Streptococcus pneumoniae possesses an unexpectedly large bacteriocin repertoire

Reza Rezaei Javan[1], Andries J van Tonder[1], James P King[1], Angela B Brueggemann*[1,2]

[1]The Peter Medawar Building for Pathogen Research, Nuffield Department of Medicine, University of Oxford, Oxford, OX1 3SY, United Kingdom.

[2]Division of Infectious Diseases, Department of Medicine, Imperial College London, London, W12 0NN, United Kingdom.

Abstract

Streptococcus pneumoniae (the 'pneumococcus') is a frequent coloniser of the human nasopharynx and a leading cause of morbidity and mortality worldwide. Understanding the composition and dynamics of the pneumococcal population is crucial, as these factors influence the observed distribution of disease in the human population. Competition among bacterial members of the nasopharynx is believed to be mediated by bacteriocins: antimicrobial toxins produced by bacteria to inhibit growth of other bacteria. The role of bacteriocins as weapons of defence also makes bacteriocins attractive candidates for the development of new antibiotics. Here, by screening >6,200 pneumococcal genomes, we identified 14 newly-discovered bacteriocin clusters, tripling the number of known bacteriocins among pneumococci. We provide evidence that many of the bacteriocin genes are transcriptionally active when the cell is induced by external stress. Using a large global and historical dataset of pneumococci isolated from 1916-2009, we determined the prevalence, molecular epidemiology and co-occurrence patterns of the identified bacteriocin clusters. Our study demonstrated extraordinary bacteriocin diversity among the pneumococcal genomes and revealed several genome hotspots for the integration of bacteriocin clusters. These findings fundamentally change our view of bacteriocins and nasopharyngeal competition among pneumococci.

The pneumococcus is most commonly an asymptomatic coloniser of the nasopharynx in healthy young children; however, colonisation is also the initial stage of the disease process. Pneumococci are among the most common causes of otitis media, sinusitis and conjunctivitis, and a leading cause of severe infections such as pneumonia, bacteraemia and meningitis. All age groups are susceptible to pneumococcal infection, but young children, the elderly, and immunocompromised persons are most at risk[1].

35

36 Despite the use of antimicrobials and pneumococcal conjugate vaccines, the pneumococcus remains
37 a major global health problem, causing approximately 14.5 million cases of serious disease and
38 826,000 deaths annually in children <5 years of age[2]. Antimicrobial-resistant pneumococci have
39 been a serious and increasing concern for several decades[3,4,5], and the pneumococcus is now
40 considered to be a 'priority pathogen' (antimicrobial-resistant bacteria that pose the greatest threat
41 to global health) by the World Health Organisation[6].

42

43 The polysaccharide capsule is the main virulence factor of the pneumococcus: nearly 100 distinct
44 capsular antigenic types (serotypes) have been described and certain serotypes are predominantly
45 associated with disease whilst others are largely associated with nasopharyngeal colonisation[7].
46 Pneumococci frequently co-colonise with other pneumococci and non-pneumococcal bacterial
47 species, and intra- and inter-specific competition can influence colonisation dynamics, strain
48 prevalence, serotype distributions and consequently, the potential for disease progression[8].

49

50 Competition among bacterial members of the nasopharyngeal microbiome is believed be mediated
51 by bacteriocins: ribosomally-synthesised antimicrobial toxins produced by bacterial species to inhibit
52 the growth of other closely-related bacteria. The producer strain also encodes an immunity protein
53 to protect itself from its own bacteriocin. Bacteriocin production has been associated with more
54 efficient colonisation of a host by the producer strain, owing to the ability of these toxins to remove
55 competitors[8,9]. Their ability to kill bacteria makes bacteriocins attractive potential candidates for the
56 development of new antimicrobials. Several bacteriocins (e.g. nisin and pediocin PA-1) have already
57 been commercialised and are widely used as food preservatives[10].

58

59 From a genetic perspective, bacteriocins are found in gene clusters, whereby the genes involved in
60 toxin production, immunity and transport (exporting and processing the toxin peptide) are situated
61 adjacent to each other in the bacterial genome[9,10,11,12]. The *blp* (bacteriocin-like peptides) cluster is
62 the best-characterised bacteriocin in pneumococcus[9,10,11,12]. Previous work by our group showed that
63 the *blp* cluster is ubiquitous among pneumococci recovered from the early 1900s onward and is
64 highly diverse. We also identified a novel bacteriocin cluster that we named pneumocyclicin[11]. Five
65 additional bacteriocins have been reported among pneumococci[13,14,15,16,17]; however, their
66 prevalence, genetic composition and molecular epidemiology in the context of the pneumococcal
67 population are unknown.

68

The explosion of whole genome sequence data expedites the detection and genetic analysis of bacteriocin clusters. Here, we report the discovery of 14 newly-discovered pneumococcal bacteriocin clusters. We describe the molecular epidemiology of all the currently known bacteriocins within a collection of pneumococci isolated over the past 90 years and reveal astonishing diversity. We demonstrate the presence of several apparent hotspots for the integration of bacteriocin clusters within the pneumococcal genome and provide transcriptomic evidence that multiple bacteriocin clusters were induced in response to external stress.

Results

Genome mining triples the number of known bacteriocins among pneumococci

Our bioinformatic investigation of a large and diverse dataset of 571 historical and modern pneumococcal genomes resulted in the identification of 14 newly-discovered bacteriocin clusters, boosting the number of known bacteriocins in this species to 21 (Table 1). We identified several clusters belonging to the lactococcin 972, sactipeptide and lassopeptide classes of bacteriocins, which were hitherto not known to be harboured by the pneumococcus (Table 1; Supplementary Table 1). We subsequently expanded our search for bacteriocins to a much larger dataset of 5,673 published pneumococcal genomes, but no additional bacteriocins were found; thus, the description of the bacteriocins here is restricted to those identified in the dataset of 571 genomes.

BLAST searches of the NCBI database of non-redundant nucleotide sequences revealed that the majority of these clusters were not exclusive to pneumococcus, but that identical or similar versions were present in other streptococci (Supplementary Table 2). Therefore, we used the prefix “strepto” and named each cluster according to its class of bacteriocin (see Methods). Among the putative bacteriocins, five were similar to lactococcin 972[18] and we called these streptococcins (Fig. 1). Despite gene synteny among the streptococcins (Fig. 1a), they shared little nucleotide sequence identity with each other (Fig. 1b). Different streptococcins were found in different, but consistent, locations within the bacterial chromosome (Fig. 1a and 1c). Eleven different streptolancidins, one streptosactin, one streptolassin and one streptocyclicin (what we previously called pneumocyclicin) were also identified among pneumococcal genomes[19] (Fig. 2). The bacteriocin clusters were commonly flanked by genes predicted to encode CAAX proteins (possibly involved in self-immunity from bacteriocin toxins[20]), Rgg and PlcR transcriptional regulators (implicated in bacterial quorum sensing[21]), transporters, lipoproteins and mobilisation proteins.

