Identification of evolutionarily conserved virulence factor by selective pressure analysis of *Streptococcus pneumoniae*

Masaya Yamaguchi[a*], Kana Goto[a, b], Yujiro Hirose[a], Yuka Yamaguchi[a], Tomoko Sumitomo[a], Masanobu Nakata[a], Kazuhiko Nakano[b], Shigetada Kawabata[a]

[a]Department of Oral and Molecular Microbiology, Osaka University Graduate School of Dentistry, Suita, Osaka, Japan

[b]Department of Pediatric Dentistry, Osaka University Graduate School of Dentistry, Suita, Osaka, Japan

*Address correspondence to Masaya Yamaguchi, yamaguchi@dent.osaka-u.ac.jp

Running title: Identification of virulence factor by evolutionary analysis

Keywords: *Streptococcus pneumoniae*, Evolutionary analysis, Choline-binding protein, Evolutionary conservation

Abstract

Streptococcus pneumoniae is a pathogen that poses one of the greatest threats to human health, causing pneumonia, sepsis, and meningitis. Here, we investigated the evolutionary selective pressures on 16 pneumococcal choline-binding cell-surface proteins. Phylogenetic and molecular analyses revealed that the *cbpJ* and *lytA* genes had the highest codon rates under negative selection among those examined. Although LytA is an important virulence factor inducing pneumococcal autolysis, the role of CbpJ in pneumococcal pathogenesis is unknown. We therefore examined whether CbpJ acts as a virulence factor *in vitro* and *in vivo*. In neutrophil bactericidal assays, the *cbpJ* mutant showed reduced survival as compared to wild-type (WT) *S. pneumoniae*. Mice intranasally infected with a *cbpJ* mutant strain showed a lower bacterial burden in bronchoalveolar lavage fluid and higher survival rate than those infected with WT cells, whereas no differences were observed between strains by intravenous infection. Thus, molecular evolutionary analysis is a powerful tool that reveals the importance of virulence factors in real-world infection and transmission, since calculations are performed based on bacterial genome diversity following transmission of infection in an uncontrolled population. In addition, evolutionarily conserved virulence factors are candidate therapeutic targets or vaccine antigens.

34 Improper use of antibiotics creates evolutionary pressures that drive bacteria to acquire drug
35 resistance by natural mutation and/or horizontal transfer of resistance genes. This is a major
36 public health threat: it is estimated that drug-resistant infections cause 10 million deaths
37 annually and may result in economic losses reaching 100 trillion US dollars by 2050[1].
38 However, a target-to-hit screen typically requires 24 works-in-process and 94 million US
39 dollars, and the baseline total cost is 1.8 billion US dollars over 13 years to launch a new
40 drug[2]. In fact, the number of new antibiotics developed and approved has steadily decreased
41 in the past three decades, leaving fewer options for treating resistant bacteria[3].
42 *Streptococcus pneumoniae* is one of the pathogens posing the greatest threat to human
43 health[4,5]. *S. pneumoniae* belongs to the mitis group[6,7] and is a major cause of pneumonia,
44 sepsis, and meningitis[8,9]. In 2015, pneumococcal pneumonia caused over 1.5 million deaths
45 in individuals of all ages, and this rate increased in people over 70 years old between 2005
46 and 2015[10], which is especially problematic since the elderly population is growing in many
47 parts of the world. Although pneumococcal conjugate vaccines have considerable benefits,
48 non-vaccine pneumococcus serotypes have increased worldwide[11,12].
49 The conflict between the host immune system and pathogens leads to an evolutionary
50 arms races known as the "Red Queen" scenario[13,14]. Protein regions at the host–pathogen

interface are subjected to the strongest selective pressure and thus evolve under positive selection. Adaptive evolution has been reported in genes related to the mammalian immune system such as pattern recognition receptors[14]. In contrast, bacterial proteins are under negative/purifying selection, which is important for the survival and/or success of the species since non-synonymous substitution of codons can lead to lineage extinction (Fig. 1). Phylogenetic and molecular evolutionary analyses can reveal the number of codons under negative/purifying selection in a species, since alterations in amino acid residues in regions under negative selective pressure are not allowed. Thus, drugs targeting these regions would be less likely to promote the development of resistance through natural mutation.

