MICROBIOLOGY

Analysis of bacterial pangenomes reduces CRISPR dark matter and reveals strong association between membranome and CRISPR-Cas systems

Alejandro Rubio[1], Maximilian Sprang[2], Andrés Garzón[1], Antonio Moreno-Rodriguez[1], Maria Eugenia Pachón-Ibáñez[3,4], Jerónimo Pachón[3,5], Miguel A. Andrade-Navarro[2], Antonio J. Pérez-Pulido[1]*

CRISPR-Cas systems are prokaryotic acquired immunity mechanisms, which are found in 40% of bacterial genomes. They prevent viral infections through small DNA fragments called spacers. However, the vast majority of these spacers have not yet been associated with the virus they recognize, and it has been named CRISPR dark matter. By analyzing the spacers of tens of thousands of genomes from six bacterial species, we have been able to reduce the CRISPR dark matter from 80% to as low as 15% in some of the species. In addition, we have observed that, when a genome presents CRISPR-Cas systems, this is accompanied by particular sets of membrane proteins. Our results suggest that when bacteria present membrane proteins that make it compete better in its environment and these proteins are, in turn, receptors for specific phages, they would be forced to acquire CRISPR-Cas.

INTRODUCTION

Bacteriophages, also known as phages, are viruses that predate bacterial cells, representing environmental burdens for their growth and spread. They can be used to control bacterial growth and are even beginning to be used to treat infections in humans (*1*). Bacteria defend themselves against infection by these phages by means of different molecular strategies. Restriction-modification systems are by far the most abundant, being present in 83% of prokaryotic genomes, followed by CRISPR-Cas with about 40% (*2*). CRISPR-Cas is an adaptive immunity system found in most archaea and in less than half of the bacteria sequenced (*3*). They provide acquired immune resistance against phages and other foreign nucleic acid molecules such as plasmids, thus restricting gene transfer (*4*). There are different types of CRISPR-Cas systems based on genes that are part of the different steps of this immune system (adaptation or spacer integration, expression, and interference) and are generically called *cas* (CRISPR-associated) genes.

The acquired immunity of CRISPR-Cas systems is based on short nucleotide fragments, called spacers. These are originated from fragments of the foreign nucleic acid sequence captured during an earlier entry into the bacterial cell, called protospacers, which were inserted into the bacterial DNA next to the *cas* genes. These spacers are mostly similar to phage sequences and, to a lesser extent, to other extrachromosomal nucleic acid molecules. However, a large proportion of them have no known protospacer (over 80 to 90%). These spacers of unknown origin are believed to originate from as yet to be sequenced phages and constitute what has been called the CRISPR "dark matter" (*5*).

Bacterial genomes with CRISPR-Cas systems have *cas* genes along with their spacers, which are separated by sequence repeats (short identical or nearly identical sequences). However, other genes have been associated with these systems because specific functionalities have been found in strains that have CRISPR-Cas systems. For example, a relationship has been demonstrated with the formation of multicellular structures called biofilms in *Pseudomonas aeruginosa*, *Streptococcus mutant*, and *Yersinia pestis* (*6*–*8*). Connections with the regulation of outer membrane proteins have also been described in *Salmonella Typhi* (*9*), and a specific relationship with virulence have also been shown in multiple bacterial species (*10*–*12*).

In a previous work with *Acinetobacter baumannii*, we found that strains with CRISPR-Cas systems had specific genes involved in biofilm, in addition to genes encoding membrane lipoproteins and proteins with signal peptides (*13*). The analysis was based on a pangenome constructed from this species (all of the different genes found in the genomes of the species). The establishment of these pangenomes is now easier because of the large number of genomes available in public databases, and they allow us to analyze the accessory genome, which is the set of genes that is not present in all the strains of a species (*14*) and are usually acquired by horizontal gene transfer (*15*). If strains with CRISPR-Cas systems have special functions that are absent in strains without these systems, then accessory genes involved in these functions should appear almost exclusively in the former.

In this work, we analyzed thousands of genomes of the group of bacteria known as ESKAPE, whose acronym refers to two Gram-positive bacteria (*Enterococcus faecium* and *Staphylococcus aureus*) and four Gram-negative bacteria (*Klebsiella pneumoniae*, *A. baumannii*, *P. aeruginosa*, and *Enterobacter cloacae*). ESKAPE bacteria have CRISPR-Cas systems of the most common types, from I to IV, but only in a minimal number of strains, with

[1]Andalusian Centre for Developmental Biology (CABD, UPO-CSIC-JA), Faculty of Experimental Sciences (Genetics Department), University Pablo de Olavide, 41013 Seville, Spain. [2]Faculty of Biology, Johannes Gutenberg-Universität Mainz, Biozentrum I, Hans-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany. [3]Institute of Biomedicine of Seville (IBiS), Virgen del Rocío Hospital/CSIC/University of Seville, Seville, Spain. [4]CIBER de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, Madrid, Spain. [5]Department of Medicine, School of Medicine, University of Seville, Seville, Spain.
*Corresponding author. Email: ajperez@upo.es

frequencies ranging from less than 1 to 60% of genomes, depending on the species (*16*). We have created a large pangenome for each species and compared strains with or without CRISPR-Cas systems to discover genes associated with the first group. Then, these genes have been functionally analyzed, and we found that they are enriched in genes encoding membrane proteins. This has motivated us to investigate whether this relationship could be mediated by phages that could take advantage of the existence of these membrane proteins as receptors or adhesion sites to infect bacteria. In addition, our results demonstrate that the study of thousands of genomes of the same species allows us to reduce the CRISPR dark matter and to trace the origin of most of the spacers found in them.

RESULTS

A large proportion of CRISPR dark matter spacers can be annotated by pangenome analysis

The genomes of the different ESKAPE species were initially obtained, and they were both structurally and functionally annotated with a special emphasis on the protein-coding genes and the elements that are part of the CRISPR-Cas systems. According to the number of genes, the smallest pangenome was found for *S. aureus* despite having started from a larger number of genomes (Fig. 1, A and B). The species with the largest number of genomes with CRISPR-Cas systems were *K. pneumoniae* and *P. aeruginosa*, which also had the largest pangenomes (along with the other Gram-negative species) and more than 3500 core genes, i.e., genes common to all genomes of the species. The difference between Gram-negative and Gram-positive bacteria is also evident when comparing the number of genes per genome, with *P. aeruginosa* showing both the largest number of genes and average number of shared genes (Fig. 1C). Last, the major difference between genes per genome and average number of shared genes is found in *E. cloacae*, which could be explained by the low number of genomes used for this species (n = 317).

The overall proportion of genomes exhibiting CRISPR-Cas systems is low, with Gram-positive bacteria having them in only 1% of their genomes, both *A. baumannii* and *E. cloacae* in 12%, and *K. pneumoniae* and *P. aeruginosa* in 30% and 47%, respectively (Fig. 1D). Because the spacers of CRISPR-Cas systems usually recognize mobile genetic elements, we first separated pangenome genes that could come from plasmids (an average of 28% of the genes) and phages (an average of 8% of the genes) (Fig. 1E). An average of 5% of the genes were annotated as originating from both genetic elements, which could come from what is known as phage-plasmids (*17*).

Next, the spacers were obtained, and their cognate protospacers were searched for within the pangenome genes. The major number of different spacers was found in the three species with CRISPR-Cas IV and/or I-F types (Fig. 1F). As expected, protospacers belonged in a much larger proportion to phage genes. It was possible to annotate more than 65% of the spacers in the same three species referred to above, with a maximum of 85% in *P. aeruginosa*, and the species with a lower number of annotated spacers (about 25%) were *E. cloacae* and *S. aureus*. It is noteworthy that about 18% of the spacers in CRISPR-Cas type I species appear to recognize genes annotated as phage-plasmids.

In addition, a small proportion of the spacers match other bacterial genes (less than 10%). When we analyzed these genes, part of them appear to be additional viral genes not previously found, as reflected by their annotated functions, including sialidases and proteins with the Ead/E22 domain typical of phage proteins (InterPro: Ead/Ea22-like protein; fig. S1). However, genes related to other functions also appear. A notable case is the flagellum and cilium annotations, which could suggest noncanonical functions of CRISPR-Cas systems. In *Salmonella enterica*, it has been shown that these systems could regulate the expression of flagellar genes, ultimately related to biofilm formation (*18*).

Because the presence of CRISPR-Cas systems prevents the entry of foreign DNA (including resistance and virulence plasmids) into the bacteria, the number of genes involved in these functions was compared between genomes with and without CRISPR-Cas systems. On average, genomes with CRISPR-Cas systems presented a lower number of resistance and virulence genes (fig. S2). The most significant difference was found with CRISPR-Cas types II and III, except for resistance genes in *S. aureus*. On the other hand, *P. aeruginosa* presented the highest number of virulence genes overall, although the genomes without CRISPR-Cas systems presented a lower number than those with CRISPR-Cas systems, a result that was repeated with the resistance genes. So, we cannot conclude that all genomes with CRISPR-Cas systems tend to carry a lower number of resistance and virulence genes.

CRISPR-Cas systems appear and disappear throughout the entire phylogeny

At this point, we wanted to know whether the CRISPR-Cas systems were linked to a branch of the phylogeny of the species studied. To define phylogenetic relationships between genomes with and without CRISPR-Cas systems, the multilocus sequence typing (MLST) was used. This is based on several housekeeping genes of each species (*19*), and two adjacent groups reflect genomes arising from a recent common ancestor. Except for *E. faecium* type II and two specific aggregations in *A. baumannii*, CRISPR-Cas systems appear to be spread throughout the phylogenetic tree (Fig. 2, top), suggesting a possible gain by horizontal gene transfer in genomes for which it could provide an evolutionary advantage and a possible subsequent loss when that advantage no longer exists.

If CRISPR-Cas systems were associated with other adaptive bacterial physiological functions that depend on the presence of certain gene functions, as in the case of genomic islands, then this would lead to genomes with CRISPR-Cas systems always having a similar collection of accessory genes, regardless of the *cas* genes. To test this idea, distance trees were constructed between the same MLST groups in the phylogeny, in this case, based on the gene profile of the genomes (gene presence/absence matrix). These gene profiles showed a similar dispersion to that found with molecular phylogeny, apart for certain aggregations of genome groups with CRISPR-Cas systems again in *E. faecium* and *A. baumannii*, suggesting that CRISPR-Cas systems do not appear in genomes with a fixed collection of accessory genes (Fig. 2, bottom).