104 ***Extraordinary bacteriocin diversity within the global pneumococcal population***

105

106 We further assessed these bacteriocins in the context of the pneumococcal population structure.
107 The study dataset consisted of a diverse collection of 571 pneumococci isolated between 1916 and
108 2009 from patients and healthy individuals of all ages residing in 39 different countries across six
109 continents. 88 pneumococcal serotypes and 99 different clonal complexes were represented (Table
110 1; Supplementary Table 3). All bacteriocins detected more than once in the dataset were identified
111 among pneumococci isolated over several decades and from a variety of different countries (Table
112 1). Some bacteriocins were present in all pneumococci, whilst others were limited to specific clonal
113 complexes (genetic lineages). Bacteriocin clusters missing one or more genes were deemed as
114 partial. The percentages of partial and complete clusters varied between different bacteriocins, e.g.
115 streptococcin E was present in all but one pneumococcal genome, but was missing one or more
116 genes in the majority (86.9%) of genomes, while streptococcin C was found as a complete cluster in
117 all genomes (Table 1 and Fig. 3a). We constructed a core genome phylogenetic tree of all isolates
118 and labelled each genome according to the presence or absence of each bacteriocin cluster
119 (Supplementary Figure 1). Overall, we found that the number of bacteriocins within each genome
120 varied from 6 to 11 per genome and that certain combinations of bacteriocins were well represented
121 in the dataset (Fig. 3b, 3c and Supplementary Figure 1).

122

123 Due to their relative simplicity (containing only three genes), we chose streptococcins as models for
124 further investigating the pattern of missing genes in partial clusters (Fig. 3a). Scrutinising these
125 clusters revealed that the majority of the partial clusters lack the toxin gene, while still retaining the
126 immunity and/or the transporter genes. This could support the general idea of a 'cheater'
127 phenotype, whereby the immunity and transporter genes are conserved to protect the
128 pneumococcus from neighbouring bacteria that express the bacteriocin toxin protein, but the
129 cheater strain does not bear the cost of producing the bacteriocin toxin[22,23].

130

131 ***Multiple hotspots for the integration of bacteriocin clusters in the pneumococcal genome***

132

133 By conducting a population genomics-based assessment of the bacteriocin cluster insertion sites, we
134 identified three genomic regions that are putative hotspots for the integration of bacteriocin clusters
135 in the pneumococcal chromosome. These **b**acteriocin **c**luster **h**otspots (BCHs) are specific locations
136 in the genome where different bacteriocin clusters were found in different pneumococci but the

137 flanking genes were the same (Fig. 4). This suggested a switching mechanism, whereby different
138 clusters can replace one another via genetic recombination. Up to three different bacteriocin
139 clusters were found to be associated with a single BCH (Fig. 4c). The acquisition of streptolancidin G
140 appears to have rendered the streptococcin E partial by replacing the toxin and part of its immunity
141 gene (Fig. 4a), which is in accord with the fact that streptolancidin G could not be found in genomes
142 that harboured a complete streptococcin E cluster (Supplementary Figure 1). Nonetheless, the
143 remnant genes of the streptococcin E partial clusters were conserved in samples collected over
144 many decades (Fig. 3a).
145
146 ***Transcriptomic analyses demonstrate temporal dynamic changes in the expression of***
147 ***bacteriocin genes in response to heat***
148
149 To explore whether the bacteriocin clusters were transcriptionally active, we analysed a whole-
150 genome RNA sequencing dataset from a broth culture of pneumococcal strain 2/2, which was
151 incubated at a higher than normal temperature (40 vs 37°C) (NCBI GEO accession number
152 GSE103778[24]). Multiple genes in the bacteriocin clusters were differentially expressed as compared
153 to the control over several time points, indicating that many of these bacteriocin genes were
154 transcribed in response to external stress (Fig. 5). Notably, there were two bacteriocin clusters
155 (streptolancidin J and streptococcin E) that were missing genes and yet the remaining genes were
156 clearly being upregulated. Moreover, the timing of gene expression varied across bacteriocin
157 clusters, e.g. genes within a cluster were induced in a specific pattern, whilst at any particular time
158 point during the sampling period several genes in different clusters were upregulated
159 simultaneously. The regulatory governance over which bacteriocin genes are expressed at any given
160 point during bacterial growth remains to be determined.
161
162 ## *Discussion*
163
164 The prevalence of different pneumococcal serotypes and genotypes in the nasopharynx fluctuates
165 over time, indicating the importance of intra-species competition to pneumococcal ecology[25,26,27].
166 We revealed here that not only do pneumococci possess a wide variety of bacteriocins, the
167 bacteriocins, often in a particular combination, are associated with specific genetic lineages. This is
168 fundamental, as it provides the framework on which to investigate the mechanisms underpinning
169 specific bacteriocin-pneumococcus combinations and the activity of specific bacteriocin and
170 immunity genes. A clear understanding of the role bacteriocins play in pneumococcal biology is

171 central to understanding microbial interactions within the ecological niche. For example,
172 understanding competition dynamics is important in the context of understanding how vaccines
173 affect the pneumococcal population structure. Pneumococcal conjugate vaccines are disruptive to
174 the pneumococcal population structure and alter the composition of microbes competing for space
175 and nutrients in the nasopharynx. The effects of this disruption are not yet fully understood, but can
176 lead to increased disease in human populations[28,29,30].
177
178 Here, we revealed that the pneumococcus possesses a substantially greater and more varied array of
179 bacteriocins than previously recognised. Using a large global and historical pneumococcal genome
180 dataset, we determined the prevalence, co-occurrence patterns and molecular epidemiology of all
181 the identified bacteriocin clusters within a collection of pneumococci recovered over the past 90
182 years. RNA sequencing of the bacteriocin clusters present in one pneumococcal isolate clearly
183 demonstrated that these bacteriocin genes were transcriptionally active when the pneumococcus
184 was under stress. The *in vivo* production of bacteriocin gene products and their functional activities
185 requires further investigation.
186
187 Interestingly, we identified several specific locations in the pneumococcal genome that harboured
188 different bacteriocin clusters, which suggested that recombination events had occurred at these
189 locations and resulted in a switching of bacteriocin clusters. Similar recombination events are well
190 documented in other locations within the pneumococcal genome, most commonly at the capsular
191 polysaccharide locus (conferring a change of serotype) and at penicillin binding protein genes
192 (conferring penicillin resistance if altered)[31,32,33,34,35]. Similarly, the DpnI, DpnII and DpnIII clusters,
193 each containing a distinct restriction modification system, can replace one another at the *dpn* locus.
194 This is believed to be of protective value in mixed pneumococcal populations against
195 bacteriophages, which are constant invaders of pneumococcal genomes[36,37,38]. We found several
196 examples of presumed bacteriocin cluster switching events that had occurred adjacent to distinct
197 quorum-sensing transcriptional regulators TprA and Rgg. Intriguingly, while the genes that are
198 known to be under the control of TprA and Rgg were replaced, the quorum-sensing transcriptional
199 regulators remained conserved. It is known that TprA controls the expression of its downstream
200 bacteriocin genes: it induces them when pneumococcal cells are at high density in the presence of
201 galactose, and represses them when under high glucose growth condition[14]. Galactose is abundant
202 in the nasopharynx, whereas glucose is absent, suggesting that the TprA may mediate the expression
203 of its adjacent bacteriocin genes to aid pneumococci to compete for resources during the
204 nasopharynx colonisation[14]. Likewise, the Rgg quorum-sensing transcriptional regulator has been

