Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis

Reza Rezaei Javan[1], Elisa Ramos-Sevillano[2], Asma Akter[3], Jeremy Brown[2] and Angela B Brueggemann[1,3]

[1]Nuffield Department of Medicine, University of Oxford
[2]UCL Respiratory, Division of Medicine, University College London
[3]Department of Medicine, Imperial College London

Corresponding author: angela.brueggemann@ndm.ox.ac.uk

Abstract

Prophages (viral genomes integrated within a host bacterial genome) are abundant within the bacterial world and are of interest because they often confer various phenotypic traits to their hosts, such as by encoding genes that increase pathogenicity. Satellite prophages are 'parasites of parasites' that rely on the bacterial host and another helper prophage for survival. We analysed >1,300 genomes of 70 different *Streptococcus* species for evidence of prophages and identified nearly 800 prophages and satellite prophages, the majority of which are reported here for the first time. We show that prophages and satellite prophages were widely distributed among streptococci, were two clearly different entities and each possessed a structured population. There was convincing evidence that cross-species transmission of prophages is not uncommon. Furthermore, *Streptococcus pneumoniae* (pneumococcus) is a leading human pathogen worldwide, but the genetic basis for its pathogenicity and virulence is not yet fully understood. Here we report that over one-third of pneumococcal genomes possessed satellite prophages and demonstrate for the first time that a satellite prophage was associated with virulence in a murine model of infection. Overall, our findings demonstrate that prophages are widespread components of *Streptococcus* species and suggest that they play a role in pneumococcal pathogenesis.

Main

The genus *Streptococcus* comprises a wide variety of pathogens responsible for causing significant morbidity and mortality worldwide[1]. *Streptococcus pneumoniae* (pneumococcus) is a leading cause of pneumonia, bacteraemia, and meningitis[2]. *Streptococcus pyogenes* (group A streptococci) is a leading cause of pharyngitis, scarlet fever and necrotising fasciitis[3]. *Streptococcus agalactiae* (group B streptococci) is the most common cause of neonatal sepsis[4]. *Streptococcus suis* and *Streptococcus equi* rarely cause disease in humans but are important animal pathogens[1].

Bacteriophages (phages) are intracellular parasites of bacteria. Lytic phages hijack the host bacterial machinery, produce new phages and destroy the infected bacterial cell. Lysogenic phages do not necessarily initiate replication immediately upon host entry and may integrate their genome within the bacterial genome to be activated at a later stage. An integrated phage is termed a prophage and those genes can be passed down to the bacterial daughter cells.

Since survival depends on their bacterial hosts, prophages often express genes that increase host cell fitness[5,6]. Prophages can exert a range of phenotypic effects on the host bacteria: encode toxins that increase virulence[5], promote binding to human platelets[7] or cells[8], evade immune defences[9,10], or protect

from oxidative stress[11]. Prophage integration can also regulate bacterial populations by altering bacterial gene expression[12,13].

Prophages and their hosts, like other predator and prey relationships, are embroiled in a complex evolutionary arms race whereby bacteria evolve various strategies to defend themselves and prophages co-evolve to overcome these barriers[14]. These coevolutionary dynamics are complicated by satellite prophages, which lack all the necessary genetic information to replicate on their own and are reliant on hijacking the machinery of another inducing 'helper' prophage to replicate. Satellite prophages might be thought of as 'parasites of parasites'[15,16].

Satellite prophages adversely interfere with helper prophage replication and thus promote bacterial survival[17-19]. Satellite prophages have been discovered through different circumstances and thus there are different terms used to describe this particular type of mobile genetic element (MGE) in the literature, including *Staphylococcus aureus* pathogenicity islands (SaPIs), phage-related chromosomal islands (PRCIs) and phage-inducible chromosomal islands (PICIs), among others[17-23].

Satellite prophages have been shown to be vectors for the spreading of toxin genes and other virulence factors, e.g. SaPI1, which possesses the gene responsible for causing toxic shock syndrome[24]. The prevalence, diversity, genetic stability and molecular epidemiology of satellite prophages in streptococcal species are largely unknown. A small number of satellite prophages have been identified in streptococcal species, although whether they are associated with virulence remains to be investigated[25]. Previous work has shown that prophage-related sequences are highly prevalent within *S. pneumoniae*[26-28], *S. pyogenes*[29,30] and *S. agalactiae* genomes[31]; however, genus-wide analyses of the genomic diversity and population structure of streptococcal prophages have not yet been reported.

Here we report the discovery of nearly 800 prophages among >1,300 streptococcal genomes and provide detailed insights into prophage genomics and population structure. Using *S. pneumoniae* as the model organism, the molecular epidemiology of satellite prophages was investigated within a large globally-distributed collection of pneumococci isolated over a 90-year period. Finally, we demonstrated that a satellite prophage was associated with virulence in a murine infection model.

Results

Prophage sequences are a significant component of the genomes of clinically-relevant *Streptococcus* species

We analysed 1,306 genomes from 70 different streptococcal species and identified 419 full-length prophages and 354 satellite prophage genomes (Supplementary Table 1). We estimated the prophage gene content within each streptococcal genome and this revealed a substantial difference in the average prophage content among various streptococcal species, ranging from 0.4% of the *Streptococcus thermophilus* genome to 9.5% of the *S. pyogenes* genome (Figure 1a; Supplementary Table 2). Furthermore, we observed significant variability in prophage content among different genomes of the same bacterial species, e.g. full-length prophages comprised up to 19% of the genes in some *S. pyogenes* genomes, while in others they made up <1% of the genome (Figure 1a). The prevalence of satellite prophages ranged from 0.1% among *Streptococcus mutans* and *Streptococcus sanguinis* genomes to 4.5% of the *Streptococcus dysgalactiae* genomes (Figure 1a).

Full-length and satellite prophages are separate entities with little effective genetic exchange between them

Satellite prophages had a lower guanine (G) and cytosine (C) content than full-length prophages and were about a third of the size in terms of both length of sequence and the number of genes they harboured (Figure 1b). Due to their relatively small genome and apparent lack of essential genes, streptococcal satellite prophage sequences have historically often been regarded as "remnant" or "defective" prophages in a state of mutational decay[13,22,32-34]. Our data reveal that satellite prophage sequences can be highly conserved over many decades, e.g. one satellite prophage was present among pneumococcal genomes with isolation dates ranging from 1939 to 2006 and had maintained >99.98% nucleotide similarity across its entire genome (Figure 1c), suggesting that it is under very strong evolutionary pressure and likely provides an important biological function.

An unrooted phylogenetic tree of all streptococcal prophage genomes in our dataset depicted full-length and satellite prophages as two clearly distinct groups (Figure 1d). We observed that the genes of satellite prophages are unique and differ to those of full-length prophages, as nearly 99% of all satellite prophage genes (>70% amino acid sequence similarity) are not found in any full-length prophages (Figure 1e). Taken together, these findings confirm that satellite prophage sequences are not recent remnants of previous lysogenisation by full-length prophages, but rather that they belong to a unique family of mobile genetic elements.

Streptococcal prophages have a structured population

We found that both full-length and satellite streptococcal prophages demonstrated well-conserved patterns in genome organisation and synteny, regardless of the species that they were isolated from (Figure 2a). Similar to other non-streptococcal prophages (Supplementary Figure 1), genes encoding specific functions were often found clustered together in the prophage genome, although note that the function of many genes is still unknown and therefore the delineation of discrete gene clusters remains problematic (Figure 2a). Whole genome comparisons of all prophage sequences in our dataset depicted several major and minor clusters for both full-length and satellite prophages (Supplementary Figure 2).

Phages are generally believed to be bacterial species-specific and even specific to genetic lineages within a single bacterial species[35]. Surprisingly, we often found prophages from different bacterial species within the same phylogenetic cluster, suggesting that cross-species transmissions are more common among streptococcal prophages than previously realised. Remarkably, despite the relatedness of their prophages, the bacterial hosts were not necessarily the closest phylogenetically-related species (Figure 2b). One possible explanation could be that streptococcal prophages are evolving separately from their microbial hosts, and therefore, other factors such as ecological relatedness may dominate over evolutionary relatedness of the host bacteria.

