	A1γ	139	152	50	57
Aeromonas	−	Diderm (lipopolysaccharide)	A1γ	95	77	26	27
Bacteroides	−	Diderm (lipopolysaccharide)	A1γ	5	6	3	3
Burkholderia	−	Diderm (lipopolysaccharide)	A1γ	72	61	27	36
Caulobacter	−	Diderm (lipopolysaccharide)	A1γ	42	58	17	17
Cellulophaga	−	Diderm (lipopolysaccharide)	A1γ	52	38	9	9
Cronobacter	−	Diderm (lipopolysaccharide)	A1γ	45	64	24	26
Enterobacter	−	Diderm (lipopolysaccharide)	A1γ			1	1
Enterobacteria	−	Diderm (lipopolysaccharide)	A1γ	797	300	3	3
Enterovibrio	−	Diderm (lipopolysaccharide)	A1γ			2	2
Escherichia	−	Diderm (lipopolysaccharide)	A1γ	684	745	211	231
Fusobacterium	−	Diderm (lipopolysaccharide)	A1γ[a]	4	3	3	3
Klebsiella	−	Diderm (lipopolysaccharide)	A1γ	266	259	67	71
Microcystis	−	Diderm (lipopolysaccharide)	A1γ	8	8	5	5
Pseudomonas	−	Diderm (lipopolysaccharide)	A1γ	625	485	89	98
Ralstonia	−	Diderm (lipopolysaccharide)	A1γ	68	21	14	15
Salmonella	−	Diderm (lipopolysaccharide)	A1γ	414	444	90	93
Serratia	−	Diderm (lipopolysaccharide)	A1γ	31	40	13	16
Shewanella	−	Diderm (lipopolysaccharide)	A1γ			1	1
Shigella	−	Diderm (lipopolysaccharide)	A1γ			3	3
Sphingobium	−	Diderm (lipopolysaccharide)	A1γ	2	2	1	2
Sphingomonas	−	Diderm (glycosphingolipid)	A1γ	7	2	1	1
Stenotrophomonas	−	Diderm (lipopolysaccharide)	A1γ	28	25	6	8
Vibrio	−	Diderm (lipopolysaccharide)	A1γ	578	225	76	85
Yersinia	−	Diderm (lipopolysaccharide)	A1γ	70	62	16	16
Arthrobacter	+	Monoderm	A3α[b]	316	400	103	161
Bacillus	+	Monoderm	A1α	348	400	150	286
Bifidobacterium[c]	+	Monoderm	A4α	10	2	1	1
Clostridioides	+	Monoderm	A1γ			23	25
Clostridium	+	Monoderm	A1γ	109	174	30	43
Corynebacterium	+	Diderm (mycolic acid)	A1γ	40	30	16	24
Cutibacterium	+	Monoderm	A3γ			65	65
Enterococcus	+	Monoderm	A3α (*E. faecalis*), A4α (*E. faecium*)[d]	127	150	50	77
Lacticaseibacillus	+	Monoderm	A4α			13	28
Lactiplantibacillus	+	Monoderm	A4α			18	42
Lactobacillus	+	Monoderm	A4α	105	133	22	41
Lactococcus	+	Monoderm	A3α	388	376	121	144
Leuconostoc	+	Monoderm	A3α	35	45	11	16
Levilactobacillus	+	Monoderm	A4α			9	14
Limosilactobacillus	+	Monoderm	A4α			6	14
Listeria	+	Monoderm	A1γ	75	87	17	32
Mycobacterium	+	Diderm (mycolic acid)	A1γ	1,990	3,596	299	374
Propionibacterium	+	Monoderm	A3γ	202	214	5	8
Rhodococcus	+	Diderm (mycolic acid)	A1γ	70	83	31	70
Staphylococcus	+	Monoderm	A3α	349	456	79	155
Streptococcus	+	Monoderm	A3α	1,064	739	246	688
Streptomyces	+	Monoderm	A3γ	240	208	97	151
Total[e]				9,539	10,206	2,182	3,303
Gram-negative				4,071	3,113	770	844
Gram-positive				5,468	7,093	1,412	2,459

[a]Contains lanthionine instead of m-DAP.

[b]Although there seems to be some variety in peptidoglycan structure among *Arthrobacter*, we assigned the most common chemotype.

[c]There are several chemotypes within *Bifidobacterium*, but A4α was assigned to the only *Bifidobacterium* host entry in our database (*Bifidobacterium thermophilus*).

[d]Phage lysins with unassigned *Enterococcus* host species were included within *E. faecalis*.

[e]In the last three rows, data are the sum of the column values (phage genomes, genes detected, and genes after curation) or weighted averages (PF hits per protein).
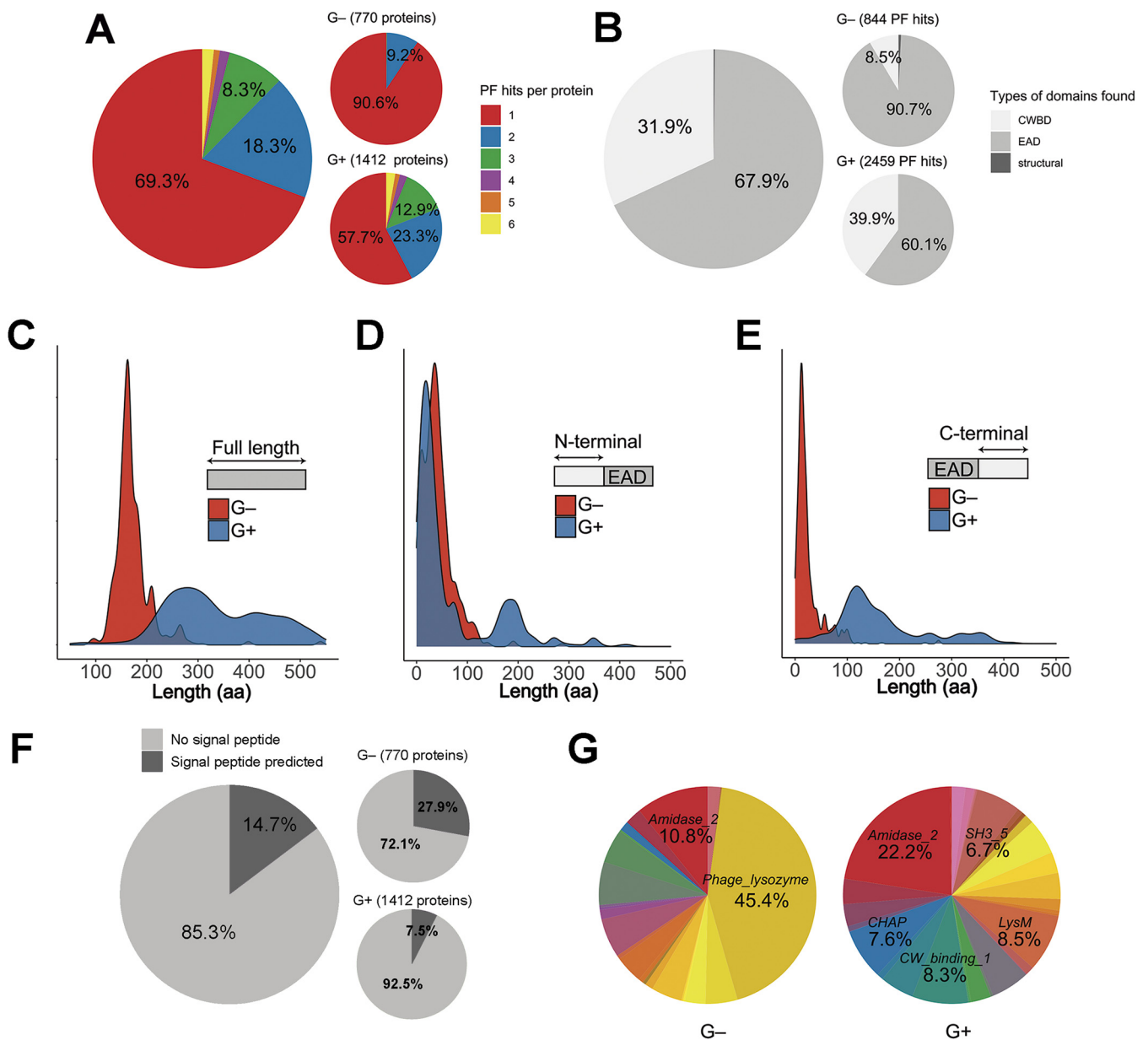
FIG 2 General properties of lysins from phages that infect G+ or G−. (A) Distribution of the number of PF hits predicted per protein. (B) Distribution of domain types. (C) Distribution of protein lengths. (D and E) Distributions of the number of amino acids before (D) or after (E) predicted EADs. (F) Percentage of lysins with a predicted N-terminal signal peptide according to Phobius. (G) PF domain variability (different colors stand for different PF domain families, corresponding to those in Table 2). In distribution charts (C, D, and E), the y axis shows an estimation of the distribution density.

after the predicted EADs (Fig. 2D and E). Figure 2D shows that EADs from G− phage lysins start at approximately the same point as G+ ones, that is, near the N-terminal end of the protein, except that the EAD starting point distribution is slightly shifted toward the C-terminal part of the enzyme in lysins from G−, perhaps due to the presence, in some cases, of CWBDs at the N terminus (28) or, most likely, because of a high representation of signal anchor release (SAR) lysins among G− phage lysins (Fig. 2F). The latter panel shows that a signal peptide was predicted by Phobius server for 27.9% (215 lysins) of the G− phage lysins, while such elements were predicted for only 7.5% (106 lysins) of the G+ ones. The prediction of a signal peptide is indicative of a SAR mechanism of lysin export into the periplasm (engaging the Sec secretion system) (12). Of note, the G+ EAD starting point distribution shows a secondary local maximum at around coordinate 200. This is consistent with the presence of EADs at a medial location within the protein,

something that has already been observed in many G+ phage lysins (13, 43). According to Fig. 2E, most G+ EAD hits have much more "space" at the C-terminal region than G− ones (the respective medians of C-terminal length after EAD hit distributions for G− and G+ are 16 and 136 aa residues). The additional length at the C-terminal part of G+ phage lysins can be occupied by CWBDs or additional EADs, and, taken together, all this evidence supports the common postulate that most detected G− lysins are monomodular.

Finally, Fig. 2G illustrates that, in contrast with the case of G− lysins, G+ lysins present a high diversity of different types of domains. There is a remarkable predominance of the EADs belonging to the *Phage_lysozyme* family of proteins in G− lysins (45.4% of total hits), whereas *Amidase_2*, the most frequent EAD among G+ phage lysins, accounted only for 22.2% of G+ PF hits.

Differential distribution of domain families among different bacterial host groups. A distribution analysis of each PF family among bacterial hosts was performed (Table 1). Of the total 3,303 PF hits analyzed, 2,460 corresponded to phages infecting G+ bacteria; 2,243 (1,477 G+ and 766 G−), 1,054 (982 G+ and 72 G−), and 6 (G−) corresponded to EADs, CWBDs, and structural domains, respectively (the sources for domain classification as EAD, CWBD, or structural can be found in Table S2 in the supplemental material). When the differential Gram group classification of each PF hit was analyzed, it was found that EADs like *Amidase_5*, *Glyco_hydro_25*, *Peptidase_C39_2*, and *Transglycosylase* were exclusive to G+, whereas *Glyco_hydro_108* and *Muramidase* were characteristic of phages infecting G−. Other EADs like *Amidase_2*, *Amidase_3*, *CHAP*, *Glucosaminidase*, *Peptidase_M15_4*, and *Peptidase_M23*, were common in G+, whereas *Glyco_hydro_19*, *Hydrolase_2*, and *Phage_lysozyme* dominated among G−. In addition, *CW_7*, *CW_binding_1*, *LGFP*, *SH3_5*, and *ZoocinA_TRD* constituted the CWBDs of G+, and although *LysM* and *PG_binding_1* were most frequently found in G+ lysins, they also appeared sometimes among G− lysins (Table 2 and Fig. 3). *PG_binding_3* was the only CWBD exclusive to G− lysins. Interestingly, all of the 40 *PG_binding_3* occurrences were accompanied by *Glyco_hydro_108* at the N-terminal moiety, yielding an architecture ([*Glyco_hydro_108*] [*PG_binding_3*]) that was widespread among gammaproteobacteria.