205 shown to mediate the expression of its adjacent genes, and this is thought to be directed by sensing
206 amino acid levels in the cellular community[39]. An explanatory hypothesis might be that bacteriocin
207 cluster switching provides a mechanism by which the existing intricate quorum-sensing signalling
208 network required for coordinating population-level behaviours is accessible by the newly-acquired
209 bacteriocin cluster. This remains to be experimentally verified.
210
211 Whilst *in vivo* assessment of the functional activity of the bacteriocin genes is pending, our current
212 work is significant for the broader community in that among the 21 different pneumococcal
213 bacteriocin clusters now identified, many homologues are found in other unrelated bacterial
214 species. Bacteriocins are possible candidates for further investigation as novel antimicrobials and
215 there may be opportunities to develop bacteriocins with cross-species activity. We also provide here
216 a unified nomenclature for the pneumococcal bacteriocin clusters and their genes. Finally, the fact
217 that any single pneumococcal genome possesses multiple bacteriocin clusters should be carefully
218 considered when designing laboratory experiments aimed at assessing the activity of an individual
219 bacteriocin cluster.
220
221 Overall, these population genomic and transcriptomic analyses reveal an extraordinary complexity of
222 bacteriocins among pneumococci, and underscore the need to determine precisely how these
223 bacteriocins drive changes within the pneumococcal population and the wider microbial community.
224 Such findings are interesting not only for their population biology and ecology perspectives, but also
225 because bacteriocins potentially have a huge impact on public health. By directly influencing
226 changes in microbial populations, bacteriocins might indirectly be affecting the effectiveness of
227 vaccines in the longer term: after vaccine use in human populations, the target pneumococcal
228 population is significantly changed and those pneumococci must now compete within the altered
229 microbiome. If bacteriocins are essential to the pneumococcal competitive strategy then the
230 composition of bacteriocins possessed by the post-vaccination pneumococcal community is
231 important to understand. Moreover, there is obvious potential for the development of bacteriocins
232 as novel antimicrobials and our study presents many new opportunities for drug development. At a
233 time when the challenges of antimicrobial-resistant microbes have never been more acute, these
234 data provide many new areas of investigation.
235
236

237 ## *Methods*

238

239 **Genome mining for bacteriocin clusters**

240

241 In total, 6,244 assembled pneumococcal genomes from studies previously published by us and
242 others[20,31,38,40,41,42,43,44,45,46] (Supplementary Table 4) were screened for the presence of bacteriocin
243 clusters using a variety of bioinformatic tools and databases, including antiSMASH[47], bactibase[48],
244 Bagel[49] and InterProScan[50] (data collected in March 2016). An in-house pipeline was developed to
245 automate part of this process. Predicted gene clusters from each of the database outputs were
246 examined manually and further scrutinised using extensive BLAST searches.

247

248 **Analyses of the putative bacteriocin genes**

249

250 Putative bacteriocin genes were annotated using homology to other known bacteriocin genes (not
251 shown) in all available databases mentioned above, as well as structure-based searches. Protein
252 domains were examined using the Conserved Domain search feature at NCBI[51]. Genes of interest
253 were screened against the STRING database[52] to search for any previously reported relationship to
254 other genes. Multiple sequence alignments of the genes in streptococcin clusters were performed in
255 Geneious version 9.1 (Biomatters Ltd) using the ClustalW algorithm[53] with default parameters (Gap
256 open cost=15, Gap extend cost=6.66). The multiple sequence alignment output was used within the
257 Geneious environment to calculate percentage identity matrices. The figures of the coding regions of
258 the bacteriocin clusters and their flanking genes were generated in Geneious and edited using
259 Inkscape (http://inkscape.org).

260

261 **Classification and nomenclature of bacteriocin clusters**

262

263 Putative bacteriocin clusters were classified based on their predicted biosynthetic machinery and
264 structural features according to Arnison *et al.*[19] and their corresponding genes were designated per
265 the standards of nomenclature for bacteriocins (Supplementary Table 1). BLAST searches in the NCBI
266 database revealed that the majority of these clusters are also present in other closely related
267 streptococci (Supplementary Table 2); therefore, the clusters were named using the prefix "strepto-"
268 followed by an abbreviation of their bacteriocin class: "streptococcins" for those belonging to the
269 lactococcin 972-like class, "streptolancidins" for lanthipeptides, "streptocyclicins" for the head-to-
270 tail cyclized peptides, "streptosactins" for the sactipeptides and "streptolassins" for the lassopeptide

271 group of bacteriocins. When more than one bacteriocin cluster from the same class was present,
272 they were lettered alphabetically by the order of their discovery in our analyses.
273
274 **Molecular epidemiology of the bacteriocin clusters**
275
276 We compiled a global and historical dataset (n=571) by selecting a diverse set of pneumococcal
277 genomes isolated between 1916 and 2009 from people of all ages residing in 39 different countries.
278 Samples from both carriage and disease, 88 pneumococcal serotypes and 99 different clonal
279 complexes were represented in this dataset (Supplementary Table 3). Genomes and their associated
280 metadata (multilocus sequence type, serotype, isolation date, country of isolation, etc.) were stored
281 in a BIGSdb database[54]. The BIGSdb database platform was used to generate a presence/absence
282 matrix of all the known bacteriocin genes in all 571 genomes. Using this matrix (79,940 genes) as an
283 input, an in-house python script was developed to calculate the prevalence, molecular epidemiology
284 and co-occurrence patterns of all the bacteriocin clusters in the study dataset.
285
286 **Construction of the core genome phylogenetic tree**
287
288 All genomes in the study dataset were annotated using the Prokka prokaryotic annotation pipeline[55].
289 The annotation files were input into Roary[56] and clustered using a sequence identity threshold of
290 90%. Genes present in every genome were selected using a core genome threshold of 100% and
291 were aligned using Roary. FastTreeMP[57] was used to construct the phylogenetic tree.
292 ClonalFrameML[58] was then applied to reconstruct the phylogenetic tree adjusted for recombination.
293 The tree was annotated using iTOL[59] and Inkscape.
294
295 **Investigation of bacteriocin cluster insertion sites**
296
297 Bacteriocin cluster sequences were used as queries to BLAST against genomes in the study dataset
298 using the custom BLAST implemented in Geneious. The matching region plus additional flanking
299 regions were visualised using the query-centric alignment feature within the Geneious environment.
300 Regions of DNA with different bacteriocin clusters but identical flanking genes among different
301 isolates were identified and further investigated using the Artemis Comparison Tool (ACT)[60]. Linear
302 comparison figures were generated using Geneious, ACT and Inkscape.
303

304 **RNA sequencing analyses**

305

306 An RNA sequencing dataset was generated in our lab from a pneumococcal strain 2/2 isolate
307 growing at a higher incubation temperature than normal (40 vs 37°C) to mimic heat stress[24] (NCBI
308 GEO accession number GSE103778). Pneumococci were grown in brain-heart infusion broth for 6
309 hours and RNA extractions were performed on samples from five time points (2, 3, 4, 5 and 6 hours
310 of incubation) using the Promega Maxwell® 16 Instrument and LEV simplyRNA Cells purification kit,
311 following the manufacturer's protocol. Extracted RNA samples were sent to the Oxford Genomics
312 Centre for sequencing. The sequenced forward and reverse reads were paired and mapped onto the
313 annotated pneumococcal strain 2/2 genome using Bowtie2[61] with the highest sensitivity option.
314 Differential gene expressions were assessed in Geneious using the DESeq[62] method. Genes with an
315 adjusted P value less than 0.05 were deemed to be differentially expressed.

316

317 **Data availability**

318

319 Representative examples of the newly-discovered bacteriocin clusters have been deposited at
320 GenBank under the accession numbers of MF990778-MF990796. Accession numbers for all genomes
321 used in this study are listed in Supplementary Table 4. Raw transcriptomic sequence data used in this
322 study is available under GEO accession no. GSE103778. Additional data that support the findings of
323 this study are available from the corresponding author upon request.