Molecular epidemiology of satellite prophages within a global pneumococcal dataset dating from 1916

We had previously determined the prevalence, diversity and molecular epidemiology of full-length prophages in a global and historical pneumococcal genome dataset[26]. Many shorter prophage sequences were also identified in that study, which were simply classified as partial prophage sequences and not characterised further at the time. Here, we used this genome dataset to further investigate satellite prophages in the context of the pneumococcal population structure. The genome collection was comprised of 482 pneumococci recovered from both healthy and diseased individuals between 1916 and 2009. Pneumococci were isolated from people of all ages residing in 36 different countries. Ninety-one serotypes and 94 different clonal complexes (genetic lineages) were represented in the dataset.

A reinvestigation of the 'partial prophage' sequences resulted in the identification of 44 representative pneumococcal satellite prophages, which clustered into five major groups (Figure 3a). The average GC content of the satellite prophages was lower than their pneumococcal host but varied among each group (Figure 3b). We found that 35% of the pneumococci in our dataset contained at least one satellite prophage and 5% of the genomes contained two. Some satellite prophages were present in up to six different clonal

complexes, whereas others were only found in Singletons (genotypes with no closely related variants; Table 1 and Supplementary Figure 3). Those satellite prophages identified in more than one genome were often found among pneumococci recovered over a decade or more (Table 1). The average prophage content for each of the major clonal complexes ranged from 2.2-6.5%, and with only one exception (CC7232), all of these are widely-circulating pneumococcal genetic lineages (Figure 1c; https://pubmlst.org/spneumoniae).

Prophages are more frequently inserted adjacent to genes involved in information storage and processing

We previously reported that pneumococcal full-length prophages were consistently integrated in specific locations within the genome[26]. Likewise, pneumococcal satellite prophages were consistently integrated in seven precise locations (a-f) within the host genome, each of which was directly associated with the integrase gene they harboured (Figure 3d; Figure 4a). The 44 representative satellite prophage integrases were divided into seven different categories with ≥95% nucleotide sequence similarity within each category. Each integrase category was associated with insertion at a single location on the pneumococcal genome, apart from integrase category I, which was associated with five different locations (Figure 3d). 28.3% of pneumococcal satellite prophages were inserted at site a, which was very close to the origin of replication (oriC) (Figure 4a) and prompted us to investigate whether factors other than the integrase sequence determined the prophage insertion site.

We investigated the location of prophage insertion sites within the genome sequences of non-pneumococcal streptococci for which at least one complete genome was available (n=29). We divided the genome of each species into 8 non-overlapping segments of equal length according to the number of base pairs, and the percentages of prophages situated in each segment were quantified. Overall, we observed no strong preference for prophage insertion in any of the 8 segments and the location of prophages residing within the genome varied greatly between different species (Supplementary Figure 4).

Among pneumococcal and non-pneumococcal streptococcal genomes, five flanking genes upstream and downstream of each prophage were retrieved for functional classification using gene ontology analyses. This revealed that nearly one-third of all the bacterial flanking genes were involved in replication, recombination, DNA repair, transcription, translation and ribosomal structure and biogenesis (Figure 4b). One-quarter of flanking genes were involved in metabolic processes, but equally, one-quarter of all flanking genes did not have a defined functional classification. The remaining flanking genes were involved in other cellular processes and signalling. A list of all prophage insertion sites and their flanking genes is available in Supplementary Table 3.

Satellite prophages and *vapE* are involved in pneumococcal pneumonia and sepsis in a murine infection model

Our investigation of pneumococcal satellite prophage genes led to the identification of a gene that is a homologue of the 'virulence-associated gene E' (*vapE*) in *S. suis*[36]. We investigated *vapE* in *S. suis* genomes and confirmed that it is carried by a satellite prophage. We searched for *vapE* in the representative pneumococcal satellite prophages and found that 30/44 (68.18%) contained *vapE*. To investigate whether the *vapE* homologue in the pneumococcal satellite prophage is also associated with virulence, we performed *in vivo* studies using a murine pneumococcal infection model.

Deletion mutant stains were constructed in a serotype 6B pneumococcal strain (BHN418) in which either *vapE* only (*ΔvapE*) or the entire satellite prophage sequence (*ΔSpnSP38)* were replaced by a spectinomycin resistance cassette (*aadA9*). Both mutant pneumococcal strains grew as well as the parental wild-type strain in Todd-Hewitt plus yeast extract broth media (data not shown). For each of the mutant strains a competitive index (CI) was determined using a highly sensitive competitive infection experiment in a mouse model of pneumonia.

The CI was significantly lower in the lungs after mixed infection with *ΔSpnSP38* vs serotype 6B or *ΔvapE* vs serotype 6B, indicating a role for the satellite prophage and *vapE* in the establishment of pneumococcal pneumonia (Figure 5a). To further assess the degree of attenuation in virulence of the *ΔSpnSP38* and *ΔvapE* strains, infection experiments were repeated with pure inocula of each strain in both the pneumonia and sepsis models. There were no significant differences in bacterial CFU recovered from the lungs of infected mice at 24 h between either mutant and the parental wild-type strain (Figure 5b) and the majority of the mice developed fatal infection by this point. However, in the sepsis model the mice infected with the wild-type serotype 6B strain had significantly greater blood and spleen CFU than the *ΔSpnSP38* mutant (Figure 5c and 5d), indicating that the satellite prophage is directly involved in pneumococcal virulence during bacterial dissemination in the systemic circulation. Although the *ΔvapE* strain had lower spleen CFU compared to the wild-type, this difference was not statistically significant, suggesting that loss of the whole satellite prophage has a more marked effect on the attenuation of virulence during sepsis than loss of VapE alone.

Discussion

In this study we revealed an extraordinarily diverse collection of full-length prophages and satellite prophages among streptococcal species. What was striking about these findings was that prophages and

satellite prophages were two clearly different entities and both had a structured population. Specifically, among pneumococci there were prophages with persistent associations to genetic lineages of pneumococci over long periods of time. This is crucial, since these data allow for the exploration of *why* certain combinations of prophages and bacteria exist and whether the prophages might be contributing to the epidemiological success of bacterial genetic lineages.

Our findings suggest that prophages are likely to be influencing bacterial biology and epidemiology to a much greater extent than previously appreciated, given the high proportion of prophage DNA present in many streptococcal species. Prophages are mobile genetic elements and genetically similar prophages were frequently detected between different streptococcal species. This implies that prophage transmission across bacterial species is more common than previously recognised, which should be taken into account when trying to understand the precise role of prophages in streptococcal biology.

Many of the streptococci we investigated are important human and animal pathogens and we demonstrated that a previously unrecognised pneumococcal satellite prophage was significantly associated with virulence in a murine model of infection. The mechanism driving virulence is not yet clear, but this work is proof-of-principle that experimental investigations of pneumococcal prophages should be pursued, and these may reveal central aspects of the bacteria/prophage relationship among pneumococci and other streptococci.

The increasingly large volume of genome sequence data in the public domain presents many new opportunities for understanding bacterial infection and pathogenesis at a depth and breadth never before experienced. Large population-level analyses such as this alter our perspective on how bacterial and prophage populations interact and drive evolution of both parasite and host. As demonstrated here, population genomics studies can and should be used to generate hypotheses, design experiments and select the most appropriate strains for testing. The findings of this study reveal numerous areas for further investigation, the results of which will increase our knowledge of prophage and bacterial biology, epidemiology and evolution.