Trends in distribution of PF domains among genera, rather than Gram group, were a bit more complex (Fig. 4 and 5), although some conclusions could be reached. To begin with G+ CWBDs, the *CW_binding_1* repeats were only encoded by phages infecting streptococci, whereas *CW_7* constitutes the CWBD of many phage lysins of *Streptococcus*, *Arthrobacter*, and *Streptomyces*. *CW_binding_1* repeats are known to bind choline residues present in the teichoic acids of *Streptococcus pneumoniae* and its relatives (i.e., streptococci of the mitis group) (47, 48) and therefore appeared only within our data set among this group of bacterial hosts (Fig. 6). *CW_7* repeats are known to bind a conserved peptidoglycan motif and are thus less restricted in the variety of bacteria they may recognize (49). *LysM* domains were also widely distributed in G+, *ZoocinA_TRD* was very common among *Streptococcus thermophilus*, and *PSA_CBD* was exclusive for *Listeria* phage lysins. As for EADs, *Amidase_5* was very frequently found among streptococci and *Amidase_2* was generally abundant among all G+.

Another exclusive trait of some G+ lysins was the concurrence of two distinct EADs. This was observed for phage lysins from *Streptococcus suis*, pyogenic-group streptococci, staphylococci, and mycobacteria. A possible explanation for multicatalytic lysins is an increased lytic efficiency over monocatalytic ones, since activities attacking different sites of the peptidoglycan are known to act synergistically in peptidoglycan degradation (50). Such synergy could also imply a decreased chance for the appearance of resistant peptidoglycan mutants (51). It has also been shown that the synergistic concurrence of both activities is sometimes needed for full activity. Thus, it has been suggested that some phages may have evolved a regulatory mechanism to avoid the lysis of other potential host cells relying on the proteolysis of bicatalytic lysins by host cell proteases. Then, both EADs would be disjointed by proteolysis upon host cell lysis, and the degraded lysins would no longer be active against the nearby bacterial population (52). This should be especially relevant for phages infecting G+, which lack a protective

TABLE 2 Distribution of PF hits of phage lysins from Gram-positive and Gram-negative bacteria[a]

Domain	Domain type	G+ (%)	G− (%)	Total	Encoded by phages of:
3D	EAD	7		7	
Amidase_2	EAD	547 (85.7)	91 (14.3)	638	Widely distributed in G+
Amidase_3	EAD	93 (81.6)	21 (1.4)	114	*Bacillus, Streptococcus, Clostridium, Staphylococcus*
Amidase_5	EAD	81 (100)		81	*Streptococcus, Lactococcus*
CHAP	EAD	186 (93.9)	12(6.1)	198	*Streptococcus, Staphylococcus, Streptomyces, Arthrobacter*
Cutinase	EAD	24		24	
FSH1	EAD	3		3	
Glucosaminidase	EAD	73 (97.3)	2 (2.7)	75	*Streptococcus*
Glyco_hydro_19	EAD	9 (14.3)	54 (85.7)	63	*Acinetobacter* and other genera
Glyco_hydro_25	EAD	142 (100)		142	*Lactobacillus, Bacillus, Streptococcus*
Glyco_hydro_108	EAD		43 (100)	43	Widely distributed in G−
GPW_gp25	EAD		12	12	
Hydrolase_2	EAD	3 (5.8)	49 (94.2)	52	*Escherichia, Pseudomonas, Vibrio*
Muramidase	EAD		35 (100)	35	*Pseudomonas, Burkholderia*
NLPC_P60	EAD	7	6	13	
PE-PPE	EAD	13		13	
Peptidase_C39_2	EAD	42 (100)		42	*Mycobacterium*
Peptidase_C93	EAD		4	4	
Peptidase_M15_3	EAD		10	10	
Peptidase_M15_4	EAD	96 (71.1)	39 (28.9)	135	*Mycobacterium, Bacillus*
Peptidase_M23	EAD	77 (97.5)	2 (2.5)	79	*Arthrobacter, Mycobacterium, Rhodococcus*
Pesticin	EAD		2	2	
Phage_lysozyme	EAD	32 (8.0)	366 (92.0)	398	Widely distributed in G−
Phage_lysozyme2	EAD	1		1	
Prok-JAB	EAD		1	1	
Prophage_tail	EAD	3		3	
SLT	EAD	6	16	22	
Transglycosylase	EAD	32 (100)		32	*Mycobacterium*
Amidase02_C	CWBD	20		20	
Big_2	CWBD	1		1	
CW_7	CWBD	125 (100)		125	*Streptococcus, Arthrobacter, Streptomyces*
CW_binding_1	CWBD (repeat)	205 (100)		205	*Streptococcus*
CW_binding_2	CWBD (repeat)	3		3	
DUF3597	CWBD	5		5	
LGFP	CWBD (repeat)	31 (100)		31	*Rhodococcus*
LysM	CWBD	210 (97.7)	5 (2.3)	215	Widely distributed in G+
PG_binding_1	CWBD	130 (83.3)	26 (16.7)	156	*Mycobacterium, Bacillus, Streptomyces*
PG_binding_3	CWBD		40 (100)	40	Widely distributed in G−
PSA_CBD	CWBD	13		13	
SH3_3	CWBD	20	1	21	
SH3_5	CWBD	164 (100)		164	*Streptococcus, Staphylococcus, Lactobacillus, Bacillus*
SPOR	CWBD	5		5	
ZoocinA_TRD	CWBD	50 (100)		50	*Streptococcus, Enterococcus*
Gp5_C	Structural		3	3	
Gp5_OB	Structural		3	3	

[a]Percentages and further remarks are shown only for domains represented by at least 30 hits.

OM hindering the lysis of other bacterial cells from without, and this accounts for the exclusiveness of the bicatalytic architecture among phages infecting G+. In some other cases, however, it is the high affinity of the CWBD that has been proposed as the mechanism that maintains lysins tightly bound to cell debris, preventing widespread lysis of the bacterial community (53), which is also an argument for the widespread presence of CWBDs among G+ and not among G−. Additionally, it has also been
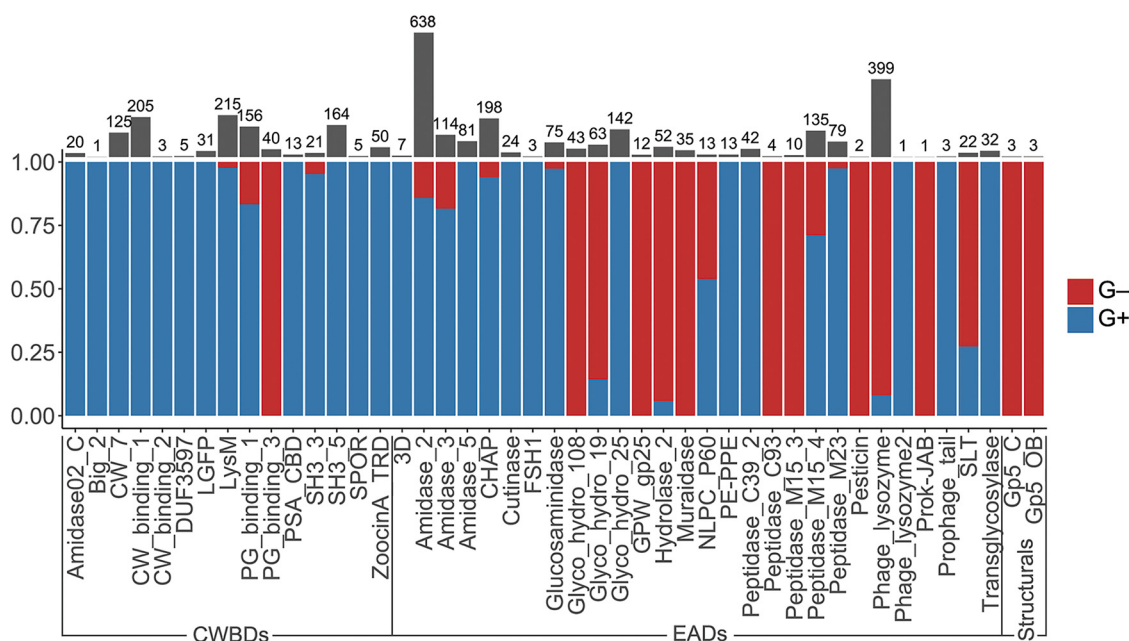
FIG 3 Differential distribution of PF hits among G− and G+ bacterial hosts. The *y* axis shows the proportion of PF hits found in G+ within a given domain family. Gray bars and numbers above the graph represent the total number of hits of each PF domain.

proposed that, at least in some cases, the central domain predicted to be an EAD may in fact have an auxiliary role in substrate binding, rather than being an actual EAD (54).

Staphylococcal phage lysins presented reduced EAD variability, normally using *Amidase_2*, *Amidase_3*, and/or *CHAP* domains, with *SH3B_5* being the preferred CWBD, in agreement with previous results (55). In some cases, the staphylococcal *SH3B_5* has been shown to bind the peptidoglycan with the characteristic pentaglycine interpeptidic bridge of *Staphylococcus* (56). Domains putatively assigned an esterase activity (*Cutinase*, *FSH1*, and *PE-PPE*) were present only in phages from *Mycobacterium* and its relatives, presumably as type B lysins. The *LGFP* repeats, quite common among *Rhodococcus* phages, might be a specific CWBD among such *Corynebacteriales*. Peptidase EADs were common and diverse among mycobacteriophages, in contrast with other G+ phages, which do not typically contain peptidase EADs other than *CHAP*. Of note, *CHAP* domains have been described sometimes as peptidases but on other occasions as *N*-acetylmuramoyl-ʟ-alanine amidases (NAM-amidases) (57, 58).

Regarding G−, the most widely spread architecture of G− phage lysins was monomodular, harboring a single *Phage_lysozyme* domain, which accounted for half (50.8%) of the identified G− lysins in our database. Another architecture that was only found in G− lysins is the localization of a CWBD at the N-terminal end (for example, [*PG_binding_1*] [*Muramidase*]), although it was not at this position in every case (e.g., the architecture [*Glyco_hydro_108*] [*PG_binding_3*] was also present).