324

325

326 ## *References*

327

328 1. Bogaert, D., de Groot, R. & Hermans, P. *Streptococcus pneumoniae* colonisation: the key to
329 pneumococcal disease. *Lancet Infect Dis* **4**, 144-154 (2004).

330 2. O'Brien, K. *et al.* Burden of disease caused by *Streptococcus pneumoniae* in children younger than
331 5 years: global estimates. *Lancet* **374**, 893-902 (2009).

332 3. Klugman, K. Pneumococcal resistance to antibiotics. *Clin Microbiol Rev* **3**, 171-196 (1990).

333 4. Doern, G. *et al.* Antimicrobial resistance among clinical isolates of *Streptococcus pneumoniae* in
334 the United States during 1999-2000, including a comparison of resistance rates since 1994-1995.
335 *Antimicrob Agents Chemother* **45**, 1721-1729 (2001).

336 5. Tadesse, B. *et al.* Antimicrobial resistance in Africa: a systematic review. *BMC Infect Dis* **17**, (2017).

337 6. Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development
338 of new antibiotics. (World Health Organization, 2017);
339 http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria

340 7. Brueggemann, A. *et al.* Clonal relationships between invasive and carriage *Streptococcus*
341 *pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis*
342 **187**, 1424-1432 (2003).

343 8. Shak, J., Vidal, J. & Klugman, K. Influence of bacterial interactions on pneumococcal colonization of
344 the nasopharynx. *Trends Microbiol* **21**, 129-135 (2013).

345 9. Dawid, S., Roche, A. & Weiser, J. The *blp* bacteriocins of *Streptococcus pneumoniae* mediate
346 intraspecies competition both *in vitro* and *in vivo*. *Infect Immun* **75**, 443-451 (2006).

347 10. Czaplewski, L. *et al.* Alternatives to antibiotics—a pipeline portfolio review. *Lancet Infect Dis* **16**,
348 239-251 (2016).

349 11. Bogaardt, C., van Tonder, A. & Brueggemann, A. Genomic analyses of pneumococci reveal a wide
350 diversity of bacteriocins – including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics* **16**,
351 (2015).

352 12. Reichmann, P. & Hakenbeck, R. Allelic variation in a peptide-inducible two-component system of
353 *Streptococcus pneumoniae*. FEMS Microbiol Lett **190**, 231-236 (2000).

354 13. Guiral, S., Mitchell, T., Martin, B. & Claverys, J. Competence-programmed predation of
355 noncompetent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements. *Proc*
356 *Natl Acad Sci* USA **102**, 8710-8715 (2005).

357 14. Hoover, S. *et al.* A new quorum-sensing system (TprA/PhrA) for *Streptococcus pneumoniae* D39
358 that regulates a lantibiotic biosynthesis gene cluster. *Mol Microbiol* **97**, 229-243 (2015).

359 15. Maricic, N., Anderson, E., Opipari, A., Yu, E. & Dawid, S. Characterization of a multipeptide
360 lantibiotic locus in *Streptococcus pneumoniae*. *mBio* **7**, e01656-15 (2016).

361 16. Kadam, A. *et al.* Promiscuous signaling by a regulatory system unique to the pandemic PMEN1
362 pneumococcal lineage. *PLOS Pathog* **13**, e1006339 (2017).

363 17. Begley, M., Cotter, P., Hill, C. & Ross, R. Identification of a novel two-peptide lantibiotic,
364 lichenicidin, following rational genome mining for LanM proteins. *Appl Environ Microbiol* **75**, 5451-
365 5460 (2009).

366 18. Martínez, B., Rodríguez, A., Suárez, J. & Fernández, M. Synthesis of lactococcin 972, a bacteriocin
367 produced by *Lactococcus lactis* IPLA 972, depends on the expression of a plasmid-encoded
368 bicistronic operon. *Microbiology* **145**, 3155-3161 (1999).

369 19. Arnison, P. *et al.* Ribosomally synthesized and post-translationally modified peptide natural
370 products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108-160
371 (2013).

372 20. Kjos, M., Snipen, L., Salehian, Z., Nes, I. & Diep, D. The Abi proteins and their involvement in
373 bacteriocin self-immunity. *J Bacteriol* **192**, 2068-2076 (2010).

374 21. Perez-Pascual, D., Monnet, V. & Gardan, R. Bacterial cell–cell communication in the host via
375 RRNPP peptide-binding regulators. *Front Microbiol* **7**, (2016).

376 22. Son, M. *et al.* Conserved mutations in the pneumococcal bacteriocin transporter gene, *blpA*,
377 result in a complex population consisting of producers and cheaters. *mBio* **2**, e00179-11-e00179-11
378 (2011).

379 23. Brown, S., West, S., Diggle, S. & Griffin, A. Social evolution in micro-organisms and a Trojan horse
380 approach to medical intervention strategies. *Phil Trans Royal Soc B* **364**, 3157-3168 (2009).

381 24. Kurioka, A. *et al.* Diverse *Streptococcus pneumoniae* strains drive a MAIT cell response through
382 MR1-dependent and cytokine-driven pathways. (2017). doi:10.1101/158204

383 25. Crook, DW. *et al.* Pneumococcal carriage. In *The Pneumococcus* (eds. Tuomanen, EI., Mitchell, TJ.,
384 Morrison, DA. & Spratt, B.) 136–147 (ASM Press, Washington (D.C.), USA, 2004).

385 26. Auranen, K., Mehtälä, J., Tanskanen, A. & S. Kaltoft, M. Between-strain competition in acquisition
386 and clearance of pneumococcal carriage—epidemiologic evidence from a longitudinal study of day-
387 care children. *Am J Epidemiol* **171**, 169-176 (2009).

388 27. Garcia-Rodriguez, J. & Fresnadillo Martínez, M. Dynamics of nasopharyngeal colonization by
389 potential respiratory pathogens. *J Antimicrob Chemother* **50**, 59-74 (2002).

390 28. Huang, S. *et al.* Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts
391 communities, 2001 and 2004. *Pediatrics* **116**, e408-e413 (2005).

392 29. Kaplan, S. *et al.* Serotype 19A is the most common serotype causing invasive pneumococcal
393 infections in children. *Pediatrics* **125**, 429-436 (2010).

394 30. Aguiar, S., Brito, M., Gonçalo-Marques, J., Melo-Cristino, J. & Ramirez, M. Serotypes 1, 7F and
395 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7
396 years of heptavalent conjugate vaccine use. *Vaccine* **28**, 5167-5173 (2010).

397 31. Coffey, T. *et al.* Horizontal transfer of multiple penicillin-binding protein genes, and capsular
398 biosynthetic genes, in natural populations of *Streptococcus pneumoniae*. *Mol Microbiol* **5**, 2255-2260
399 (1991).

400 32. Dowson, C., Coffey, T. & Spratt, B. Origin and molecular epidemiology of penicillin-binding-
401 protein-mediated resistance to β-lactam antibiotics. *Trends Microbiol* **2**, 361-366 (1994).

402 33. Wyres, K. *et al.* Pneumococcal capsular switching: a historical perspective. *J Infect Dis* **207**, 439-
403 449 (2012).

404 34. Wyres, K. *et al.* The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success.
405 *Genome Biol* **13**, R103 (2012).

406 35. Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant formed
407 through a multi-fragment recombination event. *Nat Genet* **44**, 352-355 (2012).

408 36. Johnston, C., Polard, P. & Claverys, J. The DpnI/DpnII pneumococcal system, defense against
409 foreign attack without compromising genetic exchange. *Mob Genet Elements* **3**, e25582 (2013).