We analysed pneumococcal choline-binding proteins (CBPs) localised on the bacterial cell surface through interaction with choline-binding repeats and phosphoryl choline on the cell wall. *S. pneumoniae* harbours various CBPs including *N*-acetylmuramoyl L-alanine amidase (LytA), which induces pneumococcal-specific autolysis[15]. Although several CBPs have been characterised, their phylogenetic relationships remain unclear and the unclassified gene names are confusing. We first analysed the distribution of genes encoding CBPs based on pneumococcal genome sequences. Orthologues of genes in each strain were identified by phylogenetic analysis. We then calculated the evolutionary selective pressure on each codon

68 from phylogenetic trees and aligned sequences. We found that the *cbpJ* gene contains the

69 highest rate of codons under negative selection. CbpJ has no known functional domains

70 except signal sequences and choline-binding repeats, and its role in pneumococcal

71 pathogenesis is unclear. Functional analyses revealed that CbpJ contributes to evasion of host

72 neutrophil-mediated killing in pneumococcal pneumonia. Thus, evolutionary analysis

73 focusing on negative selection can reveal novel virulence factors.

Results

Distribution of *cbp* genes among pneumococcal strains

Genes encoding CBPs among pneumococcal strains were extracted by tBLASTn search (Supplementary Table 1). Some genes were re-annotated since the search results showed that certain homologous regions were not matched to annotated open reading frames (ORFs). In strain SPNA45, the *SPNA_01670* gene contains both predicted promoter regions and intact ORF structures of the *cbpF* and *cbpJ* genes. On the other hand, *cbpG*-homologous regions in strains R6, D39, SPN034183, SPN994038, and SPN994039 did not contain promoters (Supplementary Table 1 and Supplementary Table 2). Orthologous relationships of each gene were analysed. The distribution of *cbp* genes was unrelated to capsular serotypes. Four genes—i.e., *lytA*, *lytB*, *cbpD*, and *cbpE*—were conserved as intact ORFs in all 28 pneumococcal strains (Fig. 2A). Other *cbp* genes contained frameshift mutations in the orthologues or were absent in some strains.

Phylogenetic relationships in pneumococcal CBPs

Phylogenetic relationships of genes encoding CBPs in pneumococcal species are confusing since some genes in the same cluster show high similarity to each other. To clarify the

91 relationships, we compared common nucleotide sequences among genes encoding CBPs in
92 strain TIGR4. Maximum likelihood and Bayesian phylogenetic analyses revealed two
93 common clusters: one comprising *cbpF*, *cbpG*, *cbpJ*, *cbpK*, and *cbpC*, and the other *lytA*, *lytB*,
94 *lytC*, *cbpL*, and *cbpE* (Fig. 2B and Supplementary Fig. 1). The names of some CBP genes
95 were not consistent with those of phylogenetically related genes. In particular, *cbpF*, *cbpG*,
96 *cbpJ*, and *cbpK* were located close to each other in pneumococcal genomes and showed high
97 similarity. We thus defined orthologous genes in each pneumococcal strain based on
98 maximum likelihood and Bayesian phylogenetic analyses (Fig. 3 and Supplementary Fig. 2).
99 The gene locus tag numbers in orthologous relationships are shown in Supplementary Table 1.
100 The sequence similarity of *cbpF*, *cbpG*, *cbpJ*, and *cbpK* and their close proximity within
101 genomes indicated that a common ancestral *S. pneumoniae* acquired the genes by gene
102 duplication. Phylogenetic trees showed well-separated clusters of each gene. These
103 independent relationships indicated that horizontal gene transfer did not contribute to the
104 spread of *cbpF*, *cbpG*, *cbpJ*, and *cbpK* in *S. pneumoniae* species, despite their ability to take
105 up exogenous DNA. The genetic diversity of these genes may have been established by
106 accumulation of natural mutations during pneumococcal transmission.
107