Molecular phylogenies and gene profiles do not appear to be strongly correlated in general. We found that groups of genomes that have CRISPR-Cas systems and that are phylogenetically close may have a very different gene profile (see groups of *K. pneumoniae* and *P. aeruginosa* genomes at phylogenetic distance 0 in fig. S3). Likewise, we found groups of genomes with a very similar gene profile, which contain CRISPR-Cas systems but are phylogenetically far apart (see groups of *K. pneumoniae*, *A. baumannii*, and *P.*

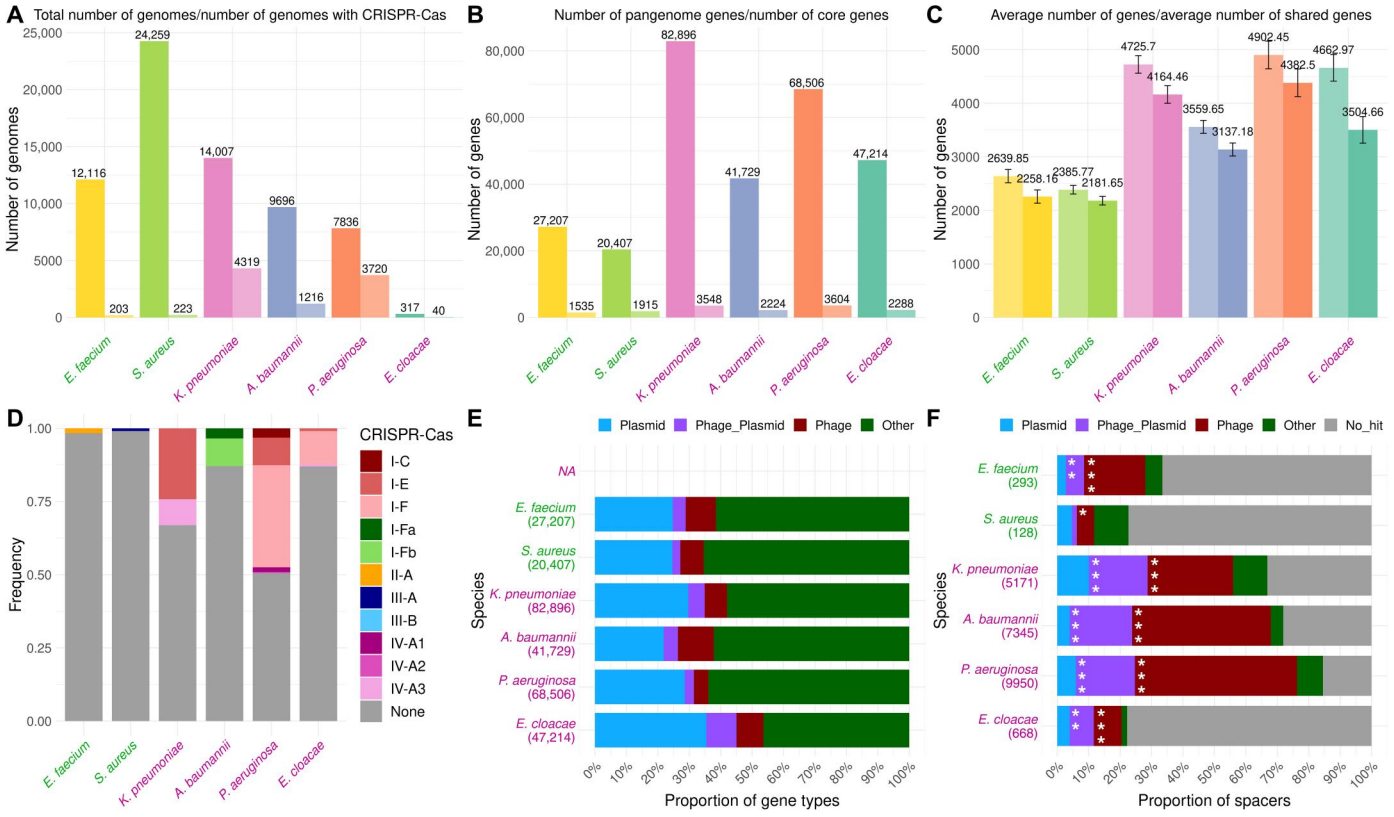

Fig. 1. Summary of analyzed genomes. (**A**) Total number of genomes analyzed and number of genomes having CRISPR-Cas systems. (**B**) Number of genes in the pangenome and number of core genes. (**C**) Average number of genes per genome and average number of shared genes (average of the number of genes shared for each genome with the remaining genomes). (**D**) Distribution of CRISPR-Cas systems among genomes. (**E**) Proportion of gene types in the pangenome. For each species, the number of total genes is shown (below the species name), as well as the proportion of them that were annotated as belonging to plasmid sequences, phages, both plasmids and phages, and other genes not included in plasmids or phages. (**F**) Spacers and assigned protospacer types by species. For each species, the number of total spacers is shown (below the species name), as well as the proportion of them that match with plasmid sequences, phages, sequences annotated as both plasmid and phage, other genes not included in plasmids or phages, and the spacers that do not match any gene (No_hit). The names of Gram-positive bacteria appear in green, and those of Gram-negative bacteria are in purple. Asterisks highlight protospacer types with significant differences according to a hypergeometric test (from $*P \leq 0.05$ to $***P \leq 1 \times 10^{-20}$).

aeruginosa genomes with a distance of 0.3 in the gene profile). Together, this reinforces the idea that CRISPR-Cas systems do not appear to be linked to strains phylogenetically related or to specific accessory genomes, but this raised the question of whether genomes with CRISPR-Cas systems presented particular genes at higher frequencies than genomes without these systems.

Genes associated with CRISPR-Cas systems mainly encode membrane proteins

We have already seen that CRISPR-Cas systems are associated with different accessory genomes and appear and disappear at any evolutionary time. However, assuming that CRISPR-Cas systems are associated with other physiological functions of bacteria, we should find some accessory genes more frequently in genomes having these systems. On the basis of this hypothesis, we searched for genes significantly associated with genomes showing CRISPR-Cas systems, excluding the *cas* genes themselves. Thus, a median of 133 ± 16 genes per species were found associated with the different CRISPR-Cas systems, while only 17 ± 25 genes were associated with the absence of CRISPR-Cas systems (table S3). To test whether the CRISPR-Cas–associated genes were significantly involved in any biological process or function, enrichment analysis was performed, and it was found that genes encoding membrane proteins were highly prominent (Fig. 3). By species and type of CRISPR-Cas system, a median of 32 ± 19 membrane proteins were found. These included pilus proteins of *P. aeruginosa* and *K. pneumoniae*, as well as outer membrane proteins of *A. baumannii* and *K. pneumoniae*, and other *A. baumannii* and *P. aeruginosa* proteins involved in type II secretion systems. This relationship with membrane proteins was especially relevant in CRISPR-Cas type I systems, while other genes involved in catabolic processes or DNA metabolism were found more notably in types II, III, and IV (fig. S4). In addition, types IV were also notable for the annotation "extrachromosomal DNA," reflecting the origin of these systems from mobile genetic elements (*20*). Last, it should be noted that only a median of 6 ± 25 membrane proteins were found associated with genomes lacking CRISPR-Cas systems.

It is noteworthy that the species that did not present any enrichment in surface proteins was *E. faecium*, which has only the CRISPR-Cas type II system. In this case, intracellular proteins stand out, such as those of the citrate lyase complex, which suggests some unknown relationship between these two elements. In

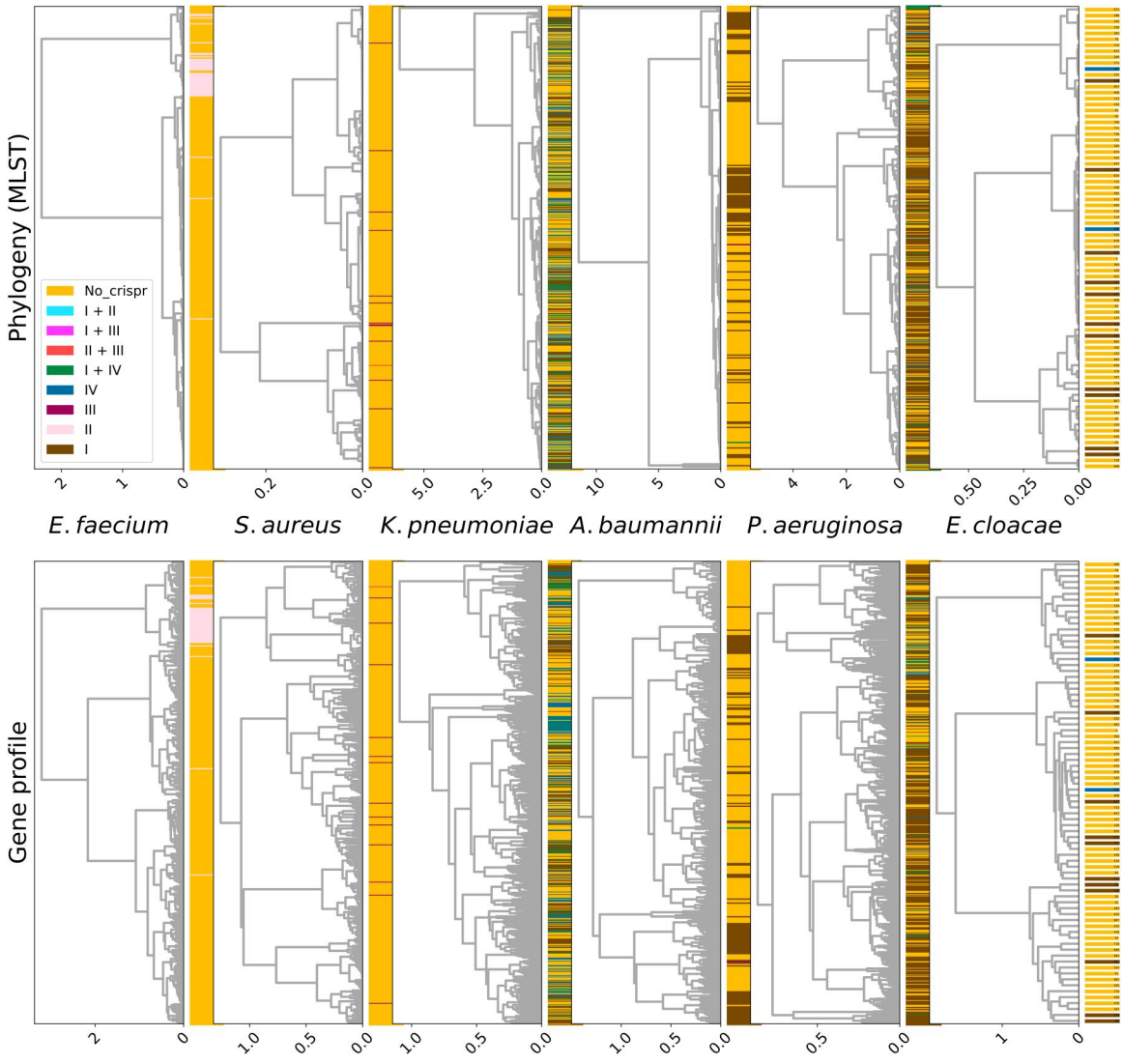

Fig. 2. Molecular phylogeny for each species based on MLST and gene profile. The legend shows the different types of CRISPR-Cas (without specifying the subtype) with orange indicating the MLST group lacking any CRISPR-Cas system.

addition, in the case of type III *S. aureus*, the relationship with membrane proteins is weaker than in the rest of the cases, which present type I. However, it should be noted that we have a low number of genomes with CRISPR-Cas systems of these types, 203 genomes with type II and 222 with type III, compared with the type I of *P. aeruginosa*, *A. baumannii*, or *K. pneumoniae* (with 3902, 1257, and 3508 genomes, respectively), and this fact could be biasing the results obtained for both types II and III.