410 37. Eutsey, R. *et al.* Genetic stabilization of the drug-resistant PMEN1 pneumococcus lineage by its
411 distinctive DpnIII restriction-modification system. *mBio* **6**, e00173-15 (2015).

412 38. Brueggemann, A. *et al.* Pneumococcal prophages are diverse, but not without structure or
413 history. *Sci Rep* **7**, 42976 (2017).

414 39. Cuevas, R. *et al.* A novel streptococcal cell-cell communication peptide promotes pneumococcal
415 virulence and biofilm formation. *Mol Microbiol* **105**, 554-571 (2017).

416 40. Croucher, N. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science*
417 331, 430-434 (2011).

418 41. Croucher, N. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology.
419 *Nat Genet* **45**, 656-663 (2013).

420 42. Croucher, N. *et al.* Dominant role of nucleotide substitution in the diversification of serotype 3
421 pneumococci over decades and during a single infection. *PLOS Genet* **9**, e1003868 (2013).

422 43. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal
423 recombination. *Nat Genet* **46**, 305-309 (2014).

424 44. van Tonder, A. *et al.* Defining the estimated core genome of bacterial populations using a
425 Bayesian decision model. *PLOS Comput* **10**, e1003788 (2014).

426 45. van Tonder, A. *et al.* Genomics reveals the worldwide distribution of multidrug-resistant serotype
427 6E pneumococci. *J Clin Microbiol* **53**, 2271-2285 (2015).

428 46. Gladstone, R. *et al.* Five winters of pneumococcal serotype replacement in UK carriage following
429 PCV introduction. *Vaccine* **33**, 2015-2021 (2015).

430 47. Weber, T. *et al.* antiSMASH 3.0—a comprehensive resource for the genome mining of
431 biosynthetic gene clusters. *Nucleic Acids Res* **43**, W237-W243 (2015).

432 48. Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J. & Fliss, I. BACTIBASE second release: a
433 database and tool platform for bacteriocin characterization. *BMC Microbiol* **10**, 22 (2010).

434 49. de Jong, A., van Hijum, S., Bijlsma, J., Kok, J. & Kuipers, O. BAGEL: a web-based bacteriocin
435 genome mining tool. *Nucleic Acids Res* **34**, W273-W279 (2006).

436 50. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**,
437 1236-1240 (2014).

438 51. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for the functional annotation of
439 proteins. global-priority-list-antibiotic-resistant **39**, D225-D229 (2010).

440 52. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of
441 life. *Nucleic Acids Res* **43**, D447-D452 (2014).

442 53. Thompson, J., Gibson, T. & Higgins, D. Multiple sequence alignment using ClustalW and ClustalX.
443 *Curr Protoc Bioinformatics* (2002). doi:10.1002/0471250953.bi0203s00

444 54. Jolley, K. & Maiden, M. BIGSdb: scalable analysis of bacterial genome variation at the population
445 level. *BMC Bioinform* **11**, 595 (2010).

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

447 56. Page, A. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-
448 3693 (2015).

449 57. Price, M., Dehal, P. & Arkin, A. FastTree 2 – approximately maximum-likelihood trees for large
450 alignments. *PLOS ONE*, e9490 (2010).

451 58. Didelot, X. & Wilson, D. ClonalFrameML: efficient inference of recombination in whole bacterial
452 genomes. *PLOS Compu.* **11**, e1004041 (2015).

453 59. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and
454 annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**, W242-W245 (2016).

455 60. Carver, T. *et al.* ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422-3423 (2005).

456 61. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359
457 (2012).

458 62. Anders, S., Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Nat*
459 *Preced* (2010). doi:10.1038/npre.2010.4282.1

460 63. Walker, G., Wescombe, P., Tagg, J., Carne, A. & Heng, N. Salivaricin E and abundant dextranase
461 activity may contribute to the anti-cariogenic potential of the probiotic candidate *Streptococcus*
462 *salivarius* JH. *Microbiology* **162**, 476-486 (2016).

463 64. Croucher, N. *et al.* Role of conjugative elements in the evolution of the multidrug-resistant
464 pandemic clone *Streptococcus pneumoniae*Spain23F ST81. *J Bacteriol* **191**, 1480-1489 (2008).

465

466 Author contributions

467 R.R.J. designed the genome mining pipelines. R.R.J. A.J.vT., J.P.K. and A.B.B. analysed the data. R.R.J.
468 and A.B.B. wrote the manuscript. All authors reviewed the manuscript.

469

470 Competing financial interests

471 The authors declare no competing financial interests.

Table 1. Bacteriocin clusters identified among a dataset of 571 pneumococci recovered since 1916 from patients of all ages residing in 39 different countries.

Bacteriocin	Genome(s)			Year(s) of Isolation	Countries (n)	CCs[a] (n)	Serotypes (n)
	Complete	Partial	Total				
streptococcin A	374 (65.5%)	31 (5.4%)	405 (70.9%)	1916 - 2009	36	80	76
streptococcin B	415 (72.7%)	156 (27.3%)	571 (100%)	1916 - 2009	39	99	88
streptococcin C	571 (100%)	0 (0.0%)	571 (100%)	1916 - 2009	39	99	88
streptococcin D	6 (1.1%)	0 (0.0%)	6 (1.1%)	1968 - 2005	5	3	4
streptococcin E	74 (13.0%)	496 (86.9%)	570 (99.8%)	1916 - 2009	39	98	88
streptolancidin A[b]	1 (0.2%)	1 (0.2%)	2 (0.4%)	1972 - 2006	2	2	2
streptolancidin B[c]	45 (7.9%)	48 (8.4%)	93 (16.3%)	1939 - 2006	16	11	10
streptolancidin C	73 (12.8%)	213 (37.3%)	286 (50.1%)	1937 - 2009	27	52	52
streptolancidin D	49 (8.6%)	0 (0.0%)	49 (8.6%)	1938 - 2006	13	13	11
streptolancidin E[d]	3 (0.5%)	161 (28.2%)	164 (28.7%)	1937 - 2009	25	21	38
streptolancidin F	23 (4.0%)	0 (0.0%)	23 (4.0%)	1937 - 2006	7	4	11
streptolancidin G[e]	186 (32.6%)	13 (2.3%)	199 (34.9%)	1916 - 2009	25	38	46
streptolancidin H	1 (0.2%)	0 (0.0%)	1 (0.2%)	2006-2008	1	0	1
streptolancidin I	1 (0.2%)	0 (0.0%)	1 (0.2%)	2009	1	1	1
streptolancidin J	140 (24.5%)	237 (41.5%)	377 (66.0%)	1916 - 2009	33	64	74
streptolancidin K	1 (0.2%)	0 (0.0%)	1 (0.2%)	2009	1	0	1
streptocyclicin[f]	205 (35.9%)	4 (0.7%)	209 (36.6%)	1937 - 2009	20	46	59
streptolassin	20 (3.5%)	0 (0.0%)	20 (3.5%)	1939 - 1996	8	7	10
streptosactin	1 (0.2%)	0 (0.0%)	1 (0.2)	2009	1	1	1
cib[g]	551 (96.5%)	14 (2.5%)	557 (97.5%)	1916 - 2009	39	97	88
blp[h]	N/A	N/A	571 (100%)	1916 - 2009	39	99	88

a. CC, clonal complex. Singletons (single genotypes with no closely related variant) were excluded from the count. Synonym(s) for the previously identified bacteriocins are as follow [reference]: **b**. pneumolancidin [15] and salivaricin E [63]; **c**. lcpAMT [16] and ICESp23FST81 lantibiotic [64]; **d**. SP23-BS72 lantibiotic [17]; **e**. phr lantibiotic [14]; **f**. pneumocyclicin [11]; **g**. cibAB [13]; **h**. spi [11] and pnc [11].