Methods

Development of PhageMiner, a bioinformatics tool for prophage identification in bacterial genomes

Some *in silico* prophage detection tools are available that identify prophages using a reference database of known prophage genomes, thus their performance is strongly influenced by the size and composition of the reference dataset[37,38]. In order to ensure a thorough discovery of previously unidentified prophages, manual

curation of annotated genomes is required, however, this is not feasible for large genome studies[26,39-40]. To address these issues, we developed a user-supervised semi-automated computational tool called PhageMiner in order to streamline the otherwise tedious manual curation process for prophage sequence discovery. PhageMiner uses a mean shift algorithm combined with annotation-based genome mining in order to rapidly identify prophage sequences within complete or draft bacterial genomes. The source code of PhageMiner is deposited in GitHub (pending).

Genomes used in this study

In total, 1,316 assembled genomes from 70 different species of the genus *Streptococcus* were selected for this study. 482 genomes belonged to a pneumococcal dataset previously characterised by us[26]. This collection was designed to be highly diverse and consisted of pneumococci recovered from both ill and healthy individuals of all ages residing in 36 different countries between 1916 and 2009. These pneumococci represented 91 serotypes and 94 different clonal complexes (Supplementary Table 4).

The remaining 834 streptococcal genomes were selected from a non-pneumococcal *Streptococcus* species genome dataset previously compiled by us[41]. In brief, 69 different *Streptococcus* species were included in this dataset and up to 50 genomes per species were selected for analyses from the ribosomal MLST database (https://pubmlst.org/rmlst)[42]. When more than 50 genomes were available, the population structure of the species was depicted using PHYLOViZ[43] and genomes were selected to maximise the population-level diversity of the species from the available genomes. All streptococcal genome sequences were stored in a BIGSdb database[44] and annotated using the RAST server (http://rast.nmpdr.org).

Sequence analyses of prophages

All putative prophage sequences were inspected manually using Geneious version 11.1 (Biomatters Ltd.) and those containing ambiguous bases (N's) and/or assembly gaps were excluded from further analyses. The total number of open reading frames (ORFs), overall sequence length and GC content of each prophage were calculated within the Geneious environment. All multiple sequence alignments were performed using ClustalW[45] with default parameters (Gap open cost = 15, Gap extend cost = 6.66). Phylogenetic trees were constructed based upon sequence alignments using FastTreeMP[46]. Unique integrase sequences were identified using the CD-HIT program[47] and a threshold of ≥95% sequence identity. Schematic diagrams of the coding regions of the prophages were produced in Geneious and edited using Adobe Illustrator.

Estimation of prophage content within bacterial genomes

The phage content was estimated based on the percentage of prophage genes within a given bacterial genome. To do this, we developed a Python script that first used Prodigal software in the Prokka annotation suite[48] to predict ORFs in three separate groups of sequences: (i) all identified full-length prophage genomes, (ii) all identified satellite prophage genomes and (iii) a single bacterial genome of interest for which the phage content is to be estimated. Next, the individual ORF nucleotide sequences from all three groups were extracted, combined and clustered using Roary[49] set at a 70% similarity threshold. Any ORFs in the bacterial genome that were also present in at least one prophage genome were deemed to be phage-related, and this information was used to output the total percentage of phage-related ORFs in the given bacterial genome. The PhageContentCalculator script is available in GitHub (pending).

Investigation of prophage insertion sites

The prophage insertion sites within the bacterial genome sequences were investigated among the representative pneumococcal prophages and any streptococcal species for which at least one complete bacterial genome was available. To investigate the bacterial genes flanking the prophage sequences, five genes both upstream and downstream of each prophage were retrieved and functional annotations were determined using eggNOG-mapper[50] based on eggNOG 4.5 orthology data[51]. Prophage insertion sites containing ambiguous bases or assembly gaps were excluded from the analyses. Bacterial genomes were divided into 8 equally-sized segments and the prevalence of prophages per segment was calculated using an in-house Python script (available upon request).

Construction of a pneumococcal core genome phylogenetic tree

The 482 pneumococcal genomes in the study dataset were annotated using Prokka in order to create GFF3 files compatible with downstream analysis scripts. Genes present in all strains were clustered at 90% sequence identity threshold and aligned using Roary. The phylogenetic tree was generated using FastTreeMP[46] using a generalized time-reversible model and then was reconstructed using ClonalFrameML[52] to account for recombination. The tree was annotated using iTOL[53] and Adobe Illustrator (Adobe Inc.).

315 **Estimate of phylogenetic relationships among *Streptococcus* species**

A phylogenetic tree was constructed using concatenated sequence data from 53 ribosomal loci among all streptococcal genomes in the study dataset using the BIGSdb PhyloTree plugin. The tree was graphically simplified to the species level by collapsing clades containing genomes from the same species into a single
320 leaf using iTOL.

Bacterial strains, media and growth conditions

S. pneumoniae strains were cultured in the presence of 5% CO_2 at 37°C on Columbia agar (Oxoid)
325 supplemented with 5% horse blood, or in Todd-Hewitt broth supplemented with 0.5% yeast-extract (THY; Oxoid). Mutant strains were selected by using appropriate antibiotics (150 µg/ml spectinomycin). Growth of *S. pneumoniae* strains in broth was monitored by measuring optical density at 580 nm (OD_{580}) and stocks of *S. pneumoniae* were stored as single use 0.5 ml aliquots of THY broth culture (OD_{580} 0.4–0.5) at −70°C in 10% glycerol.

330

Construction of Δ*vapE* and Δ*SpnSP38 S. pneumoniae* mutant strains

Strains, plasmids and primers used for this study are described in Supplementary Table 5. Both mutants, Δ*vapE* and *ΔSpnSP38* were generated by overlap extension PCR[54,55] in the *S. pneumoniae* serotype 6B
335 BHN418 strain using a transformation fragment in which the *Spn_00749* gene (*vapE*) or the entire satellite prophage, *Spn_00738-Spn_00753*, were replaced by the spectinomycin resistance cassette *aadA9*. For the satellite prophage, two products corresponding to 762 bp upstream (primers SpnSP_UpF and SpnSP_UpspecR) and 872 bp downstream (primers SpnSP_Downspec_F and SpnSP_DownR) of the satellite prophage were amplified from *S. pneumoniae* genomic DNA by PCR carrying 3' and 5' linkers
340 complementary to the 5' and 3' portion of the *aacA9* gene respectively. *aadA9* was amplified from pR412 plasmid (a gift from M. Domenech) using PCR and primers SpnSP_Upspec_F and SpnSP_Downspec_R[54].

Similarly, for the in-frame deletion of *vapE*, a construct was created in which 820 bp of flanking DNA upstream of the *vapE* ATG (primers VapE_UpF and VapE_UpspcR) and 526 bp of flanking DNA
345 downstream from the *vapE* ORF (starting from the ATG of the overlapping Spn_00750 ORF, primers VapE_DownspecF and VapE_DownR) were amplified by PCR and fused with the *aadA9* cassette by overlap extension PCR[56]. The resulting constructs were then transformed into the BHN418 strain by

homologous recombination and allelic replacement using a mix of CSP-1 and CSP-2 and standard protocols[57,58]. The mutations were confirmed by PCR analysis and sequencing.

Experimental models of infection

6-week-old female CD-1 mice were obtained from Charles River Laboratory and bred in a conventional animal facility at University College of London. Animal procedures were performed according to United Kingdom (UK) national guidelines for animal use and care and approved by the UCL Biological Services Ethical Committee and the UK Home Office (Project Licence PPL70/6510). Studies investigating pneumococcal sepsis or pneumonia were performed using 6-week-old mice and infected as previously described[59].

Briefly, in the sepsis model mice were challenged with 5×10^6 CFU/ml of the serotype 6B strain or the correspondent mutants in a volume of 150 µl by the intraperitoneal route, whereas for pneumonia, mice under anaesthesia with isofluorane were inoculated intranasally with 50 µl containing 10^7 CFU/mouse of the serotype 6B strain or the mutants. A lethal dose of pentobarbital was administered at 24 or 28 h after challenge and bacterial counts were determined from samples recovered from lung and blood. Lungs and spleens were homogenised through a 0.2 µm filter. Results were expressed as $\log_{10}$ CFU/ml of bacteria recovered from the different sites.