Genomes with specific types of CRISPR-Cas systems have different sets of membrane proteins

Because the genomes bearing CRISPR-Cas systems presented specific types of membrane proteins, we wanted to know whether all genomes with a specific CRISPR-Cas type had the same set of membrane proteins. To assess this, genes encoding membrane proteins previously associated with CRISPR-Cas were searched for in genomes bearing these systems. In general, there were at least two clusters of genomes, especially when type I was analyzed, each one presenting a different collection of genes encoding membrane proteins (Fig. 4), except for type I-F in *E. cloacae*, as the number of genomes in this species is low. However, the separation between these two clusters of genomes was not as clear as with the other types of CRISPR-Cas systems (fig. S5). Some of the two clusters that appear when analyzing genomes with the type I CRISPR-Cas system can be further divided into two (*P. aeruginosa* and I-Fa in *A. baumannii*). However, because these two new clusters share many of the membrane proteins, we decided to create and analyze henceforth only the two main clusters.

As an example, about half of the *A. baumannii* genomes with the CRISPR-Cas type I-Fb system carry the *opuD* and *betP* genes, involved in choline and glycine betaine transport, which could protect the bacterium from osmotic stress (*21*), and the other half of the genomes have porins such as *benP* or a specific variant of the TonB-dependent siderophore receptor *bauA* (*22*). The clusters found with genomes having CRISPR-Cas systems do not seem to be normally associated with the phylogeny of the corresponding

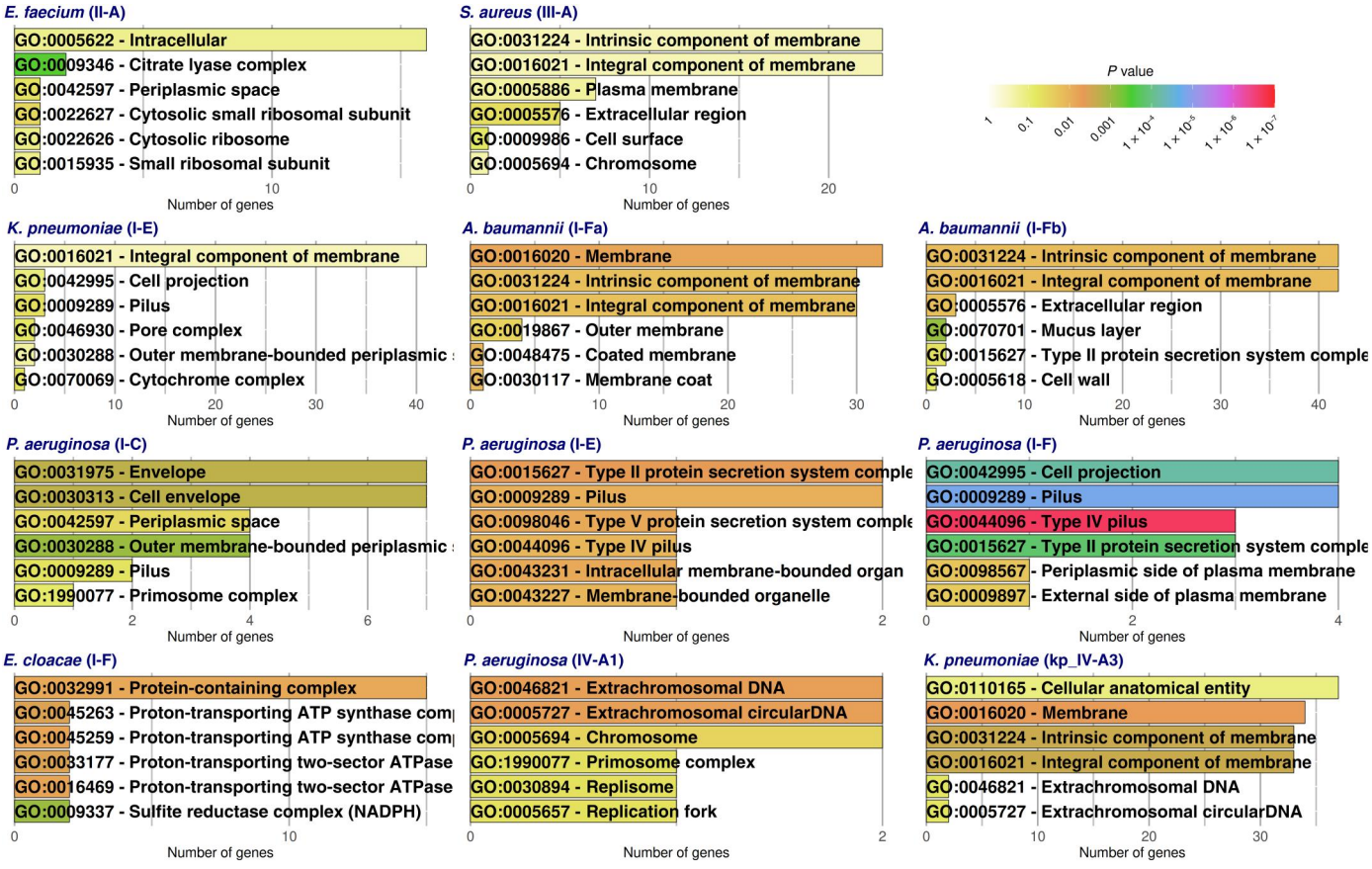

Fig. 3. Functional enrichment of genes associated with different CRISPR-Cas systems in the different species. Gene Ontology (GO) cellular component was used in these enrichments, with functional annotations of the pangenome obtained by Sma3s. Enrichment with GO biological process and molecular function can be found in fig. S4.

species because the genomes that present the membrane proteins that define them are distributed throughout this phylogeny (fig. S6), again suggesting sequential gains and losses of these genes along the phylogenetic tree just like the CRISPR-Cas systems.

Genomes with both CRISPR-Cas type I systems and specific membrane proteins show exclusive spacers and phage genes

Genomes with CRISPR-Cas type I systems show specific sets of membrane proteins. These proteins could provide the bacteria with important characteristics such as specific stress protection or detoxification that certain outer membrane proteins can offer. Because most CRISPR-Cas systems studied here seem to be oriented to phage protection as they seem to recognize phage sequences (Fig. 1F), we hypothesized that these membrane proteins could be acting as receptors or adhesion sites for specific phages. Thus, genomes that acquire these membrane proteins are forced to recruit CRISPR-Cas systems to defend against infection by these viruses while maintaining the beneficial functions of these proteins for the bacterium. The *A. baumannii* cluster 1 for type I-Fb includes the *ompA* gene, an outer membrane protein; the *P. aeruginosa* cluster 1 for type I-C shows the *fhuA* gene, and the two *A. baumannii* clusters for I-Fa show different variants of the *btuB* gene, both encoding TonB-dependent proteins. These three genes have long been known to act as phage receptors in *Escherichia coli* and *Salmonella* (*23*). In addition, we found genes involved in type IV pilus biogenesis, which are bacterial appendages that participate in different functions, such as cell adhesion, and are used by different phages to begin their infective cycle (*24*). Some of these genes were *pilX*, *pilW*, *pilE*, *pilV*, *pilA*, *fimU*, *fimT*, and *epsG*, present in *P. aeruginosa* strains with the CRISPR-Cas type I-F system. In particular, *pilA* is known to be essential for infection of phages DMS3, JDB26, and JDB68. When we searched for spacers against these viruses in *P. aeruginosa* strains with type I-F CRISPR-Cas systems, we found that 92% of strains having *pilA* have spacers against at least one of the above phages (table S4). In addition, spacers against phage JDB68, which is a specific phage that requires PilA to initiate an infection, were found in 28% of the genomes in the cluster that included *pilA* and other pilus genes (493 genomes; cluster 1 in Fig. 4). However, in the other cluster, which lacks *pilA*, only 8% of the genomes (91 genomes) presented spacers against this phage ($P = 6.94 \times 10^{-38}$). All this would support the known association between this membrane protein and the phage that makes use of it in its infective process. It should be mentioned that although PilA is the protein recognized in phage adsorption, the rest of the proteins that form the structure of the type IV pilus would also appear associated with the CRISPR-Cas system, as does PilA.