The correlation between domain distribution and peptidoglycan composition might also shed some light on the relationships of different domain families with different taxa. To that end, the chemotypes classification of peptidoglycan proposed by Schleifer and Kandler (59) was used (Table 1). Briefly, such classification hierarchically relies on (i) the site of cross-linkage of the peptide subunit of the peptidoglycan, (ii) the nature of the cross-link, and (iii) the specific residue at position 3 within such peptide subunit (Fig. 7A). Starting with CWBDs (Fig. 7B), classification by chemotype did not provide a better explanation for specificity than other genus-specific traits, as discussed above. However, some specificities could be found (e.g., *Amidase02_C* appears only in phages that infect A1α bacteria or *PG_binding_3* only in A1γ), and some CWBDs that are widespread among different chemotypes could also be observed (*PG_binding_1*, *LysM*, and *SH3_5*). In general, however, it cannot be stated that peptidoglycan composition is a major

FIG 4 Heat map of PF hits distribution across host bacterium genera. Numbers within each tile indicate the number of hits predicted for the corresponding taxon and PF family. The color scale represents the number of hits from low (red) to high (yellow). Gray bars on the right represent the total number of PF hits predicted within each genus.

determinant for CWBD specificity, except for some cases, such as *ZoocinA_TRD* domains, which have been proposed to bind A3α with two Ala residues at the cross-link (60). The poor performance of chemotype as an *a priori* predictor of the CWBD PF family ligand is more evident if we consider the CWBD types which appeared widespread among many different chemotypes, such as *LysM* and *SH3_5*. To check whether this apparent "promiscuity" may be linked to the presence of subfamilies with potentially different ligands or if it could rather be a true promiscuous binding, sequence similarity networks (SSNs) were constructed with the PF hits of *LysM* and *SH3_5* (Fig. 8). The *LysM* SSNs did not show prominent similarity clusters classified either by taxon or by chemotype of the bacterial host. This suggests that *LysM* could be a truly universal CWBD that would bind to a conserved cell wall ligand. The rather generic description of *LysM* ligands in the literature (as *N*-acetylglucosamine-containing polysaccharides) is in agreement with this observation. *SH3_5*, however, displayed at least two differentiated sequence similarity groups that correlated rather well with different taxonomic groups (namely, staphylococci versus streptococci and lactobacilli). In fact, the literature shows that while lytic enzymes with predicted *SH3_5* domains typically recognize polysaccharides (and peptidoglycan in particular), there seem to be different specializations. For example, the CWBD of the *Lactiplantibacillus plantarum* major autolysin binds many different peptidoglycans with low affinity, with glucosamine being the minimal binding motif (61), while *SH3_5*

FIG 5 Relevant architectures observed in lysins from phages infecting different taxonomic groups of bacteria. Different colors indicate different domains; brackets denote domains that appear in only some representatives of the depicted architecture.
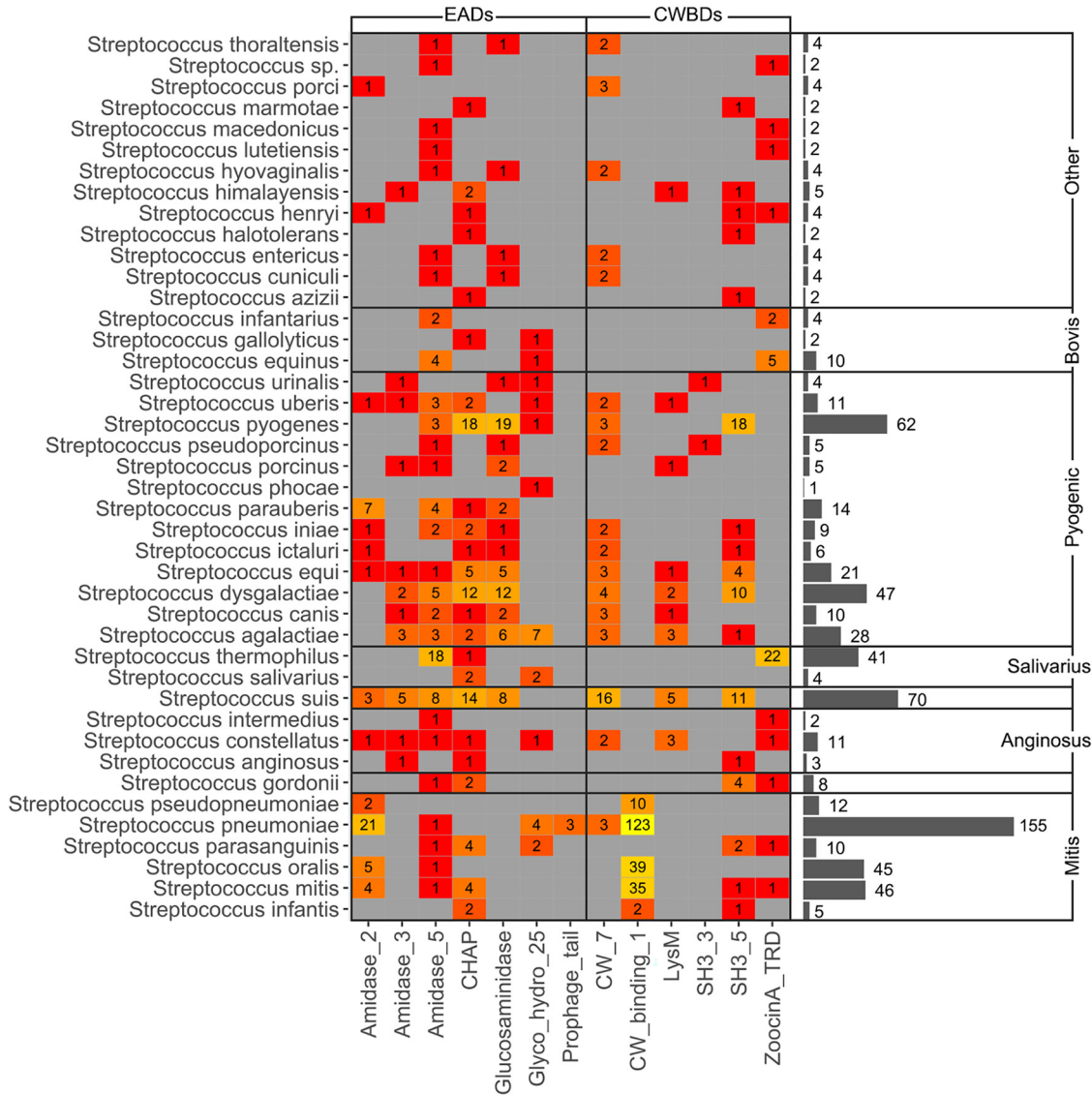
FIG 6 Heat map depicting PF hits distribution among different streptococci. Numbers within each tile indicate the number of hits for the corresponding taxon, whereas colors express a scale from lower (red) to higher (yellow) number of hits. Gray bars show the total number of PF hits for each streptococcal species.

domains from staphylolytic enzymes have been shown to be rather specific to cross-linked peptidoglycans (like the A3α peptidoglycan of *Staphylococcus* and *Streptococcus*) and that the nature of the cross-link itself determines the affinity of such CWBDs for the peptidoglycan (56, 62).

Additional information could be drawn from this analysis when it was applied to the different catalytic activities detected (Fig. 7C). First of all, NAM-amidases were the most represented type of domains and also those that appeared among more different taxonomic groups and chemotypes, even more so than lysozymes. Indeed, *Amidase_2*, the most abundant PF domain in our data set (638 hits), appeared in lysins from both G+ and G− phages. The SSN in Fig. 8 shows, however, that although *Amidase_2* seems a rather diverse group, with various observable similarity clusters, none of these clusters correlates with any of the classifiers of the bacterial hosts tested.

Muramidases were quite overrepresented among G− phages (chemotype A1γ) because of the widespread presence of *Phage_lysozyme* domains. Glucosaminidases appeared evenly both against A1 and A3 peptidoglycans, but whereas in G+ phages (which comprise all A3s and a few A1s) glucosaminidase activity was represented by
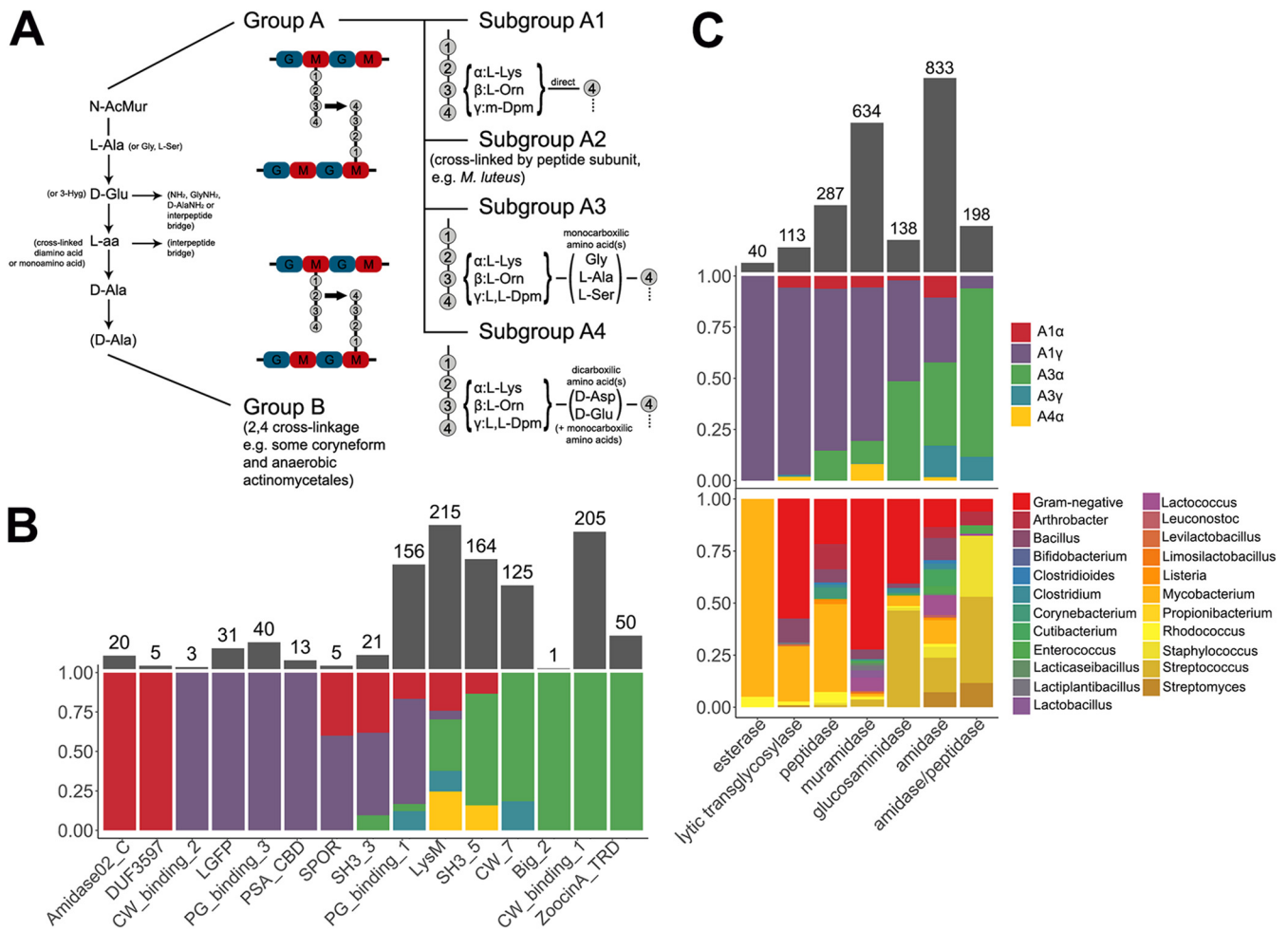
FIG 7 Differential distribution of CWBDs and catalytic activities across peptidoglycan chemotypes and taxonomic groups of bacterial hosts. (A) Schematic representation of the relevant peptidoglycan chemotypes present for the bacterial hosts in our data set. (B) Distribution of CWBD PF hits among chemotypes. (C) Distribution of catalytic activities of EAD PF hits among chemotypes and taxonomic groups.