For mixed infection experiments, mice were inoculated with a 50/50 mixture of wild-type and mutant *S. pneumoniae*. The competitive index (CI) was defined as the ratio of the test strain (mutant strain) compared to the control strain (wild-type strain) recovered from mice divided by the ratio of the test strain to the control strain in the inoculum[60,61]. A CI of <1 indicates that the test strain is attenuated in virulence compared to the control strain and the lower the CI the more attenuated the strain. Statistical analyses were performed using analysis of variance (ANOVA) for multiple comparisons. GraphPad Prism 7.0 (GraphPad Software, San Diego, CA) was used for statistical analysis.

Data availability

The newly discovered full-length and satellite prophage genomes have been deposited at GenBank (accession numbers pending). The nucleotide sequence of the *vapE* gene is available via GenBank accession number (pending). Accession numbers for all genomes used in this study are listed in the Supplementary Table 3.

Author contributions

R.R.J. and A.B.B. conceived and designed the overall study. R.R.J. wrote the computer code. E.R.S. and J.B. designed the pneumococcal mutants and animal experiments. E.R.S. and A.A. created the genetic mutants and E.R.S. performed the animal experiments. R.R.J., A.B.B., E.R.S. and J.B. analysed the data. R.R.J. and A.B.B. wrote the manuscript. All authors read and reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

References

1. Krzyściak W, Pluskwa K, Jurczak A, Kościelniak D. The pathogenicity of the *Streptococcus* genus. *Eur J Clin Microbiol Infect Dis* **32**,1361-1376 (2013).

2. O'Brien K, Wolfson L, Watt J, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T; Hib and Pneumococcal Global Burden of Disease Study Team. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet*. **374**, 893-902 (2009).

3. Carapetis J, Steer A, Mulholland E, Weber M. The global burden of group A streptococcal diseases. *Lancet Infect Dis*. **5**, 685-694 (2005).

4. Vornhagen J, Adams Waldorf K, Rajagopal L. Perinatal group B Streptococcal infections: virulence factors, immunity, and prevention strategies. *Trends Microbiol*. **25**, 919-931 (2017).

5. Boyd E, Brüssow H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*. **10**, 521-529 (2002).

6. Casjens S. Prophages and bacterial genomics: what have we learned so far? *Molec Microbiol*. **49**, 277-300 (2003).

7. Bensing B, Siboo I, Sullam P. Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect Immun*. **69**, 6186-6192 (2001).

8. Vaca Pacheco S, Garcıća González O, Paniagua Contreras G. The lom gene of bacteriophage λ is involved in *Escherichia coli* K12 adhesion to human buccal epithelial cells. *FEMS Microbiol Lett*. **156**, 129-132 (2006).

9. Mirold S, Rabsch W, Rohde M, Stender S, Tschape H, Russmann H, Igwe E, Hardt WD. Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Nat Acad Sci USA*. **96**, 9845-9850 (1999).

10. Bulgin R, Raymond B, Garnett J, Frankel G, Crepin V, Berger C, Arbeloa A. Bacterial guanine nucleotide exchange factors SopE-like and WxxxE effectors. *Infect Immun*. **78**, 1417-1425 (2010).

11. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Molec Microbiol*. **39**, 260-272 (2001).

12. Menouni R, Hutinet G, Petit M, Ansaldi M. Bacterial genome remodeling through bacteriophage recombination. *FEMS Microbiol Lett*. **362**, 1-10 (2015).

13. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits A. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol.* **13**, 641-650 (2015).

14. Koskella B, Brockhurst M. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev*. **38**, 916-931 (2014).

15. Varon M, Levisohn R. Three-membered parasitic system: a bacteriophage, *Bdellovibrio bacteriovorus*, and *Escherichia coli*. *J Virol*. **9**, 519-525 (1972).

16. Belfort M. Bacteriophage introns: parasites within parasites? *Trends Genet*. **5**, 209-213 (1989).

17. Novick R. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*. *Plasmid*. 49, 93-105 (2003).

18. Novick R, Christie G, Penadés J. The phage-related chromosomal islands of Gram-positive bacteria. *Nat Rev Microbiol*. **8**, 541-551 (2010).

19. Penadés J, Christie G. The phage-inducible chromosomal islands: a family of highly evolved molecular parasites. *Ann Rev Virol*. **2**, 181-201 (2015).

20. Frígols B, Quiles-Puchalt N, Mir-Sanchis I, Donderis J, Elena S, Buckling A, Novick RP, Marina A, Penadés JR. Virus satellites drive viral evolution and ecology. *PLOS Genetics*. **11**, e1005609 (2015).

21. O'Neill A, Larsen A, Skov R, Henriksen A, Chopra I. Characterization of the epidemic European fusidic acid-resistant impetigo clone of *Staphylococcus aureus*. *J Clin Microbiol*. **45**, 1505-1510 (2007).

22. Scott J, Nguyen S, King C, Hendrickson C, McShan W. Phage-Like *Streptococcus pyogenes* Chromosomal Islands (SpyCI) and Mutator Phenotypes: Control by Growth State and Rescue by a SpyCI-Encoded Promoter. *Front Microbiol*. **3** (2012).

23. Seed K, Lazinski D, Calderwood S, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*. **494**, 489-491 (2013).

24. Lindsay J, Ruzin A, Ross H, Kurepina N, Novick R. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Molec Microbiol*. **29**, 527-543 (1998).

25. Martínez-Rubio R, Quiles-Puchalt N, Martí M, Humphrey S, Ram G, Smyth D, Chen J, Novick RP, Penadés JR. Phage-inducible islands in the Gram-positive cocci. *ISME J*. **11**, 1029-1042 (2016).

26. Brueggemann A, Harrold C, Rezaei Javan R, van Tonder A, McDonnell A, Edwards B. Pneumococcal prophages are diverse, but not without structure or history. *Sci Rep*. **7** (2017).

27. Romero P, García E, Mitchell TJ. Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. *Appl. Environ. Microbiol*. **75**, 1642–9 (2009).

28. Ramirez M, Severina E, Tomasz A. A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J. Bacteriol*. **181**, 3618–25 (1999).

29. Beres S, Sylva G, Barbian K, Lei B, Hoff J, Mammarella N, Liu MY, Smoot JC, Porcella SF, Parkins LD, Campbell DS, Smith TM, McCormick JK, Leung DY, Schlievert PM, Musser JM. Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Nat Acad Sci USA*. **99**, 10078-10083 (2002).

30. McShan WM, Nguyen SV. The bacteriophages of *Streptococcus pyogenes*. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations. University of Oklahoma Health Sciences Center (2016). Available from: https://www.ncbi.nlm.nih.gov/books/NBK333409/

31. van der Mee-Marquet N, Diene S, Barbera L, Courtier-Martinez L, Lafont L, Ouachée A et al. Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of group B *Streptococcus* species. *Clin Microbiol Infect*. **24**, 514-521 (2018).

32. Canchaya C, Desiere F, McShan W, Ferretti J, Parkhill J, Brüssow H. Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370. *Virol*. **302**, 245-258 (2002).

33. Davies E, Winstanley C, Fothergill J, James C. The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol Lett*. **363**, fnw015 (2016).

34. Bobay L, Touchon M, Rocha E. Pervasive domestication of defective prophages by bacteria. *Proc Nat Acad Sci USA* **111**, 12127-12132 (2014).

35. Ackermann H, Audurier A, Berthiaume L, Jones L, Mayo J, Vidaver A. Guidelines for Bacteriophage Characterization. Adv Virus Res **23**, 1-24 (1978).

36. Ji X, Sun Y, Liu J, Zhu L, Guo X, Lang X, Feng S. A novel virulence-associated protein, VapE, in *Streptococcus suis* serotype 2. *Mol Med Rep*. **13**, 2871-7. (2016)

37. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*. **24**, 863-865 (2008).

38. Zhou Y, Liang Y, Lynch K, Dennis J, Wishart D. PHAST: A Fast Phage Search Tool. *Nucl Acids Res*. **39**, W347-W352 (2011).