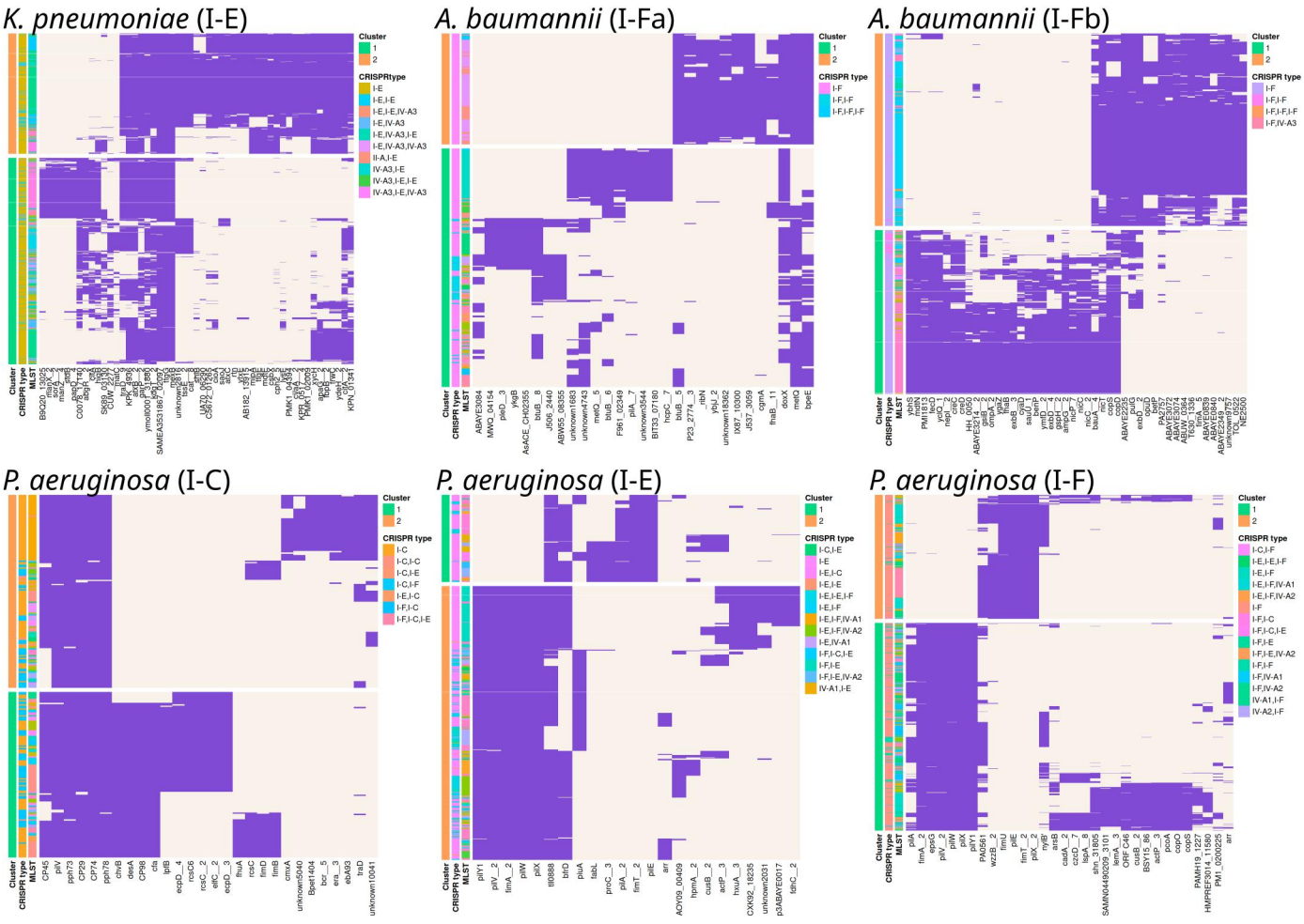

Fig. 4. Clusters of genomes with CRISPR-Cas type I systems according to the relevant genes encoding membrane proteins that they have. The purple color indicates the presence of the gene (*X* axis) in the corresponding genome (*Y* axis). On the left side of each plot, the cluster number is shown, along with the CRISPR-Cas type or the combination of them and the MLST group.

To test the above hypothesis with unknown pairs of membrane proteins and phages, we looked for spacers that recognized phage genes present in each of the two membrane protein–based clusters of genomes that did not appear (neither the spacer nor the phage gene) in the other cluster. Then, we also searched for genomes in each cluster that might carry the phage genes recognized by these spacers. Thus, within a given cluster, we could find genomes containing a cluster-specific spacer or a phage gene recognized by one of these specific spacers (Fig. 5A), but it could also be the case that both elements appear in the same genome, and this would imply having two alternatives: Either a phage gene and the cognate cluster-specific spacer appear together in the same genome, or a cluster-specific spacer and a phage gene recognized by a different cluster-specific spacer appear in the same genome. The different combinations of these elements were measured in each cluster, and two kinds of results were found. Clusters from type I showed both cluster-specific spacers and phage genes, whereas clusters from the other types of CRISPR-Cas systems showed almost no specific elements (except for cluster 1 of *P. aeruginosa* type IV-1, which showed a high proportion of genomes with cluster-specific phage genes). The CRISPR-Cas type I revealed cases such as cluster 2 of *P. aeruginosa* type I-F with three-quarters of the genomes showing phage-specific genes or cluster 1 of *K. pneumoniae* type I-E in which almost three-quarters of the genomes had spacers against cluster-specific phages. On the other hand, most genomes of the two I-F types of *A. baumannii* and cluster 2 of the types I-C and I-E of *P. aeruginosa* had both cluster-specific spacers and phages.

When evaluated in CRISPR-Cas type I systems, the number of genomes with unique spacers for each cluster was high, with a predominance of genomes with unique spacers in *K. pneumoniae* I-E and *P. aeruginosa* I-C, with spacers and phage genes in *A. baumannii* I-F types, and with phage genes in *P. aeruginosa* I-E and I-F (Fig. 5B).

At this point, we wanted to associate membrane proteins from each cluster with complete phages from the corresponding genomes. To do this, we took the type I CRISPR-Cas systems with the highest number of genomes: *A. baumannii* I-Fb, *K. pneumoniae* I-E, and *P. aeruginosa* I-F. From each of the two clusters, one membrane gene was taken and searched for complete phages that appeared in high frequency in the genomes of the cluster and not in the other cluster (Fig. 6). The representative membrane protein of each cluster was chosen as the one that appeared most

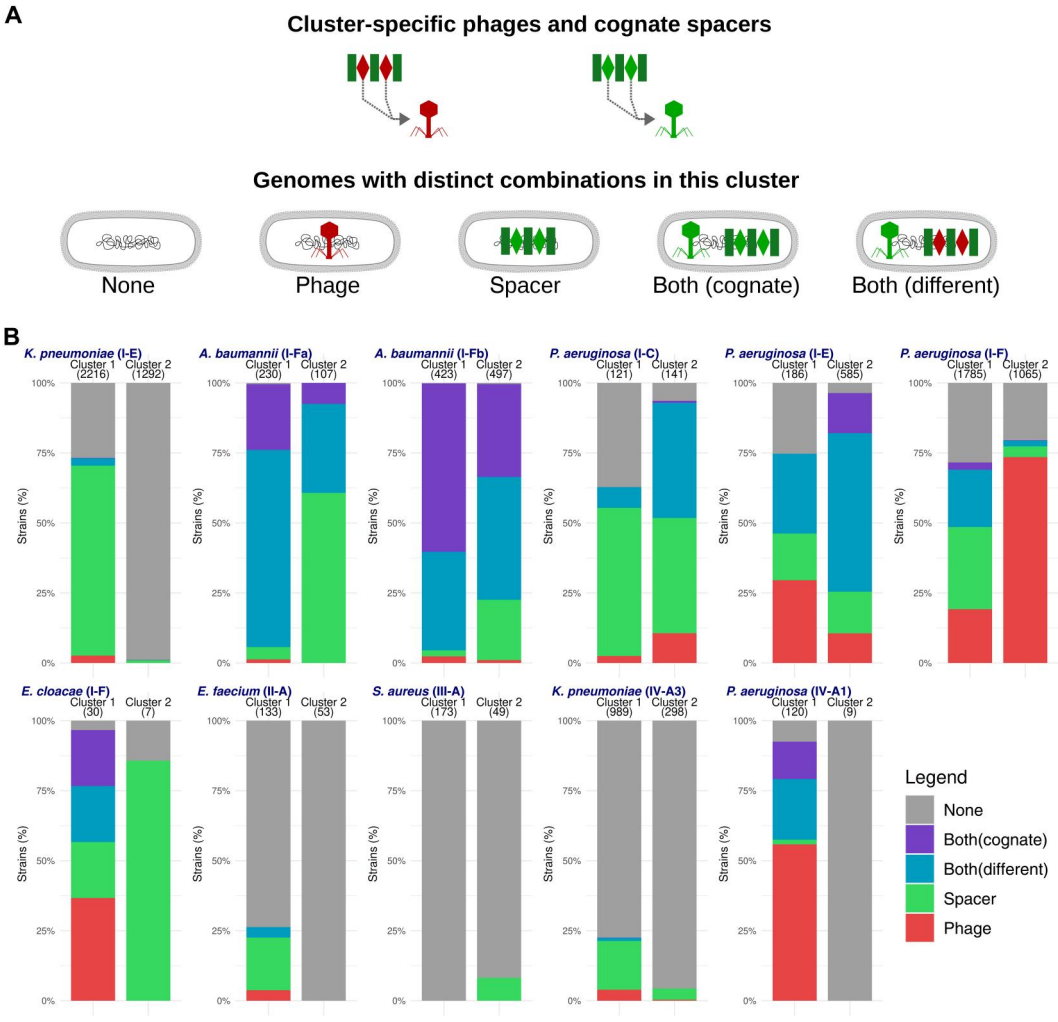

Fig. 5. Distribution of genomes in each membrane protein–based cluster with unique spacers and/or protospacers (phage genes). (**A**) Diagram representing two different cluster-specific pairs of phages and cognate spacers (top) and different combinations that can be found in a specific genome (bottom): neither cluster-specific spacers nor the phage genes are found (none), only phage genes are found (phage), only cluster-specific spacers are found (spacer), both phage genes and the cognate spacers are found (cognate), or both phage genes and spacers are found but the cluster-specific spacers are not the cognate for those phage genes. (**B**) The chart shows the proportion of genomes from each cluster that contains each combination of elements. The types of CRISPR-Cas systems that defined the best clusters are shown in the row above (in the same order as in Fig. 4), and the others are shown in the row below (in the same order as in fig. S5). The number of genomes in each cluster is shown in parentheses.

frequently in the cluster and had a lower frequency in both the other cluster and in genomes lacking CRISPR-Cas systems. Thus, we were able to find a series of phages that essentially only appear when the membrane gene is present. On the other hand, genomes presenting CRISPR-Cas systems did not normally contain the phages but did have spacers against those same phages. In addition, the cluster genomes that have the membrane gene appear with a much higher number of spacers against these phages, except in two of the six cases (cluster 2 of *K. pneumoniae* and *P. aeruginosa*), where, although the highest number of spacers appears in the other cluster, they also present spacers against the phages of the corresponding cluster.