Glucosaminidase PF domain, the only domain putatively assigned with a glucosaminidase activity among G− phages was *Glyco_hydro_19* (Fig. 4 and 7).

Another interesting observation is that peptidase activities were more common among lysins from phages infecting bacteria with subgroup A1 peptidoglycans, which, in turn, display the simplest cross-linkage of all types, lacking an interpeptide bridge. Thus, peptidases were not uncommon among G− phages and were also present in phages from A1 chemotype G+ (especially mycobacteriophages but also listeriophages and phages from *Clostridium*, *Bacillus*, or *Corynebacterium*). On the other hand, amidase/peptidases, which is the label given to *CHAP* domains (Table S2 in the supplemental material), were much more prevalent among A3 G+ and only seldom present in lysins from phages infecting A1 bacteria (namely, some G−). This suggests that if there is an A3-specific peptidase activity, it would be the one located in *CHAP* domains. It makes sense that different peptidase structures have evolved toward A1 and A3 peptidoglycans, since the complexity of their peptidoglycan peptide moieties differs significantly. Adding to this conclusion, the *CHAP* SSN (Fig. 8) did show a similarity clustering of the few *CHAP* examples in lysins from A1 phages, besides an apparent differentiation of *Staphylococcus* and *Streptococcus/Enterococcus*.

Physicochemical analysis of phage lysins from Gram-positive versus those from Gram-negative bacteria. The results analyzed so far support a distinct distribution of domain architectures and families among lysins that infect different kinds of bacteria and even hint at an association of a differential distribution with some cell
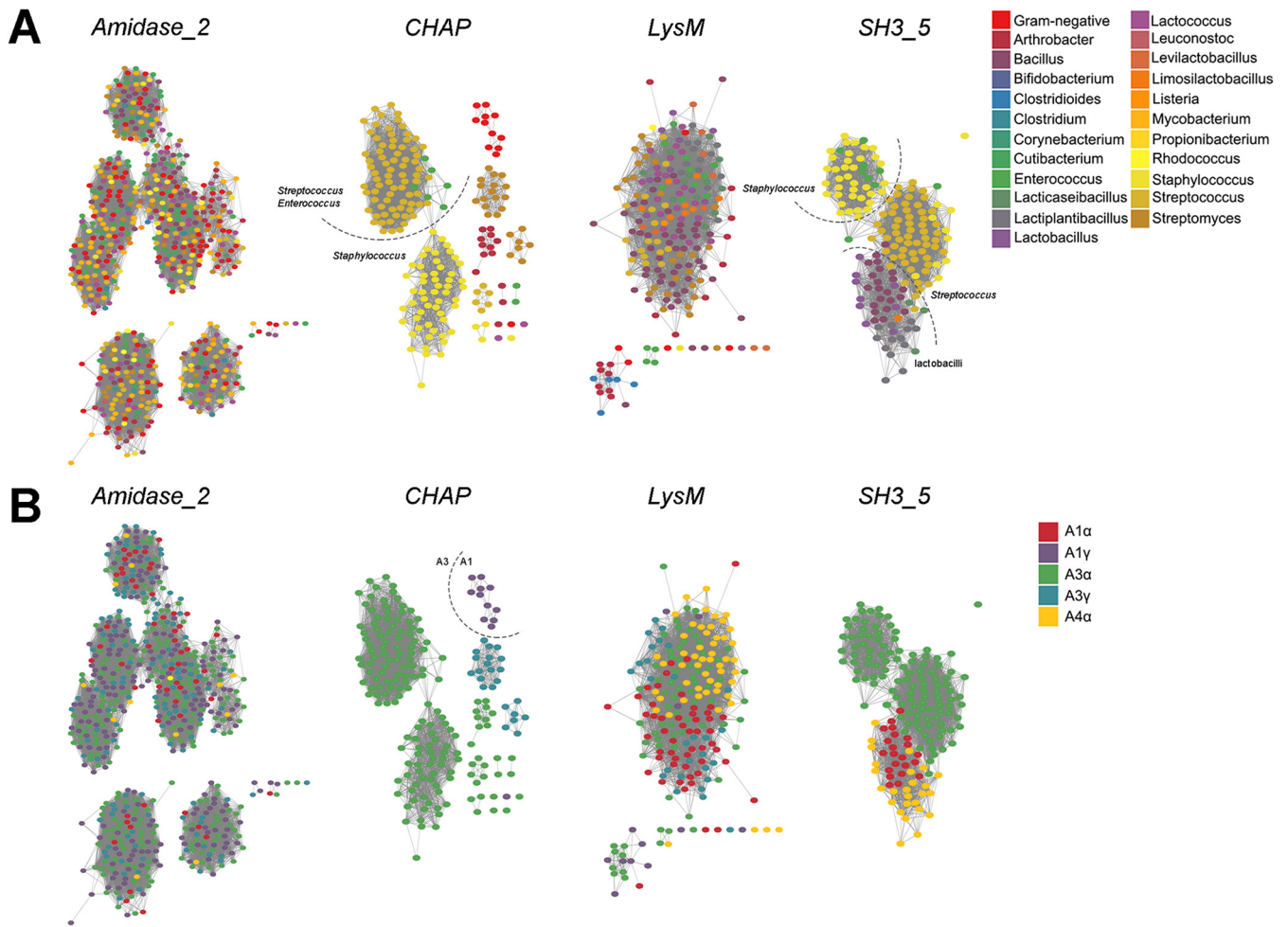
FIG 8 SSNs of the PF hits in our data set corresponding to different domain families. (A) Classification of sequences according to the taxonomic group of the corresponding bacterial host. (B) Classification by peptidoglycan chemotype of the host. Each node represents a single sequence. Dashed lines separate recognizable similarity clusters. The edge similarity cutoff was ≈40% for *CHAP*, *LysM*, and *SH3_5* and ≈30% for *Amidase_2*.

wall properties. To check whether such variations can also be correlated with a measurable difference in physicochemical properties, net charge, net charge per residue (NCPR), hydrophobicity, average hydrophobic moment, and aliphatic index were calculated and used to implement a random forest (Fig. 9). This way, the aforementioned physicochemical variables were used as classifiers for the prediction of the host bacterium Gram group of lysins. The resulting algorithm yielded a receiver operating characteristic (ROC) plot with an area under the curve (AUC) of 0.897, which can be interpreted as a good predictive ability (Fig. 9A). Using the probability threshold (0.591) derived from the best point of the ROC curve (which maximizes true-positive rate and minimizes false-positive rate), G+/G− classification upon the testing subset (Fig. 9B) yielded an accuracy of 87.9% with sensitivity and specificity of 84.1% and 81.3%, respectively (with the classification as G+ being the "positive" one). According to the subsequent analysis (Fig. 9C), NCPR was the most relevant variable to distinguish between G+ and G−, followed by average hydrophobic moment, aliphatic index, and, finally, hydrophobicity. In general, these results suggest that lysins from phages that infect G+ and G− can in fact be differentiated by their physicochemical properties in a relatively efficient manner. Nonetheless, a second random forest model constructed with additional variables (namely, the protein length, number of PF hits, and the prediction of N-terminal signal peptides) rendered a better result in classifying lysins (96.15% accuracy) (Fig. 9D to F). In the latter case, the variable with a greater impact in classification was the protein length (Fig. 9F), which is not surprising given the data
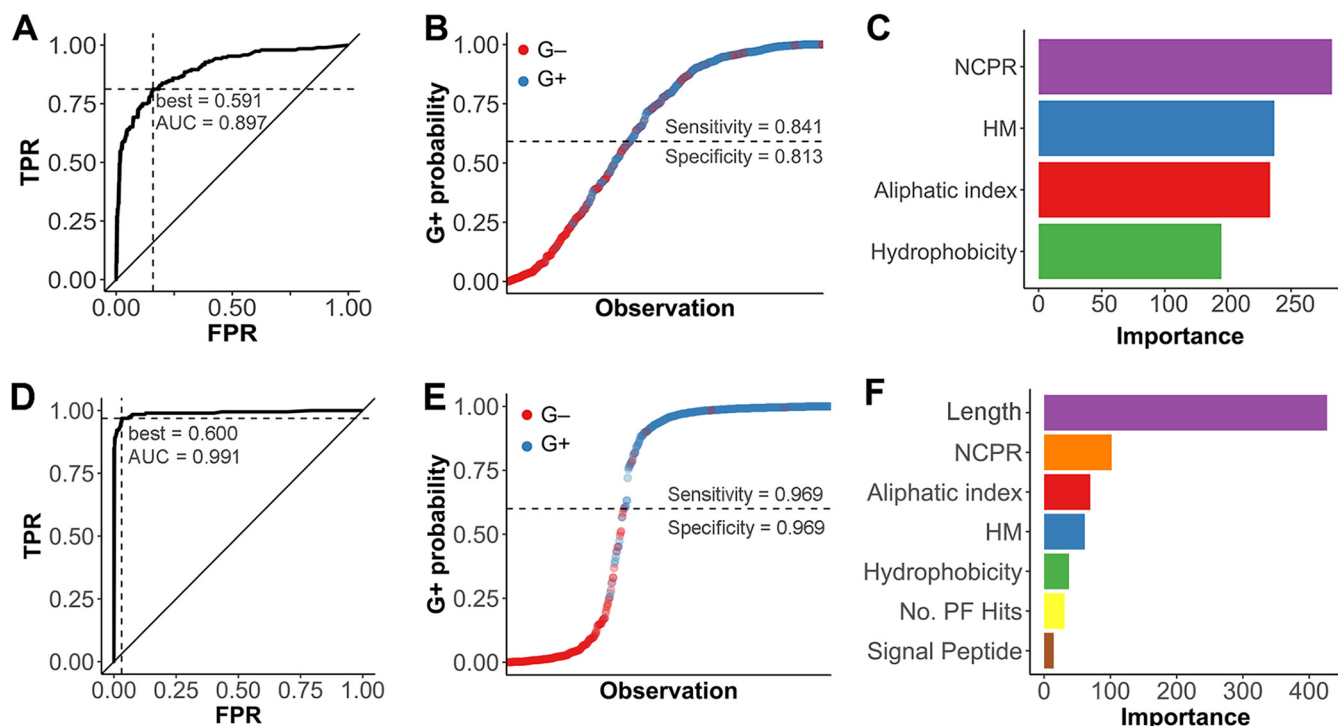
FIG 9 Random forest prediction and classification of Gram groups of bacterial hosts based on physicochemical properties of lysins (A, B, and C) or on those properties plus others relative to lysin architecture (D, E, and F). (A and D) ROC curves of the random forest predictive models (TPR, true-positive rate; FPR, false-positive rate). ROC best points of positive-group (G+) probability for outcome maximization are presented, as well as the AUCs. (B and E) Random forest castings of bacterial host Gram group on the testing subset of lysin sequences. The dashed lines represent the G+ probability threshold for classification based on the respective ROC best points. (C and F) Importance (i.e., mean Gini index decrease for each variable) of each of the four descriptors used for classification within each model. HM, hydrophobic moment.

presented in Fig. 2C, followed by the physicochemical properties. These results prove that it is possible to *a priori* predict the Gram group of the bacterial host of a given lysin with rather high accuracy and also that the physicochemical features of the lysins do play a role in the differentiation of such classes.