39. Crispim J, Dias R, Vidigal P, de Sousa M, da Silva C, Santana M, de Paula SO. Screening and characterization of prophages in Desulfovibrio genomes. *Sci Rep*. **8** (2018).

40. Langille M, Hsiao W, Brinkman F. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* **8**, 373-382 (2010).

41. Kurioka A, Wilgenburg B, Rezaei Javan R, Hoyle R, van Tonder AJ, Harrold CL, Leng T, Phalora P, Howson LJ, Shepherd D, Cerundolo V, Brueggemann AB, Klenerman P. Diverse *Streptococcus pneumoniae* strains drive a MAIT cell response through MR1-dependent and cytokine-driven pathways. *J Infect Dis* **217**, 988–999. (2018)

42. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MC. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiol* **158**,1005-15. (2012)

43. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. **13**:87 (2012).

44. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. **11**:595 (2010).

45. Thompson JD, Higgins DG, GibsonTJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).

46. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLOS ONE*. **5**:e9490 (2010).

47. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**:1658-1659 (2006).

48. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**:2068-2069 (2014).

49. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. **31**:3691-3693 (2015).

50. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115-2122 (2017).

51. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**:D286-293 (2016).

52. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Comput Biol*. **11**:e1004041 (2015).

53. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**:W475-478 (2011).

54. Khandavilli S, Homer KA, Yuste J, Basavanna S, Mitchell T, Brown JS. Maturation of *Streptococcus pneumoniae* lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence. *Mol Microbiol*. **67**, 541-57 (2008).

55. Basavanna S, Chimalapati S, Maqbool A, Rubbo B, Yuste J, Wilson RJ, Hosie A, Ogunniyi AD, Paton JC, Thomas G, Brown JS. The effects of methionine acquisition and synthesis on *Streptococcus pneumoniae* growth and virulence. *PLOS One*. **8**, e49638 (2013).

56. Heckman KL, Pease LR. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat Protoc* **2**, 924-32 (2007).

57. Håvarstein LS, Coomaraswamy G, Morrison DA. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A*. **92**, 11140-4 (1995).

58. Lau GW, Haataja S, Lonetto M, Kensit SE, Marra A, Bryant AP, McDevitt D, Morrison DA, Holden DW. A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. *Mol Microbiol* **40**, 555-71 (2001).

59. Ramos-Sevillano E, Urzainqui A, Campuzano S, Moscoso M, González-Camacho F, Domenech M, Rodríguez de Córdoba S, Sánchez-Madrid F, Brown JS, García E, Yuste J. Pleiotropic effects of cell wall amidase LytA on *Streptococcus pneumoniae* sensitivity to the host immune response. *Infect Immun*. **83**, 591-603 (2015).

60. Yuste J, Botto M, Paton JC, Holden DW, Brown JS. Additive inhibition of complement deposition by pneumolysin and PspA facilitates *Streptococcus pneumoniae* septicemia. *J Immunol*. **175**, 1813-9 (2005).

61. Beuzón CR, Holden DW. Use of mixed infections with Salmonella strains to study virulence genes and their interactions in vivo. *Microbes Infect*. **3**, 1345-52 (2001).

Table 1. Epidemiological characteristics of 44 representative satellite prophages identified among a collection of pneumococcal isolates dating from 1939 onwards.

		Pneumococci						
Satellite Prophage Name	Satellite Prophage Cluster	Clonal Complex (n)	Genomes (n)	Isolation dates	Countries (n)	Serotypes (n)	Insertion Site	Integrase Category
SpnSP16	A	3	4	1939 - 1982	2	4	b	I
SpnSP3	A	2	3	1981 - 2004	2	2	b	I
SpnSP26	A	2	3	1985 - 2000	1	2	b	I
SpnSP35	A	2	2	1952 - 1952	1	2	b	I
SpnSP43	A	2	2	1939 - 2004	2	2	b	I
SpnSP30	A	1	5	1978 - 1978	1	1	b	I
SpnSP44	A	1	2	1939 - 1962	1	1	b	I
SpnSP7	A	1	1	1968	1	1	b	I
SpnSP25	A	1	1	1999	1	1	b	I
SpnSP19	A	Singleton[a]	2	1939 - 1952	2	2	b	V
SpnSP11	A	Singleton	1	1952	1	1	b	I
SpnSP5	B	5	15	1939 - 2007	3	7	d, g	I
SpnSP29	B	1	15	1978 - 1988	1	2	b	I
SpnSP27	B	1	1	2006	1	1	b	I
SpnSP20	B	Singleton	1	1954	1	1	b	I
SpnSP2	C	2	4	1984 - 2005	3	2	f	VII
SpnSP31	C	2	2	1983 - 2005	1	2	b	I
SpnSP12	C	1	1	1968	1	1	b	I
SpnSP15	C	1	1	1943	1	1	b	I
SpnSP32	C	1	1	1986	1	1	f	VII
SpnSP37	D	5	9	1939 - 1988	4	7	c	II
SpnSP38	D	4	30	1972 - 2006	6	5	c	II
SpnSP6	D	3	8	1939 - 1991	3	3	c	II
SpnSP23	D	2	11	1962 - 2008	3	4	a	III
SpnSP39	D	1	2	2005 - 2007	1	1	a	III
SpnSP18	D	Singleton	2	1939 - 1952	2	2	c	II
SpnSP24	E	6	23	1939 - 2006	6	4	a	III
SpnSP33	E	2	3	1952 - 1998	1	2	a	III
SpnSP1	E	1	5	1978 - 1988	1	1	b	I
SpnSP40	E	1	3	2001	2	2	a	III
SpnSP8	E	1	1	1988	1	1	a	III
SpnSP9	E	1	1	1957	1	1	a	III
SpnSP13	E	1	1	1943	1	1	a	III
SpnSP14	E	1	1	1995	1	1	a	III
SpnSP17	E	1	1	1972	1	1	a	IV
SpnSP22	E	1	1	1971	1	1	a	III
SpnSP28	E	1	1	2003	1	1	a	III
SpnSP34	E	1	1	1990	1	1	a	III
SpnSP36	E	1	1	1963	1	1	a	III
SpnSP42	E	1	1	1994	1	1	a	III
SpnSP4	E	Singleton	1	1982	1	1	e	I
SpnSP10	E	Singleton	1	N/A	1	1	h	I
SpnSP21	E	Singleton	1	1954	1	1	e	VI
SpnSP41	E	Singleton	1	1983	1	1	a	III

a. Singletons are genotypes with no closely related variants.