The case of cluster 1 of *A. baumannii* stands out, in which 77 genomes appear with spacers against phages associated with the membrane protein AmpG__2, whereas no spacer is found in the cluster lacking this protein (table S4). On the other hand, cluster 2 of *A. baumannii* has specific spacers against two related phages. These phages appear in 250 genomes of this cluster (50% of them) and only in one genome without the CRISPR-Cas system. In addition, 467 genomes in this cluster show spacers against these phages (94% of them), while only 51 genomes of the other cluster have them (12% of them). Complete phages appear integrated near a tRNA-Val gene and have around 50 genes (fig. S7). In addition, the phages also appear in 249 genomes lacking CRISPR-Cas systems (3% of them). When the specific membrane proteins of cluster 2 are also searched for in genomes lacking the CRISPR-Cas system but having the phages, the best match occurs with the T630_1336 membrane protein (which we will refer to as *cam1* for CRISPR-associated membranome gene 1), while the rest of the membrane proteins appear more frequently in genomes lacking the phage gene (Fig. 7A). Specifically, 247 of the genomes that have *cam1* also have the phage gene (56%). Of the 249 genomes

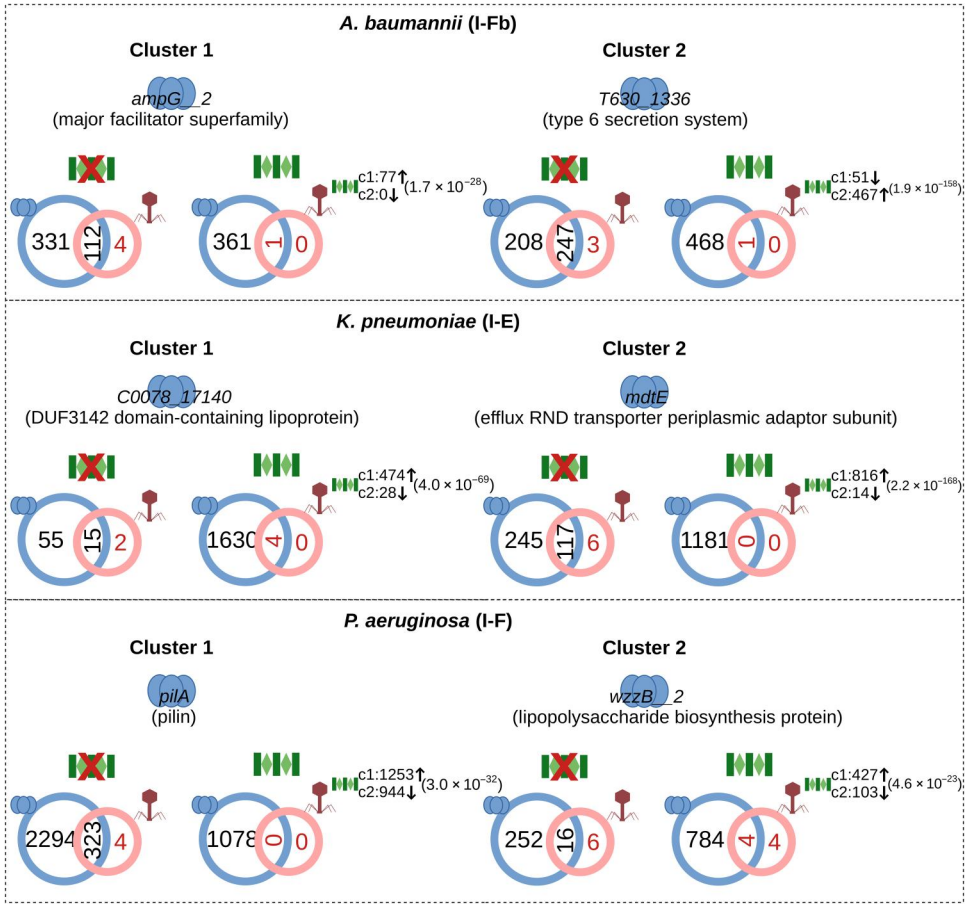

Fig. 6. Number of genomes with membrane genes and phages representative of each cluster. The two most frequent clusters of type I CRISPR-Cas systems from three species are shown. For each cluster, the name and description of the membrane gene is shown above. Below, the results for genomes lacking the CRISPR-Cas system (red cross) and those with the CRISPR-Cas system have been separated. The blue circle encompasses the number of genomes that have the membrane gene, and the red circle encompasses those that have the associated phage(s). The overlap of the two circles shows the number of genomes that have both elements. Numbers in red represent particularly low values of genomes with the phage. Last, the number of genomes containing spacers against the phage(s) in the two clusters is shown for strains that have the CRISPR-Cas system. The *P* value of the result is shown in parentheses, and the arrow indicates whether the value obtained is above or below the expected value.

without CRISPR-Cas systems that have the phages, the *cam1* gene could not be found in only 3 of them.

Cam1 is a 349–amino acid protein that shows a transmembrane region followed by a rearrangement hotspot repeat–associated core domain at its N-terminal half (InterPro: IPR022385). This domain appears in bacterial toxins involved in type VI secretion systems (T6SSs), but when present in proteins of less than 400 amino acids, found in bacteria such as *Pseudomonas putida*, it has an unknown function (*25*). The gene encoding this membrane protein is part of a cluster of six genes that is integrated next to a tRNA-Glu and the gene *fhuA*, and close to this region, there is another T6SS spike gene (*vgrG2__8*). The gene cluster also includes a *wapA* toxin gene, three genes encoding proteins with predicted signal peptides, and another gene encoding a probable membrane protein (*ABAYE2651*) not initially associated with the CRISPR-Cas system because of its slightly lower frequency in strains with this system relative to *cam1* (458 CRISPR-Cas type I-F strains have *ABAYE2651* versus 469 that have *cam1*).

Because the presence of this protein seems to be a sine qua non condition for the appearance of the phage, we could expect that the gain of this putative T6SS by the bacterium would imply that it would be exposed to infection by the phage, something that could be counteracted by the bacterium with the acquisition of a CRISPR-Cas system (Fig. 7B). This appears to be supported when analyzing the group of genomes containing this membrane protein, together with genomes that have a similar gene profile. There is a set of 65 genomes with a gene profile similar to cluster 2 CRISPR-Cas I-Fb genomes that lack the membrane protein, phage, and CRISPR-Cas systems (Fig. 7C). Then, the phylogeny of this group of genomes shows a first divergence supporting the gain of the membrane protein Cam1 that seems to allow phage entry. Later in the phylogeny, a new divergence event allows the gain of the CRISPR-Cas system that seems to prevent phage integration. While genomes with only the phage gene or CRISPR-Cas systems are rare (3 and 28 genomes, respectively), there are 247 genomes with both *cam1* and the phages and 468 with both *cam1* and the CRISPR-Cas system but not the phages (Fig. 7D), supporting the dependence

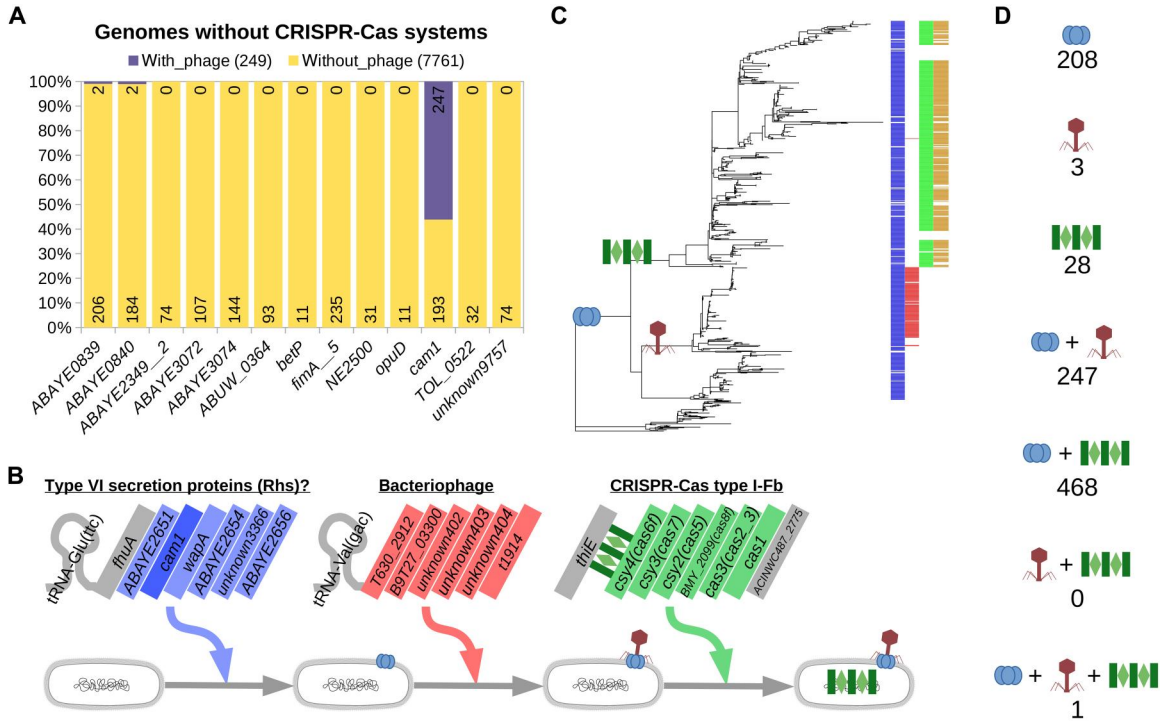

Fig. 7. Phages specific to cluster 2 of CRISPR-Cas I-Fb genomes and its co-occurrence with membrane genes. (**A**) Frequency of cluster 2–specific membrane proteins in genomes lacking CRISPR-Cas systems, separated between those with and without the phages related to this cluster. The number of genomes with and without the phages is shown in parentheses. Note that the phages appear in most genomes having the membrane protein Cam1. (**B**) Hypothesis of gain of gene groups from a genome lacking the membrane protein and the CRISPR-Cas system: first, the region of six genes involved in a T6SS entered, which would allow virus entry through the Cam1 protein, and, last, the CRISPR-Cas system would be obtained for protection against the phage. (**C**) Molecular phylogeny of 727 genomes of *A. baumannii* that belong to the membrane protein–based cluster 2 or have a similar gene profile to genomes in this cluster. Five hundred genes that appeared in all genomes were used. The metadata columns highlight the genomes with the membrane protein Cam1 (blue color), the phages (red color), the CRISPR-Cas type I-Fb (green color), and spacers against the phages (orange color). A genome of the reference MLST2 group was used to root the tree (strain XH727). The icons represent branches of the tree from which each of the three gene elements may have arisen. This tree would be consistent with the steps proposed in the hypothesis in (B). (**D**) The number of total genomes of *A. baumannii* having each possible combination of elements (Cam1, phages, and CRISPR-Cas I-Fb).

of phage on *cam1* and the dependence on CRISPR-Cas systems to prevent phage.