Indeed, the net charge distribution (normalized by protein length) was significantly higher in G− lysins than in G+ ones ($P \leq 0.0001$; effect size [ES] = 0.66) (Fig. 10A, leftmost panel). Moreover, the average prediction of local net charge suggested that this difference is located mainly at the C-terminal part of G− lysins (Fig. 10B). A more thorough comparison (Fig. 10C) seemed to confirm this. At every sequence quartile of the proteins (i.e., contiguous fragments of sequence with a length equal to 1/4 of the total number of amino acid residues in the original protein sequence), the net charge distribution of G− lysins showed a significantly higher net charge. However, the actual size of this shift was only moderate along the sequences (ES between 0.24 and 0.34), but it was, again, higher at the final quartile (ES = 0.52).

Hydrophobicity was also higher in G− lysins, but the difference relative to that of G+ ones is smaller (ES = 0.36). This might be related to the rather inconsistent pattern shown by average local hydrophobicity and sequence quartile comparison (Fig. 10B and C). G− lysins tended to have a more hydrophobic N-terminal region, which could be correlated with the greater abundance of predicted signal peptides among the G− group mentioned above (Fig. 2F). At the C-terminal moiety, the tendency was reversed, which can be explained by the relative abundance of positively charged residues shown before for G− phages. It is at the third quartile (Q3), immediately before the high positive net charge patch described above, where the difference was statistically more relevant ($P \leq 0.001$; ES = 0.59), with higher values in G− phages. There was also a statistically significant difference in the average hydrophobic moment distributions between G+ and G− phage lysins. For the G− group, the local plot (Fig. 10B) showed a higher tendency
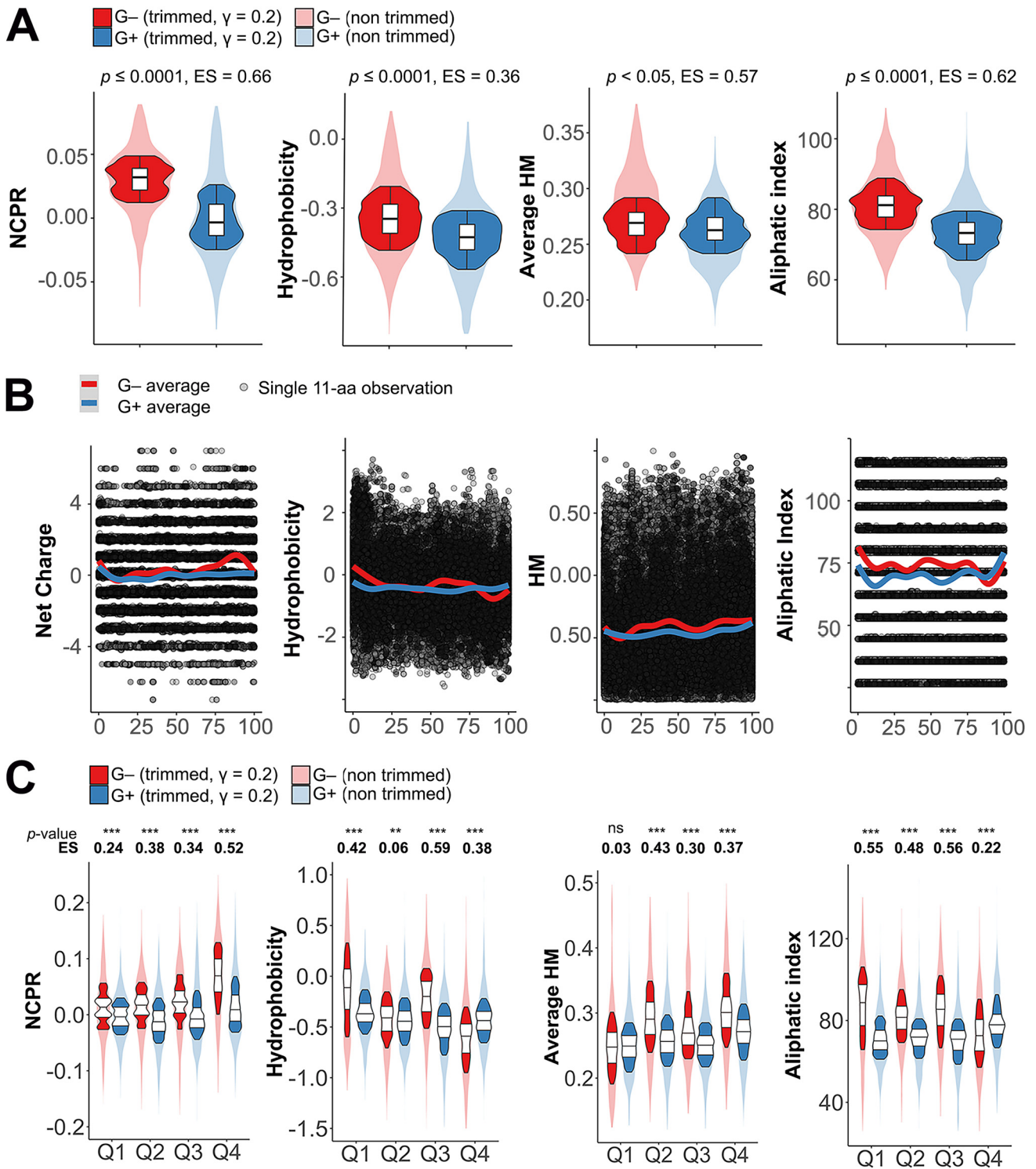
FIG 10 Differential physicochemical properties distribution among G+ and G− phage lysins. (A) Distribution of net properties calculated along the whole protein sequences of lysins from phages infecting G− or G+. (B) Local computation of physicochemical properties. Each dot represents the particular value calculated for an 11-aa window in a given lysin. Continuous lines are average tendencies based on either all G− or all G+ data points. (C) Distribution of different properties at quartiles of lysin sequences. Asterisks indicate P values (**, $P \leq 0.01$; ***, $P \leq 0.001$) obtained from the Yuen-Welch test for trimmed means with a trimming level (γ) of 0.2. ES indicates the Wilcox and Tian ζ measurement of effect size.

to present greater hydrophobic moments along the whole protein length except for the N-terminal part. Analysis of sequence quartiles confirmed a statistically significant superiority of average hydrophobic moment for G− phage lysins except at the N terminus. The aliphatic index was also significantly higher in G− phage lysins, although G+ phage
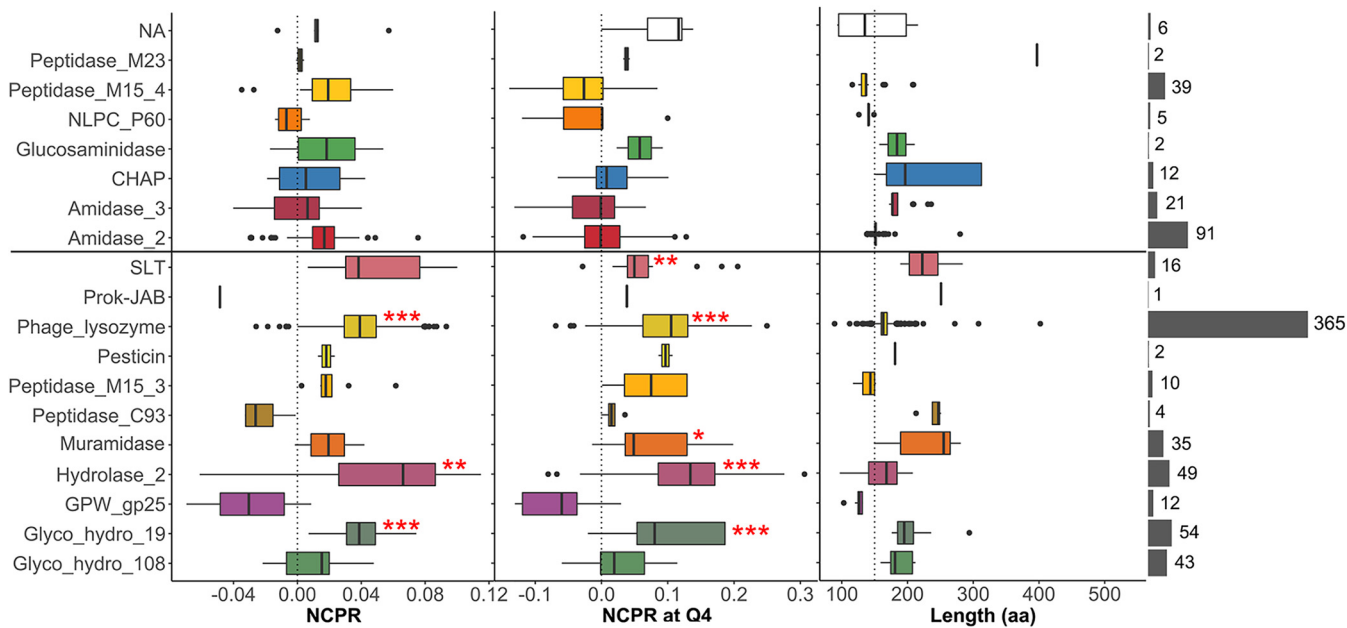
FIG 11 Net charge distribution of lysins from G−-infecting phages classified according to the predicted EAD. The rightmost gray bars depict the number of lysins classified into each EAD group (lysins within the NA group are those for which an EAD was not assigned). All groups were compared with the distribution of the *Amidase_2* domain, as a highly represented, near-neutral control using Welch's test on $\gamma = 0.2$ trimmed means with *post hoc* Bonferroni correction (*, $P \leq 0.05$; **, $P \leq 0.01$; ***, $P \leq 0.001$).

lysins showed an aliphatic index peak at their C-terminal regions that surpassed that of G− phage lysins (coincidental with the G− basic amino acid peak, which, understandably, would lower both hydrophobicity and aliphatic index at Q4) (Fig. 10C).

Taking all these observations together with the results yielded by the random forest prediction, we can conclude that the physicochemical difference between lysins from phages that infect G+ or G− is manifested as a higher positive net charge of G− phage lysins, particularly at the C-terminal end, combined with a greater propensity for incorporating aliphatic amino acids, and likely results in amphiphilic structures.

A closer examination of net charge (and C-terminal net charge) of lysins from G−-infecting phages indicated that the high positive patch trait seems specific to some domain families. As a whole, a statistically significant higher NCPR value was found in lysins bearing *Phage_lysozyme*, *Hydrolase_2*, and *Glyco_hydro_19* domain families (Fig. 11). At the C-terminal part, higher NCPR was found in lysins bearing the domains mentioned above but also in *SLT* and *Muramidase*. The average local net charge tendency showed for each EAD group (Fig. 12) confirmed that a local high-positive-charge peak appears in the protein region immediately before the C-terminal apex.