108 **Evolutionary selective pressures on each of the CBP codons**

109 To evaluate the significance of CBPs in real-life infection and transmission, we performed
110 molecular evolutionary calculations based on bacterial genome diversity established after
111 transmission of infection in an uncontrolled population. The nucleotide sequences of each
112 CBP were aligned by codon, and conserved common codons were used for phylogenetic
113 analysis (Supplementary Fig. 3). The selective pressure on each gene was calculated based on
114 the phylogenetic trees and aligned sequences (Table 1). The rates of codons under negative
115 selection are visualised in Supplementary Figure 4. Over 13% of total codons in *cbpJ* and
116 *lytA* genes were under negative selection as compared to less than 5% for other *cbp* genes,
117 indicating that these genes play an important role in the success of *S. pneumoniae* species. On
118 the other hand, the *pspA* gene encoding the genetically divergent virulence factor PspA
119 contained fewer evolutionarily conserved codons, but had the highest numbers of codons
120 under positive pressure. Additionally, there were no evolutionarily conserved codons in the
121 *cbpG*, *cbpC*, and *cbpL* genes. The latter two had no common codons since a few genes had
122 frameshift mutations. When we re-calculated selective pressure without these genes, we
123 found a low rate of codons under negative selection among CBP-encoding genes
124 (Supplementary Table 3).

125

126 **CbpJ acts as a virulence factor in pneumococcal pneumonia**

127 While CbpJ had the highest rate of codons under negative selection among pneumococcal

128 CBPs, it has no known functional domains except a choline-binding repeat in its amino acid

129 sequence; moreover, its role in pneumococcal pathogenesis is unknown. In contrast, CbpL

130 showed limited numbers of evolutionarily conserved codons even after the above-described

131 adjustment. Domain structures and codons of CbpJ and CbpL under negative selection are

132 shown in Figure 4A. To assess the roles of CbpJ and CbpL in pneumococcal pathogenesis, we

133 generated mutant strains deficient in the corresponding genes. We first performed a blood

134 bactericidal assay to investigate the role of CbpJ and CbpL in sepsis. The survival rates of

135 *ΔcbpJ* and *ΔcbpL* strains in mouse blood were comparable to that of the wild-type (WT)

136 strain (Fig. 4B). We also carried out a human neutrophil bactericidal assay without serum.

137 Strain *ΔcbpJ* had a lower survival rate than the WT. These results suggest that CbpJ

138 contributes to the evasion of neutrophil-mediated killing (Fig. 4C).

139 In a mouse intravenous infection model, the survival rates of *ΔcbpL*- and

140 *ΔcbpJ*-infected mice did not differ significantly from that of mice infected with WT *S.*

141 *pneumoniae* (Fig. 5A). We also found that incubation of *S. pneumoniae* in human plasma for

142 3 h inhibited the expression of *cbpL* and *cbpJ*, as determined by quantitative real-time

143 (q)PCR (Fig. 5B). On the other hand, mice intranasally infected with strain *ΔcbpJ* showed an

144 improved survival rate as compared to those infected with WT *S. pneumoniae*; although a

145 similar tendency was observed for *ΔcbpL*-infected relative to WT mice, the difference was

146 not statistically significant (Fig. 5C). The number of bacteria in bronchoalveolar lavage fluid

147 (BALF) from *ΔcbpJ*-infected mice was lower than that in BALF from *ΔcbpL*- and

148 WT-infected mice (Fig. 5D). We also examined whether CbpL or CbpJ contributes to the

149 association of *S. pneumoniae* to alveolar epithelial cells and found that WT *S. pneumoniae* as

150 well as *ΔcbpL* and *ΔcbpJ* mutant strains did not differ in their ability to adhere to A549

151 human alveolar epithelial cells (Fig. 5E). However, *S. pneumoniae* WT strain exhibited

152 extensive inflammatory cell infiltration and bleeding as compared to the *ΔcbpJ* strain.