DISCUSSION

A limited number of bacterial genomes have CRISPR-Cas systems to prevent the entry of foreign DNA. We have analyzed tens of thousands of genomes of bacterial species of the ESKAPE group and found that type I is the most frequent CRISPR-Cas system in the Gram-negative species of this group. Our results are consistent with previous reports that found, for example, a low proportion of genomes with CRISPR-Cas systems in *E. faecium* (*26*) and about 50% of those in *P. aeruginosa* (*4*). When a bacterial genome has a CRISPR-Cas system, it is expected to have fewer genes. Some of the missing genes may be those involved in antibiotic resistance and originating from plasmids or integrative and conjugative elements or those involved in virulence and originating from phages. This was mostly confirmed in our study with ESKAPE pangenomes and coincides with previous studies with the same species, except for the fact that we show a more complete collection of type IV systems, and we have used twice as many genomes (*16*, *27*).

Bacteria have different defense systems against phages, but to study the relationship of these systems with their target sequences in silico, CRISPR-Cas systems have the advantage of having spacers, which reflect previous encounters with exogenous sequences (*2*). Thus, spacers of CRISPR-Cas systems usually recognize sequences originating from phages (*5*) and, to a lesser extent, from plasmids, such as we previously showed in *A. baumannii* (*13*). However, most spacers have unknown origin, as the corresponding protospacer cannot be found, and have been lumped together into what is known as CRISPR dark matter (*5*, *28*). This dark matter accounts for 80 to 90% of the spacers and is expected to recognize phage sequences that are still unknown or have diverged from known phage sequences. It is believed that this percentage may decrease with the future increase of sequences in the databases (*29*). We show that by analyzing complete pangenomes, which can include all the variation of phages infecting the species, dark matter can be greatly minimized, especially in the case of type I CRISPR-Cas systems (Fig. 1F). Thus, we were able to annotate 85% of the 9950 *P. aeruginosa* spacers and 72% of the 7345 *A. baumannii* spacers, with approximately 70% of these corresponding to phages or phage-plasmids. These phage-plasmids include phages that can remain as extrachromosomal elements in the bacterium (*17*) and against which we have found more associated spacers than against plasmid genes. In addition, we also found a small proportion of spacers that could recognize other endogenous bacterial genes,

such as flagellar biogenesis genes, which would suggest regulatory functions of CRISPR-Cas systems that remain to be elucidated (*18*, *30*). We have also found that genomes with CRISPR-Cas systems do not appear phylogenetically restricted, nor do they have a unique accessory genome. This suggests that bacteria acquire these systems when they provide an important evolutionary advantage. In a study carried out to measure the impact of CRISPR-Cas systems on horizontal gene transfer in bacteria, it was concluded that these systems would play an important role at the population level but not at the evolutionary scale (*31*). These systems are often recruited by mobile genetic elements on independent phylogenetic times (*32*). All of this would support our results, positioning CRISPR-Cas systems as functional modules that are acquired and discarded under certain circumstances.

We found dozens of genes that co-occur with CRISPR-Cas systems, although not necessarily close in the bacterial chromosome, many of which encode membrane proteins. The fact that both the CRISPR-Cas systems and these membrane proteins are scattered along the phylogenetic tree discards that their association is a consequence of a fortuitous and casual incorporation into a common ancestor, emphasizing the idea of a functional relationship. The association found in this work suggests that one such circumstance could be the defense against phages that use these proteins as receptors or adhesion sites. A previous report had already suggested that misfolded membrane proteins may trigger an envelope stress response that activates a CRISPR-Cas system (*33*), and other reports have found genes with probable association to the CRISPR-Cas systems, and some of them encoded integral membrane proteins (*34*, *35*). Many of these genes were related to type III systems, which is the type in *S. aureus*, where we found more than 20 genes annotated as integral component of membrane associated with its CRISPR-Cas system. However, we mainly found this association with membrane-related accessory genes among the different classes of type I CRISPR-Cas and propose that this may be related to the acquisition of beneficial functions for the bacterium that conversely make it more vulnerable to certain phages. These membrane proteins may help form biofilms or allow for certain virulence-related advantages (*6*–*12*), but at the same time, these membrane proteins can be receptors for specific phages. The fact that we did not find this association with type II and III CRISPR-Cas systems does not rule out that it may exist. However, the low number of genomes presenting these types of CRISPR-Cas systems in the species studied, as well as the low coverage in the annotation of their spacers in contrast to type I (Fig. 1F), could be preventing us from finding this association.

It has been shown that phages can increase the virulence of the bacterium that they infect when integrated into the bacterial chromosome, as they can carry toxins, resistance genes, or adhesion factors (*36*). CRISPR-Cas systems would prevent these phages from proliferating. However, we have found that, in many cases, spacers coexist with the cognate phage gene, especially in *A. baumannii*. These cases reflect that the phage would be integrated into the bacterial genome, suggesting that the immune system has not been fully efficient. In *P. aeruginosa*, and partly in *K. pneumoniae*, we have seen that the number of virulence genes in strains carrying CRISPR-Cas systems may be higher than in those without (fig. S2). The coexistence of the protospacer with the cognate spacer has been proposed as representing autoimmunity processes with a negative effect on the bacterium (*37*), although this may also be explained by the fact that the prophage is expressing anti-CRISPR systems (*38*). However, other studies have shown that CRISPR-Cas systems can prevent the lytic cycle of phages but tolerate the virus integration as a prophage, allowing the bacteria to co-opt the phage genes for possible use as virulence factors (*39*).

We observed that genomes with specific types of CRISPR-Cas systems, especially type I, carry a particular set of genes encoding membrane proteins. Furthermore, these genomes can be separated into clusters on the basis of the membrane proteins that they have, and the spacers of these CRISPR loci would recognize different nonoverlapping phage genes (Figs. 4 and 5). By analyzing these relationships between membrane proteins, spacers, and phages, we have found a number of membrane proteins associated with type I CRISPR-Cas systems that match, for example, proteins of the type IV pili of *P. aeruginosa*, which have long been known to constitute binding sites of certain phages (*24*), and genomes that have both the membrane protein and the CRISPR-Cas system have spacers against phages known to use the pilus in their infection process. In addition, we have been able to propose new pairs of membrane proteins and associated phages (Fig. 6), including a gene encoding a member of a putative T6SS as a possible receptor or adhesion site for a phage found in genomes with and without CRISPR-Cas systems. Phylogenetic data suggest that the gain of this immune system would protect the bacterium against this phage while allowing it to maintain the secretion system, which could be useful for intra- or interspecific competition (*40*). The gene cluster to which this membrane gene belongs is integrated next to the *fhuA* gene, and it is speculated that a class of specific receptors (TonB) could be critical for phage genome injection through their interaction with FhuA (*41*). Thus, this protein could help both the entry of the phage and the proper functioning of the T6SS system. Other reports have shown that secretion systems and CRISPR-Cas systems can depend on quorum sensing (*42*, *43*), which enables the coordination of bacterial population growth and is therefore also related to biofilm formation. Thus, it could be hypothesized that the bacterium would express both systems at the same time to avoid being exposed to phages that could take advantage of the activation of the T6SS system when the bacterial population and cell-to-cell contact is increased.

In summary, we demonstrate that the use of large pangenomes allows us to annotate a great part of the spacers of CRISPR-Cas systems, especially in type I, which will allow further research in this field. Here, we describe a "membranome-phage-CRISPR" triad, in which the CRISPR-Cas systems might be especially necessary when the bacterium expresses accessory genes that encode for membrane proteins. This would give it a special advantage in that situation, but it would also represent a gateway for phages that recognize these proteins as receptors or adhesion sites in their infection of bacteria.

MATERIALS AND METHODS

Genome collection and annotation

The assembled sequences of ESKAPE species available in the National Center for Biotechnology Information Genome database on 14 June 2021, including complete and draft genomes, were collected (*44*). Genomes and metadata were downloaded with the tools datasets 12.1.0 and dataformat 12.4.0 (a total of 68,352 genomes). Genomes with a low number of total genes or a low average number of shared genes (>5 times the interquartile range) were

removed on suspicion that they did not correspond to the species studied.

The protein-coding genes were predicted using Prokka version 1.14.5 (*45*), and the pangenome was created by Roary version 3.12.0 with an identity threshold of 90% and the -s parameter for not separating paralogs at this identity threshold (*46*). Protein sequences were functionally annotated using Sma3s v2 and the UniProt bacterial taxonomic division bacteria 2019_01 as the reference database (*47*). Gene names provided by Sma3s were preferentially assigned to each protein. When a gene name was repeated, a sequential number separated by two underscores was added. In cases where Sma3s did not assign a gene name, the one proposed by Prokka was taken, if available, preceded by an underscore. A gene was classified as a core gene if it appeared in ≥99% of the genomes of the species.

CRISPR-Cas systems and their specific types were assigned using CRISPRCasTyper 1.4.1 (*48*). Types I-Fa and I-Fb of *A. baumannii* were distinguished by looking for their different integration site. To discover the spacers of CRISPR-Cas systems, CRISPRCasFinder 4.2.20 was used with default parameters (*49*). Only CRISPR arrays with an evidence level equal to 4 were considered. Identical spacers were collapsed together, taking into account both chains. The number of sequences of each type for each species is available in table S1.

Search for specific gene groups in the pangenomes

Antibiotic resistance genes were found by AMRFinderPlus 3.10.1 using the databases of the six bacterial species analyzed (*50*). Virulence genes were found by performing a similarity search with BLASTP 2.9.0+ (*51*) against the VFDB database version December 2020 (Virulence Factor Database), requiring at least 90% sequence identity and 90% database sequence coverage (*52*).

Genes encoding membrane proteins were searched for in the functional annotation performed by Sma3s, to which genes encoding outer membrane proteins were specifically added by a BLASTP similarity search against the OMPdb release 2021, requiring at least 90% sequence identity and 90% database sequence coverage (*53*).