Interestingly, all of the aforementioned domains that present a high-positive-charge patch at their C-terminal region were preferentially present in lysins from phages that infect G− (Table 2). This observation provides a basis to argue an evolutionary tendency in some G− infecting phages to develop AMP-like subdomains at the C-terminal moiety of their lysins. Indeed, such subdomains contain features typical of AMPs (such as the high net charge accompanied by a high local hydrophobic moment, hydrophobic patches, etc.) and may play a role in the interaction between lysins and cell wall in G− bacteria. Electrostatic interactions may play a significant role in phage-bacterium interplay, as suggested for modular lysins from phages that infect G+ bacteria. For example, it has been shown that the negative net charge of many G+ lysins hinders their ability to approach the negatively charged cell wall from without (18, 63). This renders the affinity-based interaction of the CWBDs with their cell wall ligands essential for lysin activity, at least when they are exogenously added. The essentiality of CWBDs has been shown for several lysins (20) and, as is the case for other carbohydrate-active
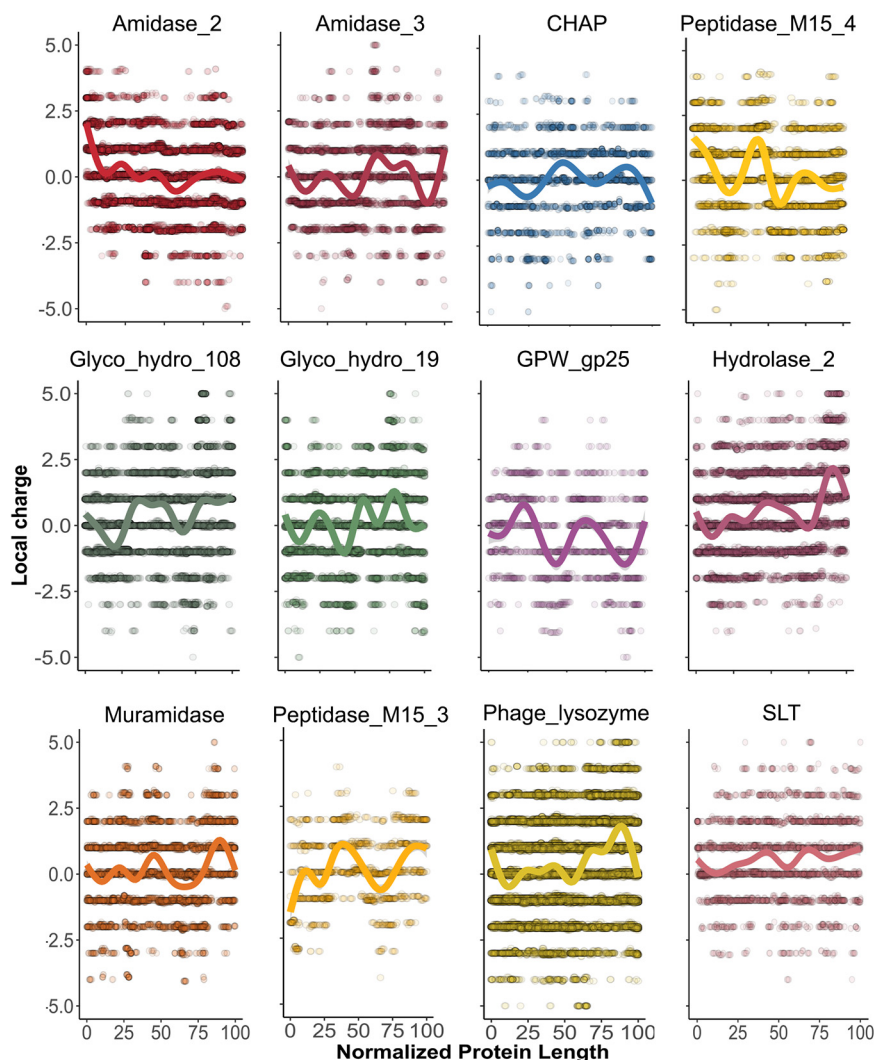
FIG 12 Local computation of physicochemical properties in lysins from G− infecting phages classified according to EAD predictions. Each dot represents the particular value predicted for an 11-aa window from a given lysin. Continuous lines are average tendency lines.

enzymes, can be related to the insoluble nature of the substrate. Therefore, the CWBD serves the purpose of providing affinity for the nondiffusible substrate (in this case, the peptidoglycan). Once again, though, generalizations should be made with caution, because there are also cases reported of single catalytic domains that lysed G+ cells from without more efficiently when their CWBD was removed (64). On the other hand, and to our knowledge, there are only a few cases reported in which CWBDs appear to increase the efficiency of cell wall-lysin interaction in G− lysins (21).

We have already shown, based on our own data, that it is safe to say that G− lysins are monomodular. Thus, taking this theoretical framework into account, it could be argued that some G− lysins could have evolved a distinct strategy to grant cell envelope interaction, namely, the AMP-like regions, rather than containing an additional CWBD, which, incidentally, might be essential for postlytic regulation in G+ but not in G− (53). The AMP-like subdomains, besides providing better anchorage to bacterial surface structures, might also act as an additional mechanism for effective lysis of G− bacteria (i.e., a spanin-like function), and it is even conceivable that the AMP-like elements could take over a holin function, since there are examples of lysins able to gain access to the peptidoglycan on its own, without the necessary co-occurrence of a holin (32, 65). There are indeed abundant examples in

literature of the ability of G− lysins to interact with bacterial membranes and permeabilize them (33, 38, 39, 66), a trait that, it is plausible to say from both our own analysis and the experimental results of many works, would reside in such AMP-like elements.

It was recently suggested that phages lacking spanins (spaninless phages) need additional gene products that behave similarly to AMPs to induce efficient host cell lysis (67). This hypothesis could also be extrapolated to the presence of AMP-like membrane-permeabilizing subdomains, which could then compensate for a missing spanin system. However, the lack of extensive evidence on the latter issue warns against this kind of generalizations for the time being. If we assume that AMP-like subdomains behave as membrane-active molecules, the identification of these elements within lysins could also provide a way to predict the ability of such lysins to better interact with the OM from without, and thus their antimicrobial potential. This kind of approach has already been successfully tested at least once, as demonstrated, for example, in the work of Larpin et al. (68), in which AMP-like subdomain bearing lysins were screened by means of a homology search based on a previously described AMP (39).

Concluding remarks. Phages and their bacterial hosts are constantly evolving in a codependent manner (69). From the point of view of phage lysins, this means that such molecules have adapted to the particular structures and features of the host cells. This adaptation can be described as the functional adjustment of the protein elements to optimally fulfil their purposes: the efficient and regulated degradation of the peptidoglycan. Therefore, lysin structures and cell wall structures must be closely correlated. A way of testing and understanding this relationship is the sequence-based classification of the constituent domains of phage lysins presented here and the analysis of their distribution among (pseudo)taxonomical and structural classes of bacterial hosts. Our procedure yielded several important associations of lysins and cell wall architectures, which can be explained in a structural-functional way.

(i) The first lies in the architectural differences between lysins from phages that infect G+ and G−. Those from G− phages are usually monomodular, whereas lysins from G+-infecting phages are multimodular. Moreover, the bicatalytic type of modular structure appears only among G+. Some possible explanations for this architecture are the requirement for tighter postlytic regulation in G+ and/or a more efficient lytic activity relying on tighter substrate binding or on the synergistic effect of combining different catalytic activities.

(ii) Second is the association of CWBDs with specific bacterial host genera in our data set, together with the literature showing that many of these CWBDs are able to recognize ligands that are specific traits of the related bacterial hosts, for example, *SH3_5* in staphylococcal phages, *CW_binding_1* in *Streptococcus mitis* group phages, *PSA_CBD* in listeriophages, and *PG_binding_3* in G−. This also manifests the genetic trading between host and parasite, since many of those CWBDs, as well as their bacterial ligands, are also used by the bacterial host surface proteins.

(iii) The third feature is the differential appearance of EAD families within phages that infect bacteria with a certain chemotype, which suggests an adaptation of the enzyme to the structure of the specific peptidoglycan it has to degrade. This is notable in the case of peptidases. The somewhat wide range of peptidases identified within our data set is mainly distributed among phages infecting bacteria with subtype A1 peptidoglycan. In phages that infect subtype A3 bacteria, the most common EAD is *CHAP*, which has been shown to function either as a NAM-amidase or as an endopeptidase and, in any case, seems to be specific for A3 peptidoglycan.

(iv) Fourth is the remarkably differential distribution of domain families among phages that infect either G+ or G−, together with the association of such domains with different physicochemical properties.

(v) The last point concerns the differential physicochemical properties between lysins from G+ and G−, which, conversely, allow prediction of the Gram group of the bacterial host of a given lysin based on its sequence. In this work, the trait of a positively charged patch at a C-terminal position was found to be widespread among at least a subpopulation of lysins from G−-infecting phages. This trait has been previously correlated with an

improved ability to interact with the G− negatively charged OM. The higher values of other physicochemical variables in G− (aliphatic index and hydrophobic moment) also suggest an analogy of certain structural segments of G− lysins with AMPs.

These observations have clear implications for the design and development of lysin-based antimicrobials, from rational search (or design) of novel lysin parts to deriving AMPs from lysins sequences. A possible setup in which specific bacterial infections are tackled in an individualized manner based on a knowledge-driven, highly efficient synthetic biology platform for lysin-based antimicrobial production in the near future can be envisioned. The conclusions of this work can contribute to the consolidation of such a framework, together with the cutting-edge research currently being carried out in the field.

MATERIALS AND METHODS

Sequence database construction and curation. Phage genomes were retrieved from the NCBI nucleotide database by searching complete phage genomes limited to several bacterial taxa of interest, mainly selected by clinical or epidemiological importance and availability. Those genomes were screened for gene products whose annotations could indicate that they are lytic enzymes. Therefore, keywords such as "lysin," "lysozyme," "murein," "amidase," "cell wall hydrolase," "peptidase," and "peptidoglycan" were used as inclusion criteria, while "structural," "tail," "holin," "baseplate," and "virion protein" were used as exclusion terms to avoid misidentifications. Associated information, such as the taxon of the bacterial host, amino acid sequence, annotations, phage denomination, and protein/genome unique identifiers, was also added to the database.

Curation included (i) a sequence length cutoff, established with a minimum of 50 and a maximum of 550 aa residues; (ii) a sequence identity cutoff using CD-HIT (70) with default parameters and a 98% identity cutoff value to avoid redundant entries; (iii) examination with PfamScan (expectation value cutoff = 10) (71, 72) to rule out sequences where no relevant significant hits were found (i.e., where no functional domains that would plausibly appear within phage lysins were detected); and (iv) bacterial host genus assignation to each entry based on literature and genome annotations. The Phobius server was used to add information on the presence of predicted N-terminal signal peptides (73). The complete lysin collection and PF hits are presented in Table S1 in the supplemental material and at Digital.CSIC (74).

Physicochemical property prediction and analysis. Prediction of physicochemical properties (net charge, aliphatic index, hydrophobicity, hydrophobic moment) based on the amino acid sequences retrieved were performed using the R package "Peptides" implementation (75). Dawson's pK$_a$ scale was used for prediction of net charge assuming a pH of 7.0 (76); the hydrophobicity scale was that proposed by Kyte and Doolittle (77), and hydrophobic moment was calculated as previously proposed (78) with a specified rotational angle of 100° (recommended angle for α-helix structures). An average value of the hydrophobic moment of each of the possible 11-aa helices within a given sequence is given whenever noted. Such properties were predicted in the whole sequences, in sequences quartiles (contiguous fragments of sequences that account in length each for a quarter of the whole sequence), or in peptides of 11 aa to provide either a global vision or more local information.

A random forest algorithm was used to check the ability of physicochemical properties to predict lysin sequences as from a G+- or G−-infecting phages. The R package "caret" was employed for creating, fitting, and testing the random forest, and further analyses on the model were performed using packages "pROC" and "randomForest." The data set was randomly partitioned into a training subset (75% of all entries) and a testing subset. The training subset was used to fit the random forest parameters (namely, the randomly selected variables for each node, which was fixed at 4) by 5-fold cross-validation with 3 repeats. Then, the constructed random forest was validated using the previously defined testing subset. Alternatively, a second random forest model was constructed with additional variables (protein length, number of PF hits, and presence or absence of an N-terminal signal peptide), using the method described above.