153 Histological examination of lung tissue from intranasally-infected mice showed that *ΔcbpJ*

154 induced milder inflammation than the WT strain. Lung tissue from Δ*cbpL*-infected mice

155 showed moderate inflammation (Fig. 5F). These results indicate that CbpJ acts as a

156 pneumococcal virulence factor in lung infection by contributing to the evasion of

157 neutrophil-mediated killing, whereas CbpJ has no role in bacterial survival in blood. In

158 addition, *cbpL* deficiency in strain TIGR4 did not significantly attenuate pathogenesis in

159 mouse lung and blood infection.

160 **Discussion**

161 In this study, we investigated evolutionarily conserved rates of CBP codons since these cell
162 surface proteins directly interact with the external environment, which induces rapid rates of
163 evolution in genes involved in genetic conflicts[14]. Evolutionary analysis based on
164 phylogenetic relationships can reveal regions in which the encoded amino acids are not
165 allowed to change even under selective pressure. The genetic diversity of *S. pneumoniae*
166 isolated from patients was the result of transmission in a real population. Thus, the
167 evolutionary conservation rate is a parameter that reflects the importance of the protein in
168 human infection. Although so-called arms races involve both the host and bacteria, most
169 studies of genetic diversity have focused on the former[14,16-18]. For example, evolutionary
170 studies based on inter-species comparisons have shown that most of the positive selection
171 targets in host receptors are located in regions that are responsible for direct interactions with
172 pathogens. Our study focused on negative selection targets in bacterial surface proteins
173 through an evolutionary analysis based on intra-species comparisons. This approach enabled
174 us to estimate the contribution of bacterial proteins to species success throughout the life
175 cycle, including inside the host and during the transmission phase.

176 We previously detected bacterial virulence factors by function prediction—e.g., by

177 searching for conserved motif/domains, constructing random transposon libraries, or

178 analysing the biochemical properties of the pathogen[19-23]. Although these laboratory-based

179 approaches are valuable, they are time-consuming and costly, and may not yield the expected

180 results. It is useful to examine the correlation between a target molecule and clinical features

181 as this can minimise the time and cost required for analysis. Furthermore, in basic studies of

182 bacterial pathogens, animal infection models are often used to determine whether a bacterial

183 molecule acts as a virulence factor. Although this is the best means of obtaining *in vivo*

184 information, it is unclear how accurately it reflects the clinical condition in humans.

185 Combining an evolutionary analysis and an animal model would be highly effective for

186 evaluating the functional significance of a putative virulence factor.

187 Genome-wide association study (GWAS) is a powerful tool for identifying the

188 relationship between genetic variants—mainly single nucleotide polymorphisms

189 (SNPs)—and phenotype, such as in diseases. As GWAS has become more prevalent, various

190 programs and software packages have been developed for this purpose[24,25]. On the other hand,

191 this approach has certain limitations, including the requirements for an appropriate control

192 group and detailed information regarding phenotype. In infectious diseases, it can be difficult

193 to quantify clinical features recorded at different medical centres. Furthermore, in the case of

194 most pathogens, there are no natural attenuated or avirulent strains that can serve as a control
195 group. Our evolutionary analysis has the advantage that it can be performed with genomic
196 information of pathogenic strains only by assuming the presence of pathogens as a phenotype
197 evading natural selection. Since synonymous and non-synonymous substitutions are
198 estimated to occur with equal probability under no selective pressure, a population in which
199 the latter has resulted in extinction by natural selection can serve as a control group. While
200 we have shown in the current study that evolutionary analysis with a small population has the
201 power to detect evolutionarily conserved proteins, a larger population would allow a
202 higher-resolution analysis, including the detection of conserved regions in some pathogenic
203 strains isolated from a specific site of infection or pathological condition. Since this analysis
204 involves simultaneous processing of aligned nucleotide and amino acid sequences, more
205 information is obtained from only SNPs extracted from nucleotide sequences. In addition,
206 automatised phylogenetic and evolutionary analyses are needed to analyse a large population.
207 Therefore, the development of software packages for meta-data is expected to aid widespread
208 application of this analytical approach.
209 There are some limitations to our evolutionary analysis. Firstly, although it can detect
210 evolutionarily conserved proteins, it cannot identify diverse virulence factors such as PspA