Genes from plasmids were searched using the annotation "Plasmid" in the UniProt keyword field. Then, genes with ≥90% sequence identity and ≥90% query coverage with a sequence of the PLSDB database v2020_06_23_v2 were added (*54*). Viral genes were searched following the same protocol but using the IMG/VR database v3 (IMG_VR_2020-10-12_5.1) (*55*) and viral genes from the functional annotation with Sma3s. Genes that appeared between two genes annotated as viral were also added.

Search for prophages

Complete prophages were searched in all the genomes using Phigaro version 2.3.0 with default parameters and *abs* mode (*56*). The prophages of each species were then grouped by similarity using the MeShClust version 3.0 program with the parameters -v (total initial sequences) and -b (1/4 of total initial sequences). (*57*). Last, the names of known phages were obtained from the literature, and their nucleotide sequence was downloaded from the GenBank database: DMS3 (NC_008717.1), JBD26 (NC_061435.1), and JBD68 (KY707339.1) (*24*).

To combine similar phage genomes with the sequence oriented in the two possible reading directions, we performed a similarity search with BLASTN by comparing the reference sequences of each cluster with each other. Then, clusters whose reference sequence shared at least 90% identity and coverage were combined.

Search for protospacers

Protospacers, putative genes recognized by spacers, were searched by performing a similarity search with BLASTN and the blastn-short option turned on, using a threshold of ≥95% sequence identity and 100% spacer coverage. Those protospacers that were also found following the same strategy but using sequences from CRISPR repeats instead of the spacers were discarded to avoid misannotated sequences (*58*). To calculate the *P* value of the number of found matches, a hypergeometric test was used (dhyper function in R).

Search for genes associated with CRISPR-Cas types with inference on random forests

We used inference of random forests in multiple iterations to search for genes associated with specific CRISPR-Cas types. For each species, we compiled one dataset containing all strains that have no CRISPR-Cas systems, and other datasets with the strains containing each CRISPR-Cas type present in the data, respectively. From 20 iterations with different random seeds, the most important features were selected and counted. In this context the features are binary indicators of the presence of the genes for each strain. After all iterations, the count of a gene can indicate how often the random forest deemed it important for the difference between the respective CRISPR-containing and CRISPR-deficient genomes. *Cas* genes were removed from consideration to be able to focus on genes not directly related to CRISPR-Cas systems. Genes identified as important features in multiple iterations were considered to be associated with CRISPR-Cas systems, when they were more abundant in CRISPR-Cas–containing genomes than in non–CRISPR-Cas–containing genomes.

The random forest implementation was done in Python with the scikit-learn package 1.0.2 (*59*). With each iteration, random parts of the datasets were divided into train and test sets with the ratio of 0.8 to 0.2. For all six species and their CRISPR-Cas systems, the random forests achieved average accuracies higher than 0.93 over all iterations. The trained random forest object has the resulting feature importance by the mean decrease in impurity available as a parameter. The permutation feature importance may be more informative for high cardinality features, but because we only have two values for each feature, the mean decrease in impurity feature importance measurement is sufficient. The default parameters of the random forest were used. The location of the code is given below. When the median has been used, the deviation value is calculated from the median absolute deviation.

Functional enrichment analysis

To discover the functional enrichment of genes associated with specific CRISPR-Cas types, we used the R package TopGO version 2.40.0 (*60*), which uses Gene Ontology (GO) terms from a specific ontology. The GO terms used were those annotated by Sma3s. Figures were created using the R ggplot2 library in a custom script.

MLST assignment, gene profiles, and molecular phylogenies

MLST numbers were assigned to each genome by compiling the genes used in the species-specific schemes in PubMLST 23 Nov

2021 (*19*) and searching them in the genome sequences using the mlst program (https://github.com/tseemann/mlst). The MLST number assigned to each genome, along with the CRISPR-Cas systems it has is available in table S2.

MLST phylogenetic trees were constructed using the MLST sequences. Nucleotide sequences were aligned with mafft v7.271 using the G-INS-I option (*61*). The phylogeny was constructed with RAxML v8.2.9 with the GTRCAT model and bootstrap of 1000 (*62*). The model was selected with ModelFinder implemented in IQ-TREE (*63*). The phylogeny in Fig. 7D was constructed with the same protocol but using all genomes, the PROTGAMMAWAG model, and 500 core proteins.

The gene profiles for each species were constructed using a binary representation of the bacterial genome, where a gene is either absent or present in a strain, without accounting for the number of paralogs. These data are condensed to MLST level by assigning 0 or 1 to the gene in the MLST group by majority vote of all strains in the group. The MLST groups are then subjected to a pairwise Jaccard distance measurement, resulting in an $N \times N$ matrix of Jaccard distances between MLST groups, with N equal to the number of MLST groups for the respective species dataset. The pairwise Jaccard distances were computed with scikit-learn.

These pairwise distances were used to construct a profile of genetic distances between the MLST groups for each species. We used ward linkage and descending distance sort for the hierarchal clustering and the dendrogram. Dendrograms were produced using SciPy 1.6.2 (*64*), and correlation plots were plotted with seaborn 0.11.2 and Matplotlib 3.5.0 (*65*, *66*).

Cluster analysis

Heatmaps to compare genomes presenting different combinations of membrane proteins were performed with the R library pheatmap 1.0.12, and clusters were created with the cutree function.

Supplementary Materials

This PDF file includes:
Figs. S1 to S7
Legends for tables S1 to S4

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S4

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. R. M. Dedrick, B. E. Smith, M. Cristinziano, K. G. Freeman, D. Jacobs-Sera, Y. Belessis, A. Whitney Brown, K. A. Cohen, R. M. Davidson, D. van Duin, A. Gainey, C. B. Garcia, C. R. Robert George, G. Haidar, W. Ip, J. Iredell, A. Khatami, J. S. Little, K. Malmivaara, B. J. McMullan, D. E. Michalik, A. Moscatelli, J. A. Nick, M. G. Tupayachi Ortiz, H. M. Polenakovik, P. D. Robinson, M. Skurnik, D. A. Solomon, J. Soothill, H. Spencer, P. Wark, A. Worth, R. T. Schooley, C. A. Benson, G. F. Hatfull, Phage therapy of *Mycobacterium* infections: Compassionate-use of phages in twenty patients with drug-resistant mycobacterial disease. *Clin. Infect. Dis.* **76**, 103–112 (2022).
2. F. Tesson, A. Hervé, E. Mordret, M. Touchon, C. d'Humières, J. Cury, A. Bernheim, Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.* **13**, 2561 (2022).
3. L. A. Marraffini, E. J. Sontheimer, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.* **11**, 181–190 (2010).
4. R. M. Wheatley, R. C. MacLean, CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa. *ISME J.* **15**, 1420–1433 (2021).
5. S. A. Shmakov, V. Sitnik, K. S. Makarova, Y. I. Wolf, K. V. Severinov, E. V. Koonin, The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *MBio* **8**, e01397–17 (2017).
6. G. E. Heussler, K. C. Cady, K. Koeppen, S. Bhuju, B. A. Stanton, G. A. O'Toole, Clustered regularly interspaced short palindromic repeat-dependent, biofilm-specific death of *Pseudomonas aeruginosa* mediated by increased expression of phage-related genes. *mBio* **6**, e00129–15 (2015).
7. B. Tang, T. Gong, X. Zhou, M. Lu, J. Zeng, X. Peng, S. Wang, Y. Li, Deletion of cas3 gene in Streptococcus mutans affects biofilm formation and increases fluoride sensitivity. *Arch. Oral Biol.* **99**, 190–197 (2019).
8. K. C. Cady, G. A. O'Toole, Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. *J. Bacteriol.* **193**, 3433–3445 (2011).
9. L. Medina-Aparicio, S. Rodriguez-Gutierrez, J. E. Rebollar-Flores, Á. G. Martínez-Batallar, B. D. Mendoza-Mejía, E. D. Aguirre-Partida, A. Vázquez, S. Encarnación, E. Calva, I. Hernández-Lucas, The CRISPR-Cas system is involved in OmpR genetic regulation for outer membrane protein synthesis in *Salmonella* Typhi. *Front. Microbiol.* **12**, 657404 (2021).
10. T. R. Sampson, S. D. Saroj, A. C. Llewellyn, Y.-L. Tzeng, D. S. Weiss, A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**, 254–257 (2013).
11. M. A. B. Shabbir, Y. Tang, Z. Xu, M. Lin, G. Cheng, M. Dai, X. Wang, Z. Liu, Z. Yuan, H. Hao, The involvement of the *Cas9* gene in virulence of *Campylobacter jejuni*. *Front. Cell. Infect. Microbiol.* **8**, 285 (2018).
12. J. Solbiati, A. Duran-Pinedo, F. Godoy Rocha, F. C. Gibson, J. Frias-Lopez, Virulence of the pathogen *Porphyromonas gingivalis* is controlled by the CRISPR-Cas protein Cas3. *mSystems* **5**, e00852–20 (2020).
13. E. L. Mangas, A. Rubio, R. Álvarez-Marín, G. Labrador-Herrera, J. Pachón, M. E. Pachón-Ibáñez, F. Divina, A. J. Pérez-Pulido, Pangenome of *Acinetobacter baumannii* uncovers two groups of genomes, one of them with genes involved in CRISPR/Cas defence systems associated with the absence of plasmids and exclusive genes for biofilm formation. *Microb. Genom.* **5**, e000309 (2019).
14. Y. Kim, C. Gu, H. U. Kim, S. Y. Lee, Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotechnol.* **63**, 54–62 (2020).
15. R. J. Hall, F. J. Whelan, J. O. McInerney, Y. Ou, M. R. Domingo-Sananes, Horizontal gene transfer as a source of conflict and cooperation in prokaryotes. *Front. Microbiol.* **11**, 1569 (2020).
16. K. Mortensen, T. J. Lam, Y. Ye, Comparison of CRISPR-Cas immune systems in healthcare-related pathogens. *Front. Microbiol.* **12**, 758782 (2021).
17. E. Pfeifer, J. A. Moura de Sousa, M. Touchon, E. P. C. Rocha, Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).
18. N. Sharma, A. Das, P. Raja, S. A. Marathe, The CRISPR-Cas system differentially regulates surface-attached and pellicle biofilm in Salmonella enterica serovar typhimurium. *Microbiol. Spectr.* **10**, e0020222 (2022).
19. K. A. Jolley, J. E. Bray, M. C. J. Maiden, Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).
20. A. Moya-Beltrán, K. S. Makarova, L. G. Acuña, Y. I. Wolf, P. C. Covarrubias, S. A. Shmakov, C. Silva, I. Tolstoy, D. B. Johnson, E. V. Koonin, R. Quatrini, Evolution of type IV CRISPR-Cas systems: Insights from CRISPR loci in integrative conjugative elements of *Acidithiobacillia*. *CRISPR J.* **4**, 656–672 (2021).
21. J. Breisch, B. Averhoff, Identification of osmo-dependent and osmo-independent betaine-choline-carnitine transporters in Acinetobacter baumannii: Role in osmostress protection and metabolic adaptation. *Environ. Microbiol.* **22**, 2724–2735 (2020).
22. L. Moynié, I. Serra, M. A. Scorciapino, E. Oueis, M. G. Page, M. Ceccarelli, J. H. Naismith, Preacinetobactin not acinetobactin is essential for iron uptake by the BauA transporter of the pathogen *Acinetobacter baumannii*. *eLife* **7**, e42270 (2018).
23. J. Bertozzi Silva, Z. Storms, D. Sauvageau, Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**, fnw002 (2016).
24. H. Harvey, J. Bondy-Denomy, H. Marquis, K. M. Sztanko, A. R. Davidson, L. L. Burrows, Pseudomonas aeruginosa defends against phages through type IV pilus glycosylation. *Nat. Microbiol.* **3**, 47–52 (2018).
25. P. Bernal, L. P. Allsopp, A. Filloux, M. A. Llamas, The Pseudomonas putida T6SS is a plant warden against phytopathogens. *ISME J.* **11**, 972–987 (2017).
26. K. D. Mlaga, V. Garcia, P. Colson, R. Ruimy, J.-M. Rolain, S. M. Diene, Extensive comparative genomic analysis of *Enterococcus faecalis* and *Enterococcus faecium* reveals a direct association between the absence of CRISPR-Cas Systems, the presence of anti-endonuclease