Sequence similarity networks. SSNs were generated for visually assessing the similarity clustering of sequence sets. For this purpose, the Enzyme Similarity tool from the Enzyme Function Initiative server (EFI-EST) was employed (79). Briefly, this tool performs a local alignment from which every possible pair of sequences receives a score similar to the E value obtained from a typical BLAST analysis. A threshold score value was selected for each SSN so that below this threshold, sequence pairs were considered nonsimilar, and therefore, the pair would not be connected in the resulting representation. Scores were selected so that sequence pairs whose similarity was below 30 to 40% were deemed nonsimilar. The SSN graphs were produced using Cytoscape 3 with the yFiles organic layout (80).

Statistical analysis. Default methods for data representation implemented in the ggplot2 R package, such as kernel density estimation and GAM smoothing, were used throughout this work for data visualization (81). For comparison of nonnormal, heteroskedastic data populations, robust statistical methods were used (82). Specifically, a generalization of Welch's test with trimmed means (default trimming level $\gamma = 0.20$) was used with a Bonferroni adjustment when multiple comparisons were performed. Effect sizes were estimated according to Wilcox and Tian's ζ (83). A general interpretation for ζ is given in that reference, with values of around 0.10 indicating a small effect size, those around 0.30 a medium effect, and those of 0.50 and above a large one. A P value of ≤ 0.05 was considered significant. All robust methods used were from the implementation in the R package WRS2 (84).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
SUPPLEMENTAL FILE 1, XLSX file, 0.6 MB.
SUPPLEMENTAL FILE 2, PDF file, 0.2 MB.

REFERENCES

1. Silver LL. 2011. Challenges of antibacterial discovery. Clin Microbiol Rev 24:71–109. https://doi.org/10.1128/CMR.00030-10.
2. Sabtu N, Enoch DA, Brown NM. 2015. Antibiotic resistance: what, why, where, when and how? Br Med Bull 116:105–113. https://doi.org/10.1093/bmb/ldv041.
3. European Antimicrobial Resistance Surveillance Network. 2019. Surveillance of antimicrobial resistance in Europe 2018.
4. Centers for Disease Control and Prevention. 2019. Antibiotic resistance threats in the United States 2019.
5. Hofer U. 2019. The cost of antimicrobial resistance. Nat Rev Microbiol 17:3. https://doi.org/10.1038/s41579-018-0125-x.
6. Organisation for Economic Co-operation and Development. 2018. Stemming the superbug tide: just a few dollars more.
7. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, Pulcini C, Kahlmeter G, Kluytmans J, Carmeli Y, Ouellette M, Outterson K, Patel J, Cavaleri M, Cox EM, Houchens CR, Grayson ML, Hansen P, Singh N, Theuretzbacher U, Magrini N, the WHO Pathogens Priority List Working Group. 2018. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis 18:318–327. https://doi.org/10.1016/S1473-3099(17)30753-3.
8. O'Neill J. 2016. Tackling drug-resistant infections globally: final report and recommendations. The review of antimicrobial resistance. The Wellcome Trust, London, United Kingdom.
9. Theuretzbacher U, Outterson K, Engel A, Karlén A. 2020. The global preclinical antibacterial pipeline. Nat Rev Microbiol 18:275–285. https://doi.org/10.1038/s41579-019-0288-0.
10. McCallin S, Sacher JC, Zheng J, Chan BK. 2019. Current state of compassionate phage therapy. Viruses 11:343. https://doi.org/10.3390/v11040343.
11. Abdelkader K, Gerstmans H, Saafan A, Dishisha T, Briers Y. 2019. The preclinical and clinical progress of bacteriophages and their lytic enzymes: the parts are easier than the whole. Viruses 11:96. https://doi.org/10.3390/v11020096.
12. Young R. 2014. Phage lysis: three steps, three choices, one outcome. J Microbiol 52:243–258. https://doi.org/10.1007/s12275-014-4087-z.
13. Payne KM, Hatfull GF. 2012. Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities. PLoS One 7:e34052. https://doi.org/10.1371/journal.pone.0034052.
14. Kongari R, Rajaure M, Cahill J, Rasche E, Mijalis E, Berry J, Young R. 2018. Phage spanins: diversity, topological dynamics and gene convergence. BMC Bioinformatics 19:326. https://doi.org/10.1186/s12859-018-2342-8.
15. Dams D, Briers Y. 2019. Enzybiotics: enzyme-based antibacterials as therapeutics. Adv Exp Med Biol 1148:233–253. https://doi.org/10.1007/978-981-13-7709-9_11.

16. Pastagia M, Schuch R, Fischetti VA, Huang DB. 2013. Lysins: the arrival of pathogen-directed anti-infectives. J Med Microbiol 62:1506–1516. https://doi.org/10.1099/jmm.0.061028-0.
17. Cheng M, Zhang Y, Li X, Liang J, Hu L, Gong P, Zhang L, Cai R, Zhang H, Ge J, Ji Y, Guo Z, Feng X, Sun C, Yang Y, Lei L, Han W, Gu J. 2017. Endolysin LysEF-P10 shows potential as an alternative treatment strategy for multi-drug-resistant *Enterococcus faecalis* infections. Sci Rep 7:10164. https://doi.org/10.1038/s41598-017-10755-7.
18. Díez-Martínez R, de Paz HD, Bustamante N, García E, Menéndez M, García P. 2013. Improving the lethal effect of Cpl-7, a pneumococcal phage lysozyme with broad bactericidal activity, by inverting the net charge of its cell wall-binding module. Antimicrob Agents Chemother 57:5355–5365. https://doi.org/10.1128/AAC.01372-13.
19. Kusuma C, Jadanova A, Chanturiya T, Kokai-Kun JF. 2007. Lysostaphin-resistant variants of *Staphylococcus aureus* demonstrate reduced fitness in vitro and in vivo. Antimicrob Agents Chemother 51:475–482. https://doi.org/10.1128/AAC.00786-06.
20. Vázquez R, García E, García P. 2018. Phage lysins for fighting bacterial respiratory infections: a new generation of antimicrobials. Front Immunol 9:2252. https://doi.org/10.3389/fimmu.2018.02252.
21. Gerstmans H, Criel B, Briers Y. 2018. Synthetic biology of modular endolysins. Biotechnol Adv 36:624–640. https://doi.org/10.1016/j.biotechadv.2017.12.009.
22. Heselpoth RD, Euler CW, Schuch R, Fischetti VA. 2019. Lysocins: bioengineered antimicrobials that deliver lysins across the outer membrane of Gram-negative bacteria. Antimicrob Agents Chemother 63:e00342-19. https://doi.org/10.1128/AAC.00342-19.
23. Diaz E, López R, Garcia JL. 1991. Chimeric pneumococcal cell wall lytic enzymes reveal important physiological and evolutionary traits. J Biol Chem 266:5464–5471. https://doi.org/10.1016/S0021-9258(19)67618-0.
24. Gerstmans H, Grimon D, Gutiérrez D, Lood C, Rodríguez A, van Noort V, Lammertyn J, Lavigne R, Briers Y. 2020. A VersaTile-driven platform for rapid hit-to-lead development of engineered lysins. Sci Adv 6:eaaz1136. https://doi.org/10.1126/sciadv.aaz1136.
25. Seijsing J, Sobieraj AM, Keller N, Shen Y, Zinkernagel AS, Loessner MJ, Schmelcher M. 2018. Improved biodistribution and extended serum half-life of a bacteriophage endolysin by albumin binding domain fusion. Front Microbiol 9:2927. https://doi.org/10.3389/fmicb.2018.02927.
26. Díez-Martínez R, De Paz HD, García-Fernández E, Bustamante N, Euler CW, Fischetti VA, Menendez M, García P. 2015. A novel chimeric phage lysin with high *in vitro* and *in vivo* bactericidal activity against *Streptococcus pneumoniae*. J Antimicrob Chemother 70:1763–1773. https://doi.org/10.1093/jac/dkv038.

27. Abouhmad A, Mamo G, Dishisha T, Amin MA, Hatti-Kaul R. 2016. T4 lysozyme fused with cellulose-binding module for antimicrobial cellulosic wound dressing materials. J Appl Microbiol 121:115–125. https://doi.org/10.1111/jam.13146.

28. Ghose C, Euler CW. 2020. Gram-negative bacterial lysins. Antibiotics (Basel) 9:74. https://doi.org/10.3390/antibiotics9020074.

29. Briers Y, Lavigne R. 2015. Breaking barriers: expansion of the use of endolysins as novel antibacterials against Gram-negative bacteria. Future Microbiol 10:377–390. https://doi.org/10.2217/fmb.15.8.

30. Briers Y, Walmagh M, Van Puyenbroeck V, Cornelissen A, Cenens W, Aertsen A, Oliveira H, Azeredo J, Verween G, Pirnay J-P, Miller S, Volckaert G, Lavigne R. 2014. Engineered endolysin-based "Artilysins" to combat multidrug-resistant Gram-negative pathogens. mBio 5:e01379-14. https://doi.org/10.1128/mBio.01379-14.

31. Zampara A, Sørensen MCH, Grimon D, Antenucci F, Vitt AR, Bortolaia V, Briers Y, Brøndsted L. 2020. Exploiting phage receptor binding proteins to enable endolysins to kill Gram-negative bacteria. Sci Rep 10:12087. https://doi.org/10.1038/s41598-020-68983-3.

32. Morita M, Tanji Y, Orito Y, Mizoguchi K, Soejima A, Unno H. 2001. Functional analysis of antibacterial activity of *Bacillus amyloliquefaciens* phage endolysin against Gram-negative bacteria. FEBS Lett 500:56–59. https://doi.org/10.1016/s0014-5793(01)02587-x.

33. Guo M, Feng C, Ren J, Zhuang X, Zhang Y, Zhu Y, Dong K, He P, Guo X, Qin J. 2017. A novel antimicrobial endolysin, LysPA26, against *Pseudomonas aeruginosa*. Front Microbiol 8:293. https://doi.org/10.3389/fmicb.2017.00293.

34. Lood R, Winer BY, Pelzek AJ, Diez-Martinez R, Thandar M, Euler CW, Schuch R, Fischetti VA. 2015. Novel phage lysin capable of killing the multidrug-resistant Gram-negative bacterium *Acinetobacter baumannii* in a mouse bacteremia model. Antimicrob Agents Chemother 59:1983–1991. https://doi.org/10.1128/AAC.04641-14.

35. Düring K, Porsch P, Mahn A, Brinkmann O, Gieffers W. 1999. The non-enzymatic microbicidal activity of lysozymes. FEBS Lett 449:93–100. https://doi.org/10.1016/s0014-5793(99)00405-6.

36. Rotem S, Radzishevsky I, Inouye RT, Samore M, Mor A. 2006. Identification of antimicrobial peptide regions derived from genomic sequences of phage lysins. Peptides 27:18–26. https://doi.org/10.1016/j.peptides.2005.07.001.

37. Ibrahim HR, Higashiguchi S, Koketsu M, Juneja LR, Kim M, Yamamoto T, Sugimoto Y, Aoki T. 1996. Partially unfolded lysozyme at neutral pH agglutinates and kills Gram-negative and Gram-positive bacteria through membrane damage mechanism. J Agric Food Chem 44:3799–3806. https://doi.org/10.1021/jf960133x.

38. Orito Y, Morita M, Hori K, Unno H, Tanji Y. 2004. *Bacillus amyloliquefaciens* phage endolysin can enhance permeability of *Pseudomonas aeruginosa* outer membrane and induce cell lysis. Appl Microbiol Biotechnol 65:105–109. https://doi.org/10.1007/s00253-003-1522-1.

39. Thandar M, Lood R, Winer BY, Deutsch DR, Euler CW, Fischetti VA. 2016. Novel engineered peptides of a phage lysin as effective antimicrobials against multidrug-resistant *Acinetobacter baumannii*. Antimicrob Agents Chemother 60:2671–2679. https://doi.org/10.1128/AAC.02972-15.