a

Species	Full-length Prophages			Satellite Prophages		
	Min	Max	Average	Min	Max	Average
S. pyogenes(50)	1.0%	19.0%	9.5%	0.4%	3.3%	1.3%
S. urinalis (4)	6.8%	9.1%	8.0%	1.5%	1.6%	1.5%
S. pseudopneumoniae (16)	3.9%	7.0%	5.4%	0.7%	2.8%	1.7%
S. parauberis (16)	0.4%	10.5%	5.4%	0.4%	3.2%	1.6%
S. canis (11)	2.0%	9.5%	4.5%	2.0%	2.6%	2.3%
S. dysgalactiae (50)	0.7%	11.9%	3.8%	0.3%	4.5%	1.3%
S. suis (50)	0.5%	9.1%	3.7%	0.4%	3.7%	1.3%
S. equi (50)	0.5%	11.1%	4.4%	0.2%	2.4%	0.6%
S. agalactiae (50)	0.3%	8.7%	4.0%	0.2%	2.4%	1.0%
S. infantis (4)	1.5%	7.6%	4.2%	0.4%	1.3%	0.8%
S. uberis (13)	0.5%	6.7%	3.2%	0.9%	2.9%	1.7%
S. constellatus (10)	1.4%	6.7%	3.9%	0.4%	1.3%	0.9%
S. azizii (3)	3.7%	5.5%	4.3%	0.5%	0.5%	0.5%
S. pseudoporcinus (5)	1.9%	6.7%	3.8%	0.5%	1.0%	0.7%
S. mitis (49)	0.4%	9.3%	2.5%	0.3%	3.5%	1.6%
S. pasteurianus (5)	1.3%	5.2%	2.5%	0.8%	2.3%	1.4%
S. gallolyticus (17)	0.5%	6.2%	2.9%	0.4%	1.8%	0.9%
S. pneumoniae (482)	0.4%	8.4%	2.9%	0.2%	3.3%	0.9%
S. anginosus (24)	0.4%	7.6%	2.4%	0.4%	2.3%	1.2%
S. oralis (49)	0.5%	6.4%	2.1%	0.2%	2.8%	1.0%
S. iniae (8)	0.5%	8.6%	2.3%	0.2%	1.5%	0.6%
S. parasanguinis (31)	0.4%	5.4%	2.2%	0.3%	2.0%	0.7%
S. equinus (27)	0.4%	7.2%	1.9%	0.2%	2.7%	0.9%
S. cristatus (4)	0.8%	3.1%	1.7%	0.5%	1.5%	0.8%
S. intermedius (9)	0.5%	3.4%	1.1%	0.5%	1.8%	1.3%
S. gordonii (22)	0.4%	5.6%	1.9%	0.2%	1.3%	0.5%
S. vestibularis (6)	0.4%	0.8%	0.5%	0.5%	2.3%	1.6%
S. salivarius (32)	0.4%	3.0%	1.2%	0.2%	1.7%	0.5%
S. thermophilus (32)	0.3%	0.9%	0.4%	0.2%	1.4%	0.9%
S. sanguinis (37)	0.3%	3.4%	0.8%	0.1%	1.3%	0.4%
S. sobrinus (42)	0.4%	0.7%	0.5%	0.3%	0.6%	0.4%
S. mutans (50)	0.2%	1.5%	0.5%	0.1%	0.3%	0.2%

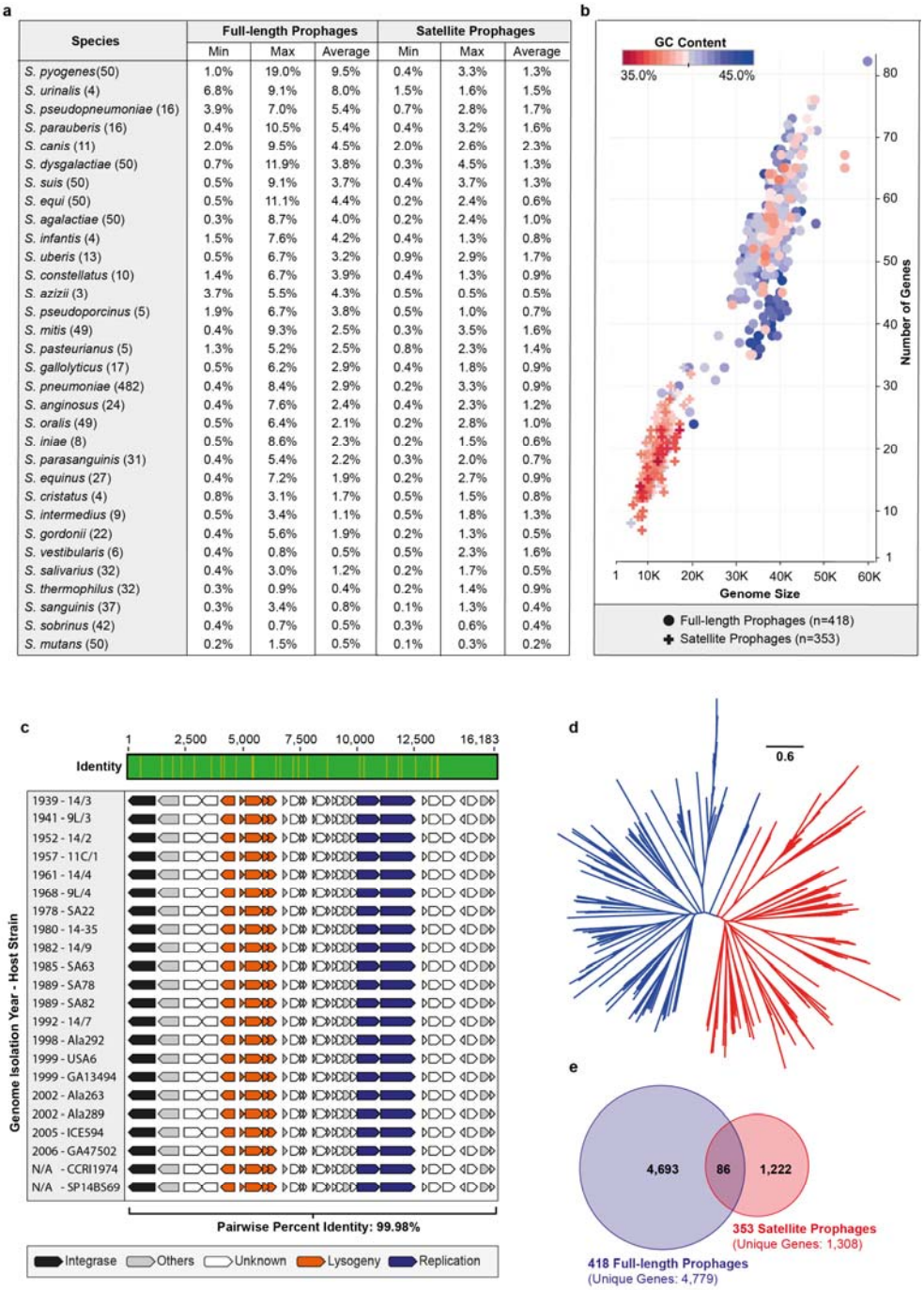

Fig. 1. Full-length and satellite prophages identified among streptococcal genomes. a, Average prophage content within each streptococcal species. **b**, Graphical representation of all prophages by average genome size and number of genes. Each prophage is coloured to represent its average guanine (G) and cytosine (C) content. **c**, Satellite prophage SpnSP24 was represented among pneumococci isolated between 1939 and 2006 and all of these satellite prophages were nearly identical at the nucleotide level one to another. **d**, An unrooted phylogenetic tree of all streptococcal prophage genomes identified in the dataset. Blue branches mark full-length prophages and red branches mark satellite prophages. **e**, Genes found in full-length prophages and in satellite prophages (at a threshold of >70% amino acid sequence similarity) are rarely shared.