211 and CbpA within species[26-28]. Similarly, virulence factors recently acquired by horizontal
212 gene transfer have not been under selective pressure for a sufficiently long period of time to
213 perform this analysis. In addition, the high rate of codons under negative selection indicate
214 their universal importance in bacterial species. In other words, a molecule under relaxed
215 selective pressure could contribute to the virulence of some populations of the species.
216 However, these features of molecular evolutionary analysis can be advantages when
217 screening for therapeutic target sites or vaccine antigens with a low frequency of missense
218 mutations, which could reduce the virulence or survivability of the pathogen. Evolutionary
219 analysis could also be an effective alternative strategy for overcoming drug resistance
220 through antigen replacement, and could reduce costs associated with drug discovery and
221 development.
222 The *lytA* gene, which was conserved among virtually all pneumococcal strains,
223 showed the highest rates of codons under negative selection, except for the *cbpJ* gene that
224 was only present in some strains. LytA is known to induce pneumococcal-specific autolysis[29]
225 and contributes to pneumococcal virulence[30,31]. Our evolutionary analysis supports previous
226 reports that the *lytA* gene is a suitable genetic marker[32,33] due to its evolutionary conservation.
227 We also showed that the *pspA* and *cbpA* genes contain relatively high rates of codons under

228 positive selection, and both encode polymorphic virulent proteins[26,27,34] that are candidate

229 vaccine antigens, although the genes are not universally present within a global serotype 1

230 collection[28]. In addition, selective pressure by vaccines can easily cause differentiation or

231 deficiency of these proteins since the corresponding genes contain few codons under negative

232 selection. A multivalent system would be required for vaccines using these antigens.

233 Although the difference was not statistically significant, mice intranasally infected

234 with TIGR4 *ΔcbpL* strain showed a trend towards improved survival relative to the

235 WT-infected mice. In a previous study, a D39*lux cbpL*-deficient strain showed reduced

236 virulence as compared to the WT strain[35]. Since CbpL sequences in TIGR4 and D39 strains

237 are similar, the discrepancy between the previous study and our findings is likely due to

238 differences in other surface proteins in each strain. For example, the absence of CbpJ, which

239 contributes to the evasion of neutrophil killing, could affect the survivability of D39.

240 Recently, antivirulence drugs have been developed as an additional strategy to treat or

241 prevent bacterial infections. Drugs targeting bacterial virulence factors are expected to reduce

242 the selective pressure of conventional antibiotics since they would not affect the natural

243 survival of targeted bacteria[36]. Furthermore, the abundance of candidate targets is a major

244 advantage of antivirulence strategies. Effective design of vaccines and antivirulence drugs

245 requires a thorough understanding of virulence factors; combining our evolutionary analysis

246 and traditional molecular microbiological approaches can improve the detection of potential

247 drug targets. In this study, we identified CbpJ as a novel evolutionarily conserved virulence

248 factor. Thus, molecular evolutionary analysis is a powerful system that can reveal the

249 importance of virulence factors in real-world infections and transmission.

250 **Methods**

251 **Phylogenetic and evolutionary analyses**

252 Phylogenetic and evolutionary analyses were performed as previously described[37,38], with

253 minor modifications. Homologues and orthologues of *cbp* genes were searched using the

254 tBLASTn function of NCBI BLAST. Bacterial ORFs and promoters were predicted by

255 FGENESB (Bacterial Operon and Gene Prediction) and BPROM, respectively[39]. To prevent

256 node density artefacts, sequences with 100% identity were treated as the same sequence in

257 Phylogears2[40,41]. The sequences were aligned using MAFFT v.7.221 with an L-INS-i

258 strategy,[42] and ambiguously aligned regions were removed using Jalview[43,44]. Calculated

259 orthologous regions were used for further phylogenetic analysis, and edited codon sequences

260 were re-aligned using MAFFT with an L-INS-i strategy. The best-fitting codon evolutionary

261 models for MrBayes and RAxML analyses were determined using Kakusan4[45]. Bayesian

262 Markov chain Monte Carlo analyses were performed with MrBayes v.3.2.5[46], and 2×10^6

263 generations were sampled after confirming that the standard deviation of split frequencies

264 was < 0.01 for up to 8×10^6 generations. To validate phylogenetic inferences, maximum

265 likelihood phylogenetic trees with bootstrap values were generated with RAxML v.8.1.20[47].