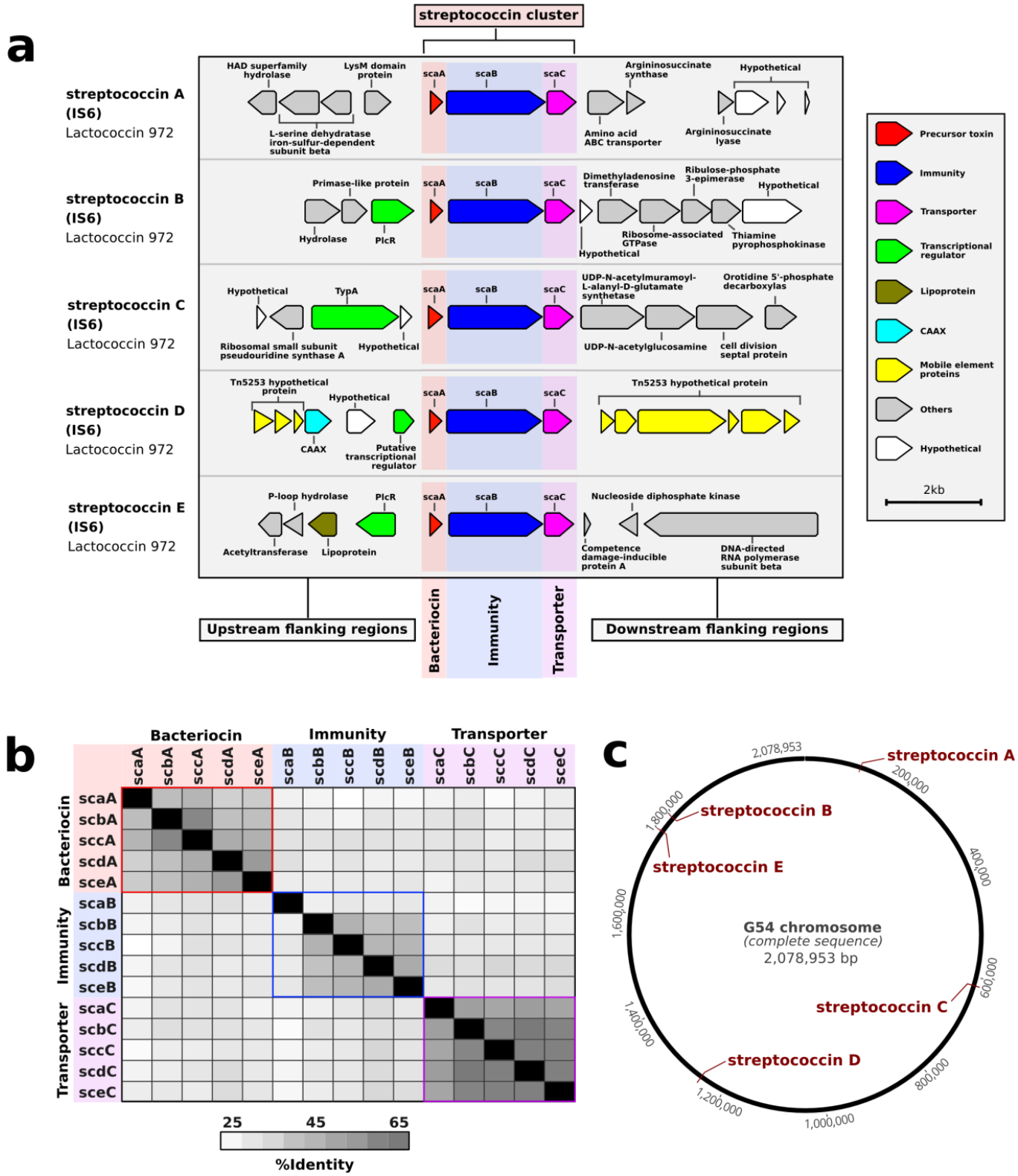


Figure 1 | Discovery of streptococcin bacteriocins in the pneumococcus. a, A schematic representation of streptococcin clusters and their flanking regions. The coding regions were derived from the genome of pneumococcal strain IS6, which harbours five different complete streptococcin clusters. **b**, A distance matrix depicting the nucleotide sequence identity shared between genes of different streptococcin clusters found in pneumococcal strain IS6. **c**, Schematics of the finished genome of pneumococcal strain G54 with the locations of the five streptococcin clusters highlighted in red.