(ardA) and the acquisition of vancomycin resistance in *E. faecium*. *Microorganisms* **9**, 1118 (2021).
27. E. Pursey, T. Dimitriu, F. L. Paganelli, E. R. Westra, S. van Houte, CRISPR-Cas is associated with fewer antibiotic resistance genes in bacterial pathogens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200464 (2022).
28. J. McGinn, L. A. Marraffini, Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat. Rev. Microbiol.* **17**, 7–12 (2019).
29. S. A. Shmakov, Y. I. Wolf, E. Savitskaya, K. V. Severinov, E. V. Koonin, Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun. Biol.* **3**, 321 (2020).
30. B. Bozic, J. Repac, M. Djordjevic, Endogenous gene regulation as a predicted main function of type I-E CRISPR/Cas system in *E. coli*. *Molecules* **24**, 784 (2019).
31. U. Gophna, D. M. Kristensen, Y. I. Wolf, O. Popa, C. Drevet, E. V. Koonin, No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J.* **9**, 2021–2027 (2015).
32. G. Faure, S. A. Shmakov, W. X. Yan, D. R. Cheng, D. A. Scott, J. E. Peters, K. S. Makarova, E. V. Koonin, CRISPR-Cas in mobile genetic elements: Counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
33. R. Perez-Rodriguez, C. Haitjema, Q. Huang, K. H. Nam, S. Bernardis, A. Ke, M. P. DeLisa, Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol. Microbiol.* **79**, 584–599 (2011).
34. S. A. Shmakov, K. S. Makarova, Y. I. Wolf, K. V. Severinov, E. V. Koonin, Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E5307–E5316 (2018).
35. S. A. Shah, O. S. Alkhnbashi, J. Behler, W. Han, Q. She, W. R. Hess, R. A. Garrett, R. Backofen, Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-*cas* gene cassettes reveals 39 new *cas* gene families. *RNA Biol.* **16**, 530–542 (2019).
36. L.-C. Fortier, O. Sekulovic, Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence* **4**, 354–365 (2013).
37. A. Stern, L. Keren, O. Wurtzel, G. Amitai, R. Sorek, Self-targeting by CRISPR: Gene regulation or autoimmunity? *Trends Genet.* **26**, 335–340 (2010).
38. S. Govindarajan, A. Borges, S. Karambelkar, J. Bondy-Denomy, Distinct subcellular localization of a type I CRISPR complex and the Cas3 nuclease in bacteria. *J. Bacteriol.* **204**, e0010522 (2022).
39. G. W. Goldberg, W. Jiang, D. Bikard, L. A. Marraffini, Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* **514**, 633–637 (2014).
40. N.-H. Le, V. Pinedo, J. Lopez, F. Cava, M. F. Feldman, Killing of gram-negative and gram-positive bacteria by a bifunctional cell wall-targeting T6SS effector. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2106555118 (2021).
41. L. Letellier, P. Boulanger, L. Plançon, P. Jacquot, M. Santamaria, Main features on tailed phage, host recognition and DNA uptake. *Front. Biosci.* **9**, 1228–1339 (2004).
42. L. Cui, X. Wang, D. Huang, Y. Zhao, J. Feng, Q. Lu, Q. Pu, Y. Wang, G. Cheng, M. Wu, M. Dai, CRISPR-*cas3* of *Salmonella* upregulates bacterial biofilm formation and virulence to host cells by targeting quorum-sensing systems. *Pathogens* **9**, E53 (2020).
43. A. D. Maharajan, E. Hjerde, H. Hansen, N. P. Willassen, Quorum sensing controls the CRISPR and type VI secretion systems in *Aliivibrio wodanis* 06/09/139. *Front. Vet. Sci.* **9**, 799414 (2022).
44. P. A. Kitts, D. M. Church, F. Thibaud-Nissen, J. Choi, V. Hem, V. Sapojnikov, R. G. Smith, T. Tatusova, C. Xiang, A. Zherikov, M. DiCuccio, T. D. Murphy, K. D. Pruitt, A. Kimchi, Assembly: A resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–D80 (2016).
45. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
46. A. J. Page, C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. G. Holden, M. Fookes, D. Falush, J. A. Keane, J. Parkhill, Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
47. C. S. Casimiro-Soriguer, A. Muñoz-Mérida, A. J. Pérez-Pulido, Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. *Proteomics* **17**, 1700071 (2017).
48. J. Russel, R. Pinilla-Redondo, D. Mayo-Muñoz, S. A. Shah, S. J. Sørensen, CRISPRCasTyper: Automated identification, annotation, and classification of CRISPR-Cas loci. *CRISPR J.* **3**, 462–469 (2020).
49. D. Couvin, A. Bernheim, C. Toffano-Nioche, M. Touchon, J. Michalik, B. Néron, E. P. C. Rocha, G. Vergnaud, D. Gautheret, C. Pourcel, CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
50. M. Feldgarden, V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, M. Hoffmann, J. B. Pettengill, A. B. Prasad, G. E. Tillman, G. H. Tyson, W. Klimke, AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12728 (2021).
51. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
52. B. Liu, D. Zheng, Q. Jin, L. Chen, J. Yang, VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692 (2019).
53. K. D. Tsirigos, P. G. Bagos, S. J. Hamodrakas, OMPdb: A database of β-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res.* **39**, D324–D331 (2011).
54. V. Galata, T. Fehlmann, C. Backes, A. Keller, PLSDB: A resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).
55. S. Roux, D. Páez-Espino, I.-M. A. Chen, K. Palaniappan, A. Ratner, K. Chu, T. B. K. Reddy, S. Nayfach, F. Schulz, L. Call, R. Y. Neches, T. Woyke, N. N. Ivanova, E. A. Eloe-Fadrosh, N. C. Kyrpides, IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
56. E. V. Starikova, P. O. Tikhonova, N. A. Prianichnikov, C. M. Rands, E. M. Zdobnov, E. N. Ilina, V. M. Govorun, Phigaro: High-throughput prophage sequence annotation. *Bioinformatics* **36**, 3882–3884 (2020).
57. B. T. James, B. B. Luczak, H. Z. Girgis, MeShClust: An intelligent tool for clustering DNA sequences. *Nucleic Acids Res.* **46**, e83 (2018).
58. A. Rubio, P. Mier, M. A. Andrade-Navarro, A. Garzón, J. Jiménez, A. J. Pérez-Pulido, CRISPR sequences are sometimes erroneously translated and can contaminate public databases with spurious proteins containing spaced repeats. *Database* **2020**, baaa088 (2020).
59. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
60. A. Alexa, J. Rahnenführer, T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
61. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
62. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
63. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
64. P. Peterson, F2PY: A tool for connecting Fortran and Python programs. *Int. J. Comput. Sci. Eng.* **4**, 296–305 (2009).
65. M. L. Waskom, seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
66. J. D. Hunter, Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgments: We thank C3UPO and its admin Cristina Moral, and the HPC group of the JGU for the HPC support. **Funding:** This work was supported by MCIN/AEI/ PID2020-114861GB-I00 (Agencia Estatal de Investigación/Ministry of Science and Innovation of the Spanish Government) and by the European Regional Development Fund and the Consejeria de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucia (PY20_00871). **Author contributions:** Conceptualization: A.J.P.-P., M.A.A.-N., and A.G. Data curation: A.J.P.-P., A.M.-R., M.S., and A.R. Formal analysis: A.J.P.-P., A.M.-R., M.S., and A.R. Funding acquisition: A.J.P.-P. Investigation: A.J.P.-P., A.G., M.S., and A.R. Methodology: A.J.P.-P., M.A.A.-N., A.M.-R., M.S., and A.R. Project administration: A.J.P.-P. Resources: A.J.P.-P., M.A.A.-N., J.P., and M.E.P.-I. Software: A.J.P.-P., M.S., and A.R. Supervision: A.J.P.-P. and M.A.A.-N. Validation: A.J.P.-P., M.A.A.-N., J.P., M.E.P.-I., A.G., M.S., and A.R. Visualization: A.J.P.-P., M.S., and A.R. Writing (original draft): A.J.P.-P. Writing (review and editing): A.J.P.-P., M.A.A.-N., J.P., M.E.P.-I., A.G., M.S., and A.R. **Competing interests:** The authors declare they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The code used to analyze the data, the pangenomes, and the phylogenetic distances are available in the repository Zenodo: https://zenodo.org/record/7224593#.Y0-_KExBxPZ.

Submitted 11 July 2022
Accepted 17 February 2023
Published 24 March 2023
10.1126/sciadv.add8911