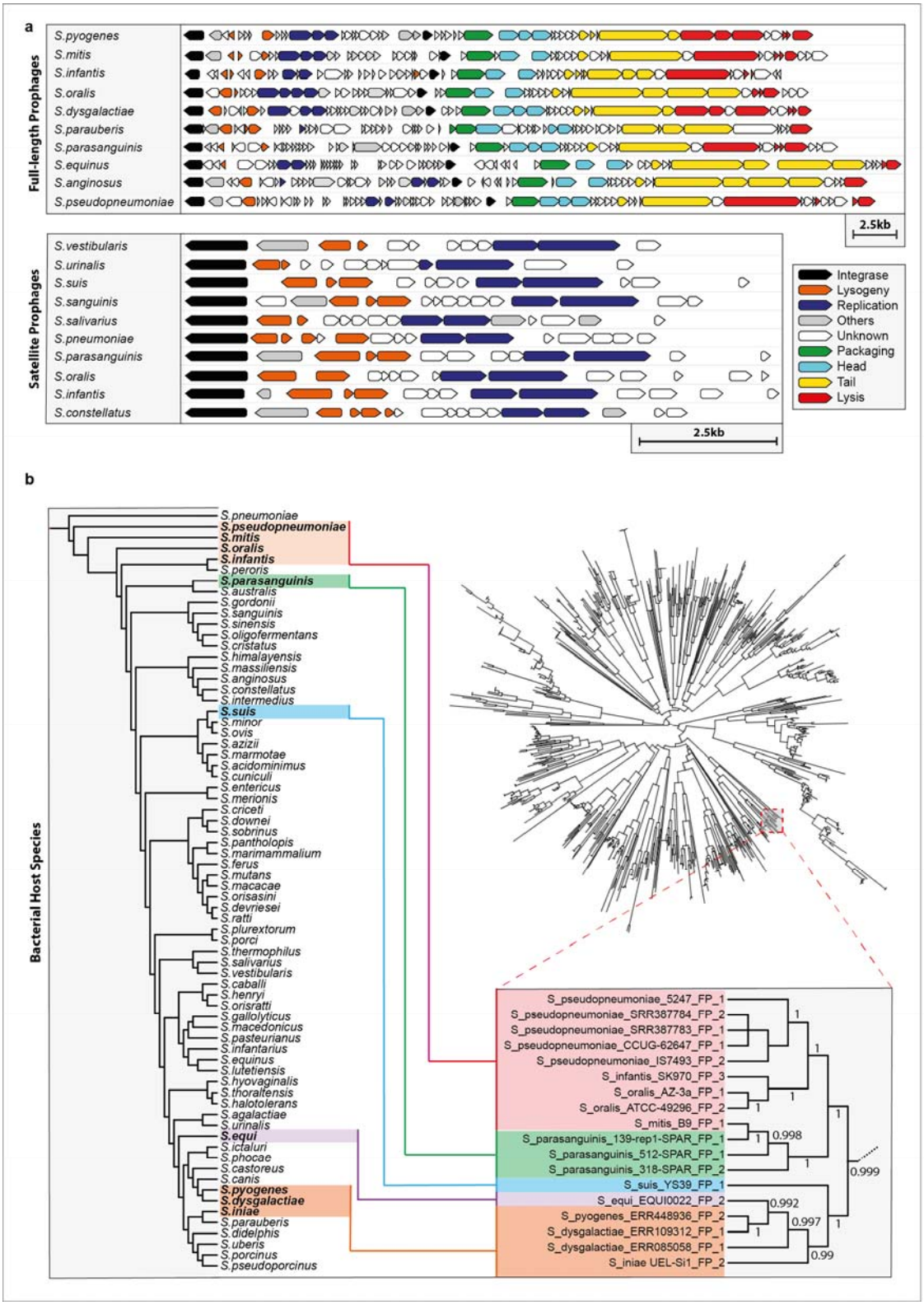

Fig. 2. Similarities among streptococcal prophages and evidence for cross-species transmission of prophages. a, Full-length and satellite prophages identified among different streptococcal species shared an identical pattern in gene orientation and synteny. **b**, Phylogenetic tree depicting all prophages detected in this study (see Supplementary Fig. 2 for a larger version of the tree) and a zoomed-in branch depicting one example of a cluster of full-length prophages that were found among multiple streptococcal species.

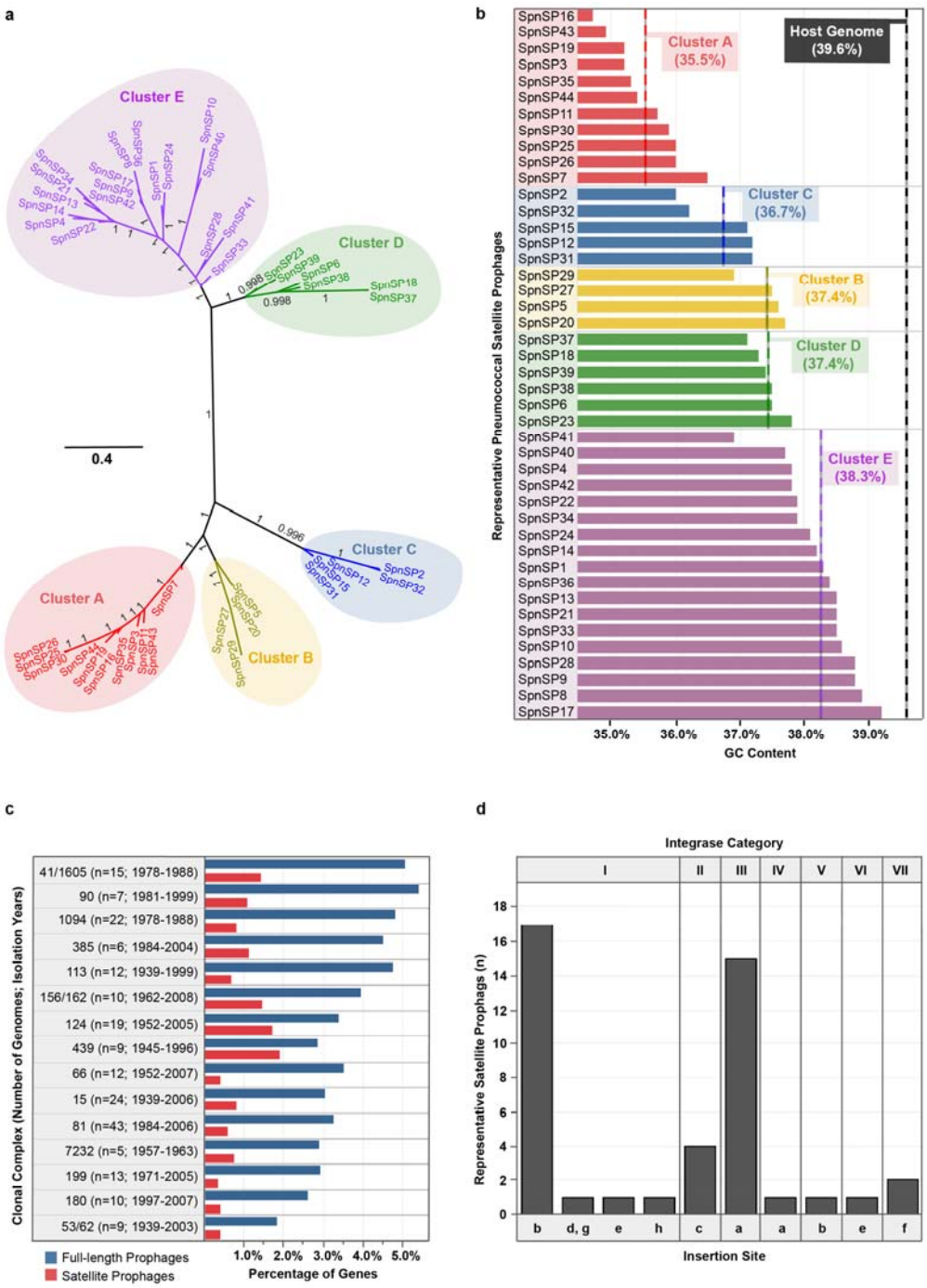

Fig. 3. Satellite prophages found among a large collection of nearly 500 diverse pneumococcal genomes. a, An unrooted phylogenetic tree demonstrated that the 44 representative satellite prophages could be clustered into five major groups based upon nucleotide similarity. **b**, The average guanine/cytosine (GC) content (stated in brackets) of the satellite prophages varied by genetic cluster and was lower than the GC content of the pneumococcal host. **c**, The average prophage content for each of the major clonal complexes (genetic lineages) is depicted as a percentage of the total number of genes in the host pneumococcal genome (~2 Mb). **d**, The integrase sequences of the 44 representative satellite prophages were divided into seven different categories based upon ≥95% nucleotide similarity.