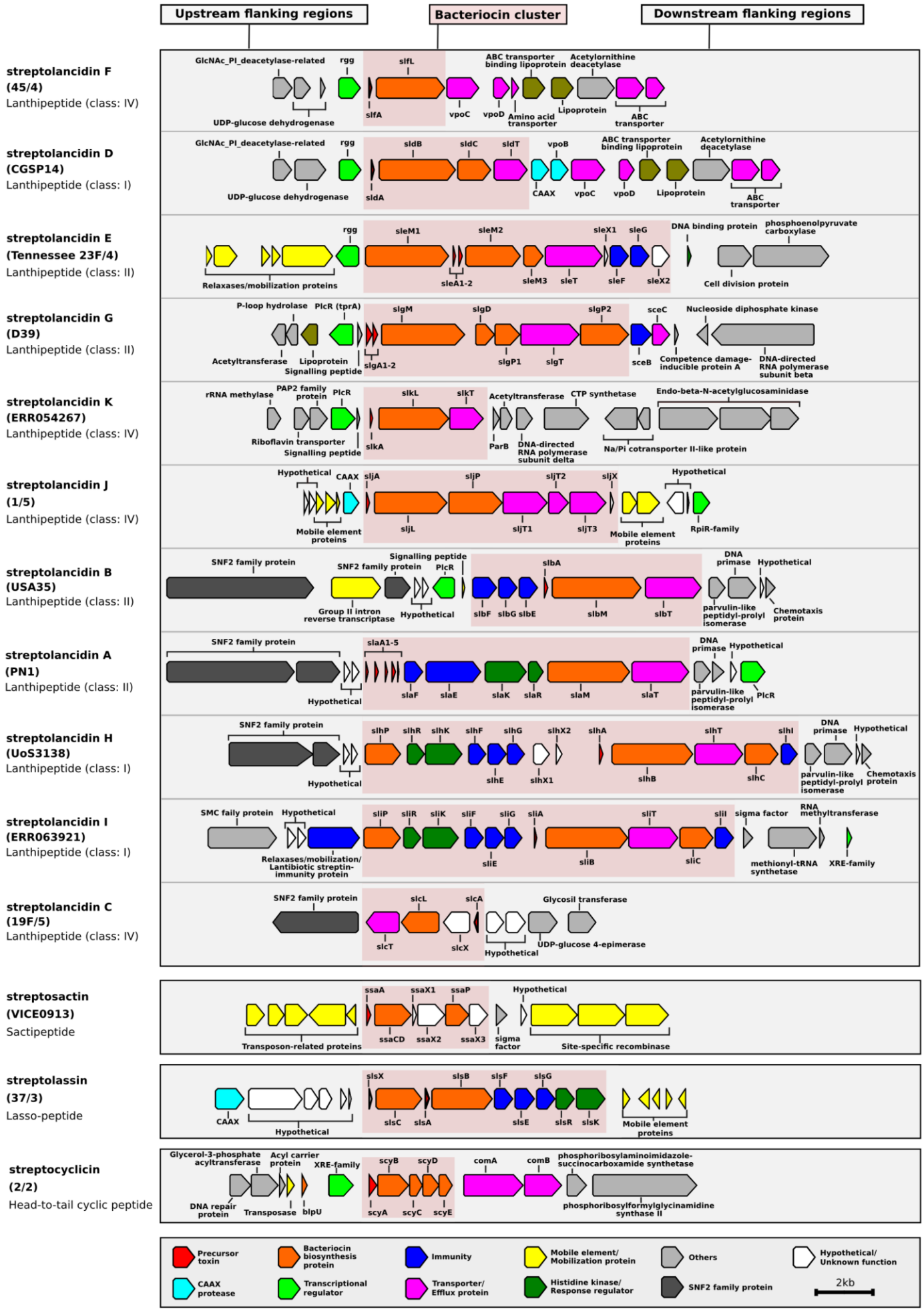


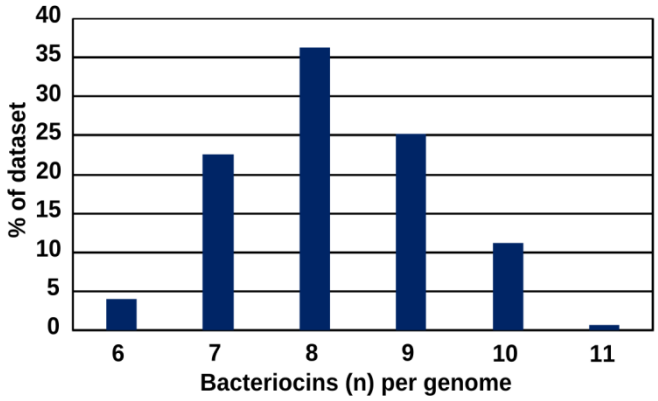
Figure 2 | Genetic composition of streptolancidin, streptocyclicin, streptosactin, and streptolassin bacteriocins in the pneumococcal genomes. A schematic representation of 11 different streptolancidins, one streptocyclicin, one

streptosactin and one streptolassin identified among pneumococcal genomes. The names of isolate containing each bacteriocin cluster are given in brackets. The family of each bacteriocin cluster is given underneath the names of isolates.

a

	streptococcin A	*streptococcin B*	*streptococcin C*	*streptococcin D*	*streptococcin E*
ABC	374	415	571	6	74
AB	0	0	0	0	0
AC	0	0	0	0	0
BC	31	154	0	0	300
A	0	0	0	0	0
B	0	0	0	0	0
C	0	2	0	0	196
Total	**405**	**571**	**571**	**6**	**570**

b



c

blp	*cib*	*streptococcin A*	*streptococcin B*	*streptococcin C*	*streptococcin E*	*streptocyclicin*	*streptolancidin B*	*streptolancidin C*	*streptolancidin D*	*streptolancidin E*	*streptolancidin F*	*streptolancidin G*	*streptolancidin J*	*CC (n)*	*Serotypes (n)*	*Countries (n)*	*Isolation dates*	*Genomes (n)*	*% of dataset*
■	■	■	■	■	■	■	□	□	□	□	□	□	■	13	19	8	1938 - 2008	44	7.7
■	■	■	■	■	■	□	□	■	□	□	□	□	■	12	16	12	1939 - 2006	40	7.0
■	■	■	■	■	■	■	□	□	□	□	□	■	■	8	12	3	1939 - 2007	15	2.6
■	■	■	■	■	■	□	□	■	□	□	□	□	□	7	9	11	1952 - 2007	48	8.4
■	■	■	■	■	■	□	□	□	□	□	□	□	□	6	6	5	1939 - 1989	8	1.4
■	■	■	■	■	■	□	□	□	□	□	□	■	■	5	7	3	1916 - 2004	10	1.8
■	■	■	■	■	■	□	□	■	■	□	□	□	□	4	4	11	1967 - 2006	26	4.6
■	■	■	■	■	■	□	□	□	□	□	□	□	■	3	10	6	1946 - 2003	18	3.2
■	■	□	■	■	■	■	□	□	□	■	□	□	■	3	6	5	1968 - 2005	14	2.5
■	■	■	■	■	■	■	□	□	□	■	□	■	■	3	9	4	1939 - 2007	12	2.1
■	■	■	■	■	■	□	□	■	□	□	□	■	■	3	5	3	1952 - 2006	11	1.9
■	■	□	■	■	■	□	□	■	□	□	□	□	□	3	4	3	1962 - 2004	7	1.2
■	■	□	■	■	■	■	□	■	□	□	□	■	■	2	4	2	1952 - 1988	24	4.2
■	■	■	■	■	■	■	□	■	□	□	□	□	■	2	2	2	1989 - 1992	11	1.9
■	■	■	■	■	■	■	□	□	□	■	□	□	■	2	5	3	1939 - 2005	8	1.4
■	■	□	■	■	■	□	□	■	□	□	□	□	■	2	3	2	1950 - 2002	7	1.2
■	■	□	■	■	■	□	■	□	□	■	□	■	■	1	3	7	1987 - 2005	27	4.7
■	■	■	■	■	■	□	■	□	□	■	□	■	■	1	5	7	1970 - 2006	23	4.0
■	■	□	■	■	■	□	□	■	□	■	□	□	□	1	2	1	1962 - 1966	13	2.3
■	■	■	■	■	■	■	□	□	□	■	□	□	□	1	2	3	1945 - 1991	12	2.1
■	■	□	■	■	■	□	■	■	□	□	□	□	■	1	1	2	1978 - 1985	11	1.9
■	□	■	■	■	■	□	□	■	□	□	■	□	■	1	2	6	1937 - 2006	9	1.6
■	■	□	■	■	■	■	□	□	□	□	□	□	■	1	1	3	1989 - 2005	6	1.1

■ **Present (complete or partial)** □ **Absent**

Figure 3 | The prevalence, molecular epidemiology and co-occurrence patterns of bacteriocins within the global pneumococcal population. **a**, The gene patterns (A, toxin; B, immunity; C, transporter) of different streptococcins within the study dataset. **b**, The number of bacteriocins (of any type) per genome among the 571 genomes. **c**, Epidemiological characteristics of the pneumococci that possessed bacteriocin combinations found in >1 percent of the genomes in the dataset.