40. Maciejewska B, Żrubek K, Espaillat A, Wiśniewska M, Rembacz KP, Cava F, Dubin G, Drulis-Kawa Z. 2017. Modular endolysin of *Burkholderia* AP3 phage has the largest lysozyme-like catalytic subunit discovered to date and no catalytic aspartate residue. Sci Rep 7:14501. https://doi.org/10.1038/s41598-017-14797-9.

41. Sykilinda NN, Nikolaeva AY, Shneider MM, Mishkin DV, Patutin AA, Popov VO, Boyko KM, Klyachko NL, Miroshnikov KA. 2018. Structure of an *Acinetobacter* broad-range prophage endolysin reveals a C-terminal α-helix with the proposed role in activity against live bacterial cells. Viruses 10:309. https://doi.org/10.3390/v10060309.

42. Peng S-Y, You R-I, Lai M-J, Lin N-T, Chen L-K, Chang K-C. 2017. Highly potent antimicrobial modified peptides derived from the *Acinetobacter baumannii* phage endolysin LysAB2. Sci Rep 7:11477. https://doi.org/10.1038/s41598-017-11832-7.

43. Oliveira H, Melo LD, Santos SB, Nobrega FL, Ferreira EC, Cerca N, Azeredo J, Kluskens LD. 2013. Molecular aspects and comparative genomics of bacteriophage endolysins. J Virol 87:4558–4570. https://doi.org/10.1128/JVI.03277-12.

44. Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC, Funk K, Ketter A, Kim S, Kimchi A, Kitts PA, Kuznetsov A, Lathrop S, Lu Z, McGarvey K, Madden TL, Murphy TD, O'Leary N, Phan L, Schneider VA, Thibaud-Nissen F, Trawick BW, Pruitt KD, Ostell J. 2020. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 48:D9–D16. https://doi.org/10.1093/nar/gkz899.

45. Li Q, Cheng W, Morlot C, Bai X-H, Jiang Y-L, Wang W, Roper DI, Vernet T, Dong Y-H, Chen Y, Zhou C-Z. 2015. Full-length structure of the major autolysin LytA. Acta Crystallogr Sect D Biol Crystallogr 71:1373–1381. https://doi.org/10.1107/S1399004715007403.

46. Lai WCB, Chen X, Ho MKY, Xia J, Leung SSY. 2020. Bacteriophage-derived endolysins to target gram-negative bacteria. Int J Pharm 589:119833. https://doi.org/10.1016/j.ijpharm.2020.119833.

47. Fernández-Tornero C, López R, García E, Giménez-Gallego G, Romero A. 2001. A novel solenoid fold in the cell wall anchoring domain of the pneumococcal virulence factor LytA. Nat Struct Biol 8:1020–1024. https://doi.org/10.1038/nsb724.

48. Fernández-Tornero C, Ramón A, Fernández-Cabrera C, Giménez-Gallego G, Romero A. 2002. Expression, crystallization and preliminary X-ray diffraction studies on the complete choline-binding domain of the major pneumococcal autolysin. Acta Crystallogr D Biol Crystallogr 58:556–558. https://doi.org/10.1107/s0907444902000537.

49. Bustamante N, Iglesias-Bexiga M, Bernardo-García N, Silva-Martín N, García G, Campanero-Rhodes MA, García E, Usón I, Buey RM, García P, Hermoso JA, Bruix M, Menéndez M. 2017. Deciphering how Cpl-7 cell wall-binding repeats recognize the bacterial peptidoglycan. Sci Rep 7:16494. https://doi.org/10.1038/s41598-017-16392-4.

50. Vázquez R, García P. 2019. Synergy between two chimeric lysins to kill *Streptococcus pneumoniae*. Front Microbiol 10:1251. https://doi.org/10.3389/fmicb.2019.01251.

51. Rodríguez-Rubio L, Martínez B, Rodríguez A, Donovan DM, Götz F, García P. 2013. The phage lytic proteins from the *Staphylococcus aureus* bacteriophage vB_SauS-phiIPLA88 display multiple active catalytic domains and do not trigger staphylococcal resistance. PLoS One 8:e64671. https://doi.org/10.1371/journal.pone.0064671.

52. Oechslin F, Menzi C, Moreillon P, Resch G. 2021. The multi-domain architecture of a bacteriophage endolysin enables intramolecular synergism and regulation of bacterial lysis. J Biol Chem. https://doi.org/10.1016/j.jbc.2021.100639.

53. Loessner MJ, Kramer K, Ebel F, Scherer S. 2002. C-terminal domains of *Listeria monocytogenes* bacteriophage murein hydrolases determine specific recognition and high-affinity binding to bacterial cell wall carbohydrates. Mol Microbiol 44:335–349. https://doi.org/10.1046/j.1365-2958.2002.02889.x.

54. Son B, Kong M, Ryu S. 2018. The auxiliary role of the amidase domain in cell wall binding and exolytic activity of staphylococcal phage endolysins. Viruses 10:284. https://doi.org/10.3390/v10060284.

55. Oliveira H, Sampaio M, Melo LDR, Dias O, Pope WH, Hatfull GF, Azeredo J. 2019. Staphylococci phages display vast genomic diversity and evolutionary relationships. BMC Genomics 20:357. https://doi.org/10.1186/s12864-019-5647-8.

56. Lu JZ, Fujiwara T, Komatsuzawa H, Sugai M, Sakon J. 2006. Cell wall-targeting domain of glycylglycine endopeptidase distinguishes among peptidoglycan cross-bridges. J Biol Chem 281:549–558. https://doi.org/10.1074/jbc.M509691200.

57. Bateman A, Rawlings ND. 2003. The CHAP domain: a large family of amidases including GSP amidase and peptidoglycan hydrolases. Trends Biochem Sci 28:234–237. https://doi.org/10.1016/S0968-0004(03)00061-6.

58. Rossi P, Aramini JM, Xiao R, Chen CX, Nwosu C, Owens LA, Maglaqui M, Nair R, Fischer M, Acton TB, Honig B, Rost B, Montelione GT. 2009. Structural elucidation of the Cys-His-Glu-Asn proteolytic relay in the secreted CHAP domain enzyme from the human pathogen *Staphylococcus saprophyticus*. Proteins 74:515–519. https://doi.org/10.1002/prot.22267.

59. Schleifer KH, Kandler O. 1972. Peptidoglycan types of bacterial cell walls and their taxonomic implications. Bacteriol Rev 36:407–477. https://doi.org/10.1128/BR.36.4.407-477.1972.

60. Chen Y, Simmonds RS, Timkovich R. 2013. Proposed docking interface between peptidoglycan and the target recognition domain of zoocin A. Biochem Biophys Res Commun 441:297–300. https://doi.org/10.1016/j.bbrc.2013.09.087.

61. Beaussart A, Rolain T, Duchene MC, El-Kirat-Chatel S, Andre G, Hols P, Dufrene YF. 2013. Binding mechanism of the peptidoglycan hydrolase Acm2: low affinity, broad specificity. Biophys J 105:620–629. https://doi.org/10.1016/j.bpj.2013.06.035.

62. Grundling A, Schneewind O. 2006. Cross-linked peptidoglycan mediates lysostaphin binding to the cell wall envelope of *Staphylococcus aureus*. J Bacteriol 188:2463–2472. https://doi.org/10.1128/JB.188.7.2463-2472.2006.

63. Low LY, Yang C, Perego M, Osterman A, Liddington R. 2011. Role of net charge on catalytic domain and influence of cell wall binding domain on bactericidal activity, specificity, and host range of phage lysins. J Biol Chem 286:34391–34403. https://doi.org/10.1074/jbc.M111.244160.

64. Horgan M, O'Flynn G, Garry J, Cooney J, Coffey A, Fitzgerald GF, Ross RP, McAuliffe O. 2009. Phage lysin LysK can be truncated to its CHAP domain and retain lytic activity against live antibiotic-resistant staphylococci. Appl Environ Microbiol 75:872–874. https://doi.org/10.1128/AEM.01831-08.

65. Cloutier I, Paradis-Bleau C, Giroux A-M, Pigeon X, Arseneault M, Levesque RC, Auger M. 2010. Biophysical studies of the interactions between the phage ϕKZ gp144 lytic transglycosylase and model membranes. Eur Biophys J 39:263–276. https://doi.org/10.1007/s00249-009-0530-1.

66. Morita M, Tanji Y, Mizoguchi K, Soejima A, Orito Y, Unno H. 2001. Antibacterial activity of *Bacillus amyloliquefaciens* phage endolysin without holin conjugation. J Biosci Bioeng 91:469–473. https://doi.org/10.1263/jbb.91.469.

67. Holt AJ, Ramsey J, O'Leary C, Moreland R, Martin C, Galbadage DT, Sharan R, Sule P, Bettridge K, Xiao J, Cirillo J, Young R. 2019. Phage-encoded cationic antimicrobial peptide used for outer membrane disruption in lysis. bioRxiv https://doi.org/10.1101/515445.

68. Larpin Y, Oechslin F, Moreillon P, Resch G, Entenza JM, Mancini S. 2018. In vitro characterization of PlyE146, a novel phage lysin that targets Gram-negative bacteria. PLoS One 13:e0192507. https://doi.org/10.1371/journal.pone.0192507.

69. Safari F, Sharifi M, Farajnia S, Akbari B, Ahmadi MKB, Negahdaripour M, Ghasemi Y. 2020. The interaction of phages and bacteria: the co-evolutionary arms race. Crit Rev Biotechnol 40:119–137. https://doi.org/10.1080/07388551.2019.1674774.

70. Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682. https://doi.org/10.1093/bioinformatics/btq003.

71. Madeira F, Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636–W641. https://doi.org/10.1093/nar/gkz268.

72. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448.

73. Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 35:W429–W432. https://doi.org/10.1093/nar/gkm256.

74. Vázquez R, García E, García P. 2020. Curated phage lysins collection including identifiers, amino acid sequences, functional domain predictions, architectures and physicochemical properties calculations. DIGITAL.CSIC. http://hdl.handle.net/10261/221469.

75. Osorio D, Rondón-Villarreal P, Torres R. 2015. Peptides: a package for data mining of antimicrobial peptides. R J 7:4–14. https://doi.org/10.32614/RJ-2015-001.

76. Dawson RMC. 1986. Data for biochemical research, 3rd ed. Clarendon Press, Oxford, United Kingdom.

77. Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132. https://doi.org/10.1016/0022-2836(82)90515-0.

78. Eisenberg D, Weiss RM, Terwilliger TC. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A 81:140–144. https://doi.org/10.1073/pnas.81.1.140.

79. Zallot R, Oberg N, Gerlt JA. 2019. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 58:4169–4182. https://doi.org/10.1021/acs.biochem.9b00735.

80. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303.

81. Wickham H. 2016. ggplot2: elegant graphics for data analysis, 2nd ed. Springer-Verlag, New York, NY.

82. Wilcox RR. 2012. Introduction to robust estimation and hypothesis testing, 3rd ed. Elsevier Inc., Amsterdam, The Netherlands.

83. Wilcox RR, Tian TS. 2011. Measuring effect size: a robust heteroscedastic approach for two or more groups. J Appl Stat 38:1359–1368. https://doi.org/10.1080/02664763.2010.498507.

84. Mair P, Wilcox R. 2020. Robust statistical methods in R using the WRS2 package. Behav Res Methods 52:464–488. https://doi.org/10.3758/s13428-019-01246-w.