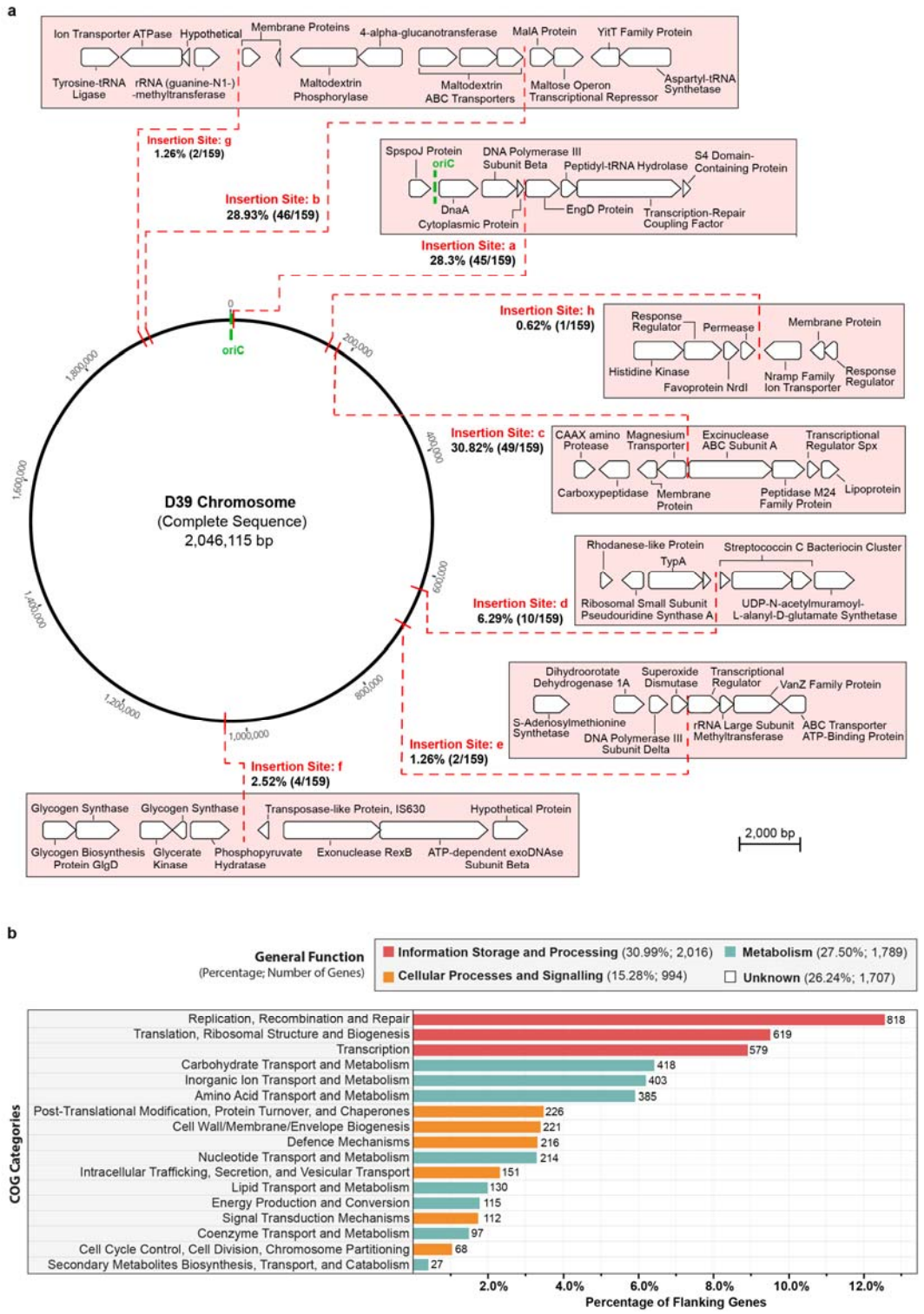

Fig. 4. Insertion sites of prophages within the pneumococcal genome. a, Pneumococcal satellite prophages were integrated in seven locations (a-f) within the host genome. Percentages and numbers in brackets refer to the proportion and number out of all 159 satellite prophages that were inserted in that particular location. **b**, The pneumococcal flanking genes upstream and downstream of all integrated full-length and satellite prophages were retrieved for functional classification and are depicted here based upon their COG (clusters of orthologous groups) classifications.

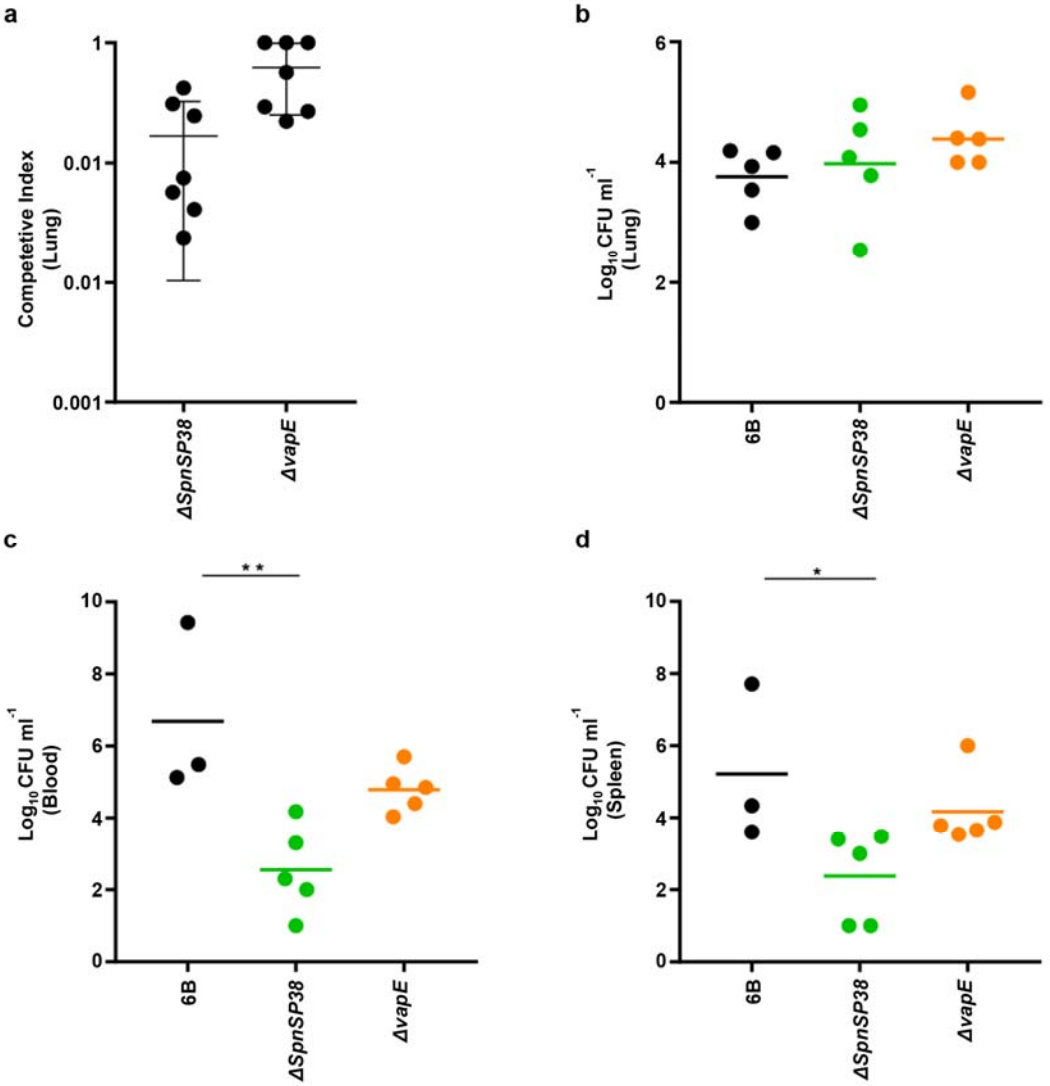

Fig. 5. Assessment of the virulence of *ΔSpnSP38* (deletion of entire satellite prophage) and *ΔvapE* (deletion of *vapE* only) mutant pneumococcal strains in murine infection. **a**, A comparison of the competitive index (CI) in the lungs in a model of pneumonia of mixed infection with the parental wild-type strain versus the mutant. Each point represents the CI for a single animal. **b**, Bacterial levels recovered from lung homogenates at 24h (expressed as colony-forming units (CFU) per ml) after pneumococcal pneumonia produced with the wild-type and mutant strains. **c**, **d**, Bacterial levels recovered at 28h from blood or spleen homogenate. Error bars represent standard error of the mean (Kruskall-Wallis with Dunn's post hoc test).