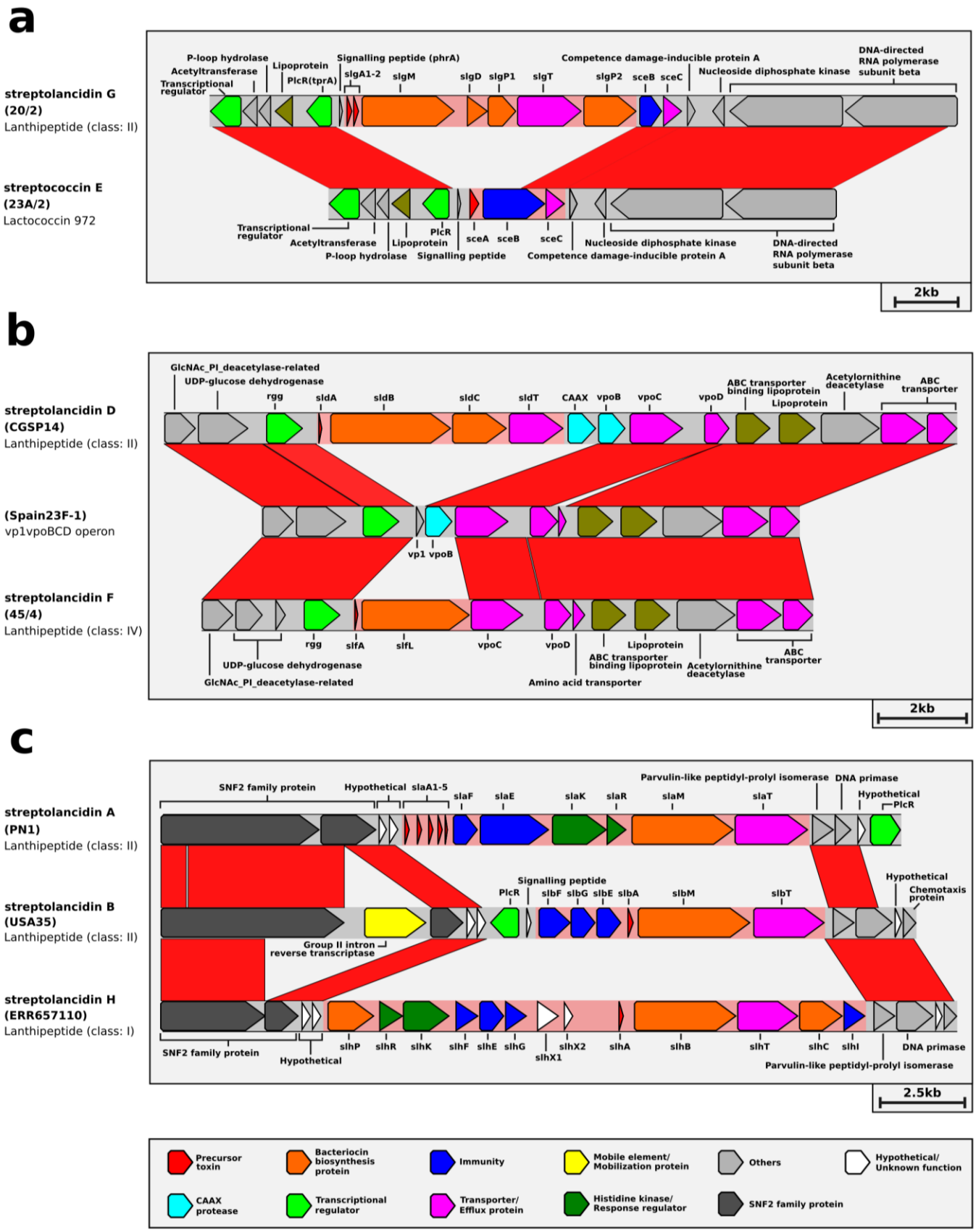


Figure 4 | Whole-genome-based population analysis reveals a bacteriocin switching mechanism. Bacteriocin cluster hotspots (BCHs) were defined as regions of DNA where different bacteriocin clusters with identical flanking genes were found among pneumococcal genomes. Linear comparisons of BCH-1 (**a**), BCH-2 (**b**) and BCH-3 (**c**) are shown. The isolate names are given in brackets. The family of each bacteriocin cluster is given underneath the isolate name.

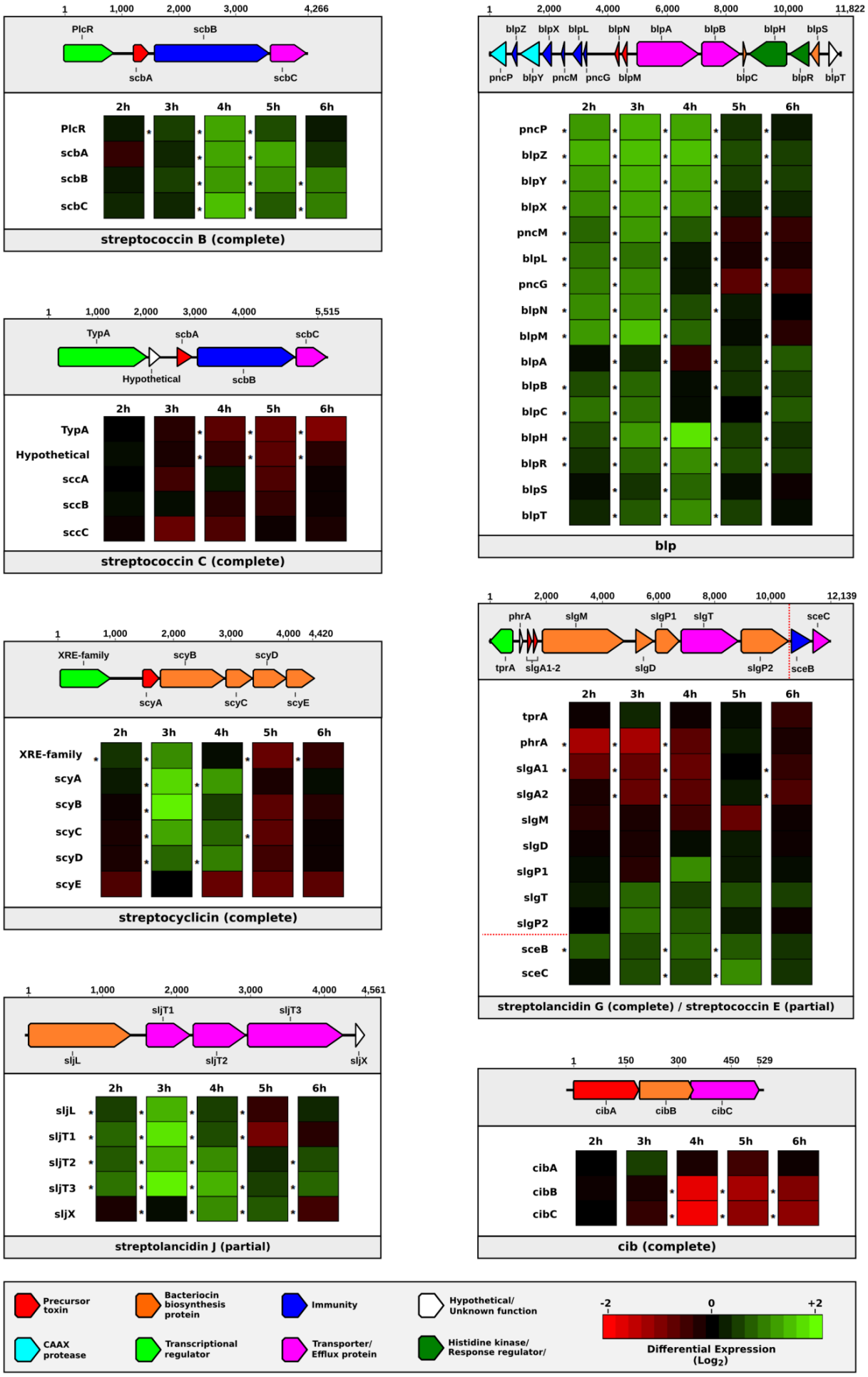


Figure 5 | Dynamic changes in the expression of bacteriocin genes in response to heat stress. The differential expression levels of bacteriocin gene clusters found in pneumococcal strain 2/2 are displayed in individual boxes. A

schematic representation of each bacteriocin cluster is provided above each box. Genes are represented by rows and differential expression levels at different time points are indicated in columns. An asterisk to the left of a cell indicates a statistically significant differential expression level ($p < 0.05$). A streptolancidin C cluster was also present in the reference genome, but was excluded from this analysis due to sequence/assembly issues.