266 Phylogenetic trees were generated using FigTree v.1.4.2[48] based on the calculated data.

267 Evolutionary analyses were performed based on aligned orthologous regions of *cbp*

268 genes and Bayesian phylogenetic trees. Whole-gene non-synonymous/synonymous ratio

269 calculations as well as statistical tests for negative or positive selection of individual codons

270 were performed using the two-rate fixed-effects likelihood function in HyPhy software

271 package[49].

272

273 **Bacterial strains and construction of mutant strains**

274 *Streptococcus pneumoniae* strains were cultured in Todd-Hewitt broth (BD Biosciences,

275 Franklin Lakes, NJ, USA) supplemented with 0.2% yeast extract (BD Biosciences) (THY

276 medium) at 37°C. For mutant selection and maintenance, spectinomycin (Wako Pure

277 Chemical Industries, Osaka, Japan) was added to the medium at a concentration of 120

278 μg/ml.

279 *S. pneumoniae* TIGR4 isogenic *cbpJ* (*ΔcbpJ*) and *cbpL* (*ΔcbpL*) mutant strains were

280 generated as previously described[22]. Briefly, the upstream region of *cbpJ* or *cbpL*, an *aad9*

281 cassette, and the downstream region of *cbpJ* or *cbpL* were combined by PCR using primers

282 shown in Supplementary Table 4. The products were used to construct the mutant strains by

283 double-crossover recombination with synthesised CSP2[50]. All mutations were confirmed by

284 PCR amplification of genomic DNA isolated from the mutant strains.

285

286 **Blood and neutrophil bactericidal assays**

287 A blood bactericidal assay was performed as previously described[20,22,51]. Mouse blood was

288 obtained via cardiac puncture from healthy female CD-1 mice (Slc:ICR, 6 weeks old; Japan

289 SLC, Hamamatsu, Japan). For human neutrophil isolation, blood was collected via

290 venopuncture from healthy donors after obtaining written, informed consent according to a

291 protocol approved by the institutional review board of Osaka University Graduate School of

292 Dentistry (H26-E43). Neutrophils were isolated from fresh human blood by density gradient

293 centrifugation using Polymorphprep (Alere Technologies, Jena, Germany). Pneumococcal

294 cells grown to the mid-log phase were washed and resuspended in phosphate-buffered saline

295 (PBS). Bacterial cells (1×10^4 CFU/20 μl) were combined with fresh mouse blood (180 μl)

296 or human neutrophils (2×10^5 cells/180 μl), and the mixture was incubated at 37°C in 5%

297 CO_2 for 1, 2, and 3 h. Viable cell counts were determined by seeding diluted samples onto

298 THY blood agar. The percent of original inoculum was calculated as the number of CFU at

299 the specified time point divided by the number of CFU in the initial inoculum.

300

301 **Mouse infection assays**

302 All mouse experiments were conducted in accordance with animal protocols approved by the
303 Animal Care and Use Committee of Osaka University Graduate School of Dentistry
304 (28-002-0). Female CD-1 mice (Slc:ICR, 6 weeks old) were intranasally infected with $5 \times$
305 10^7 or 2×10^6 CFU of *S. pneumoniae* via the tail vein. Mouse survival was monitored for 14
306 days. At 24 h after intranasal infection, animals were euthanized by lethal intraperitoneal
307 injection of sodium pentobarbital and lung tissue or BALF samples were collected. Bacterial
308 counts in BALF were determined by plating serial dilutions. Lung tissue specimens were
309 fixed with 4% formaldehyde, embedded in paraffin, and cut into sections that were stained
310 with haematoxylin and eosin solution (Applied Medical Research, Osaka, Japan) and
311 visualized with a BZ-X710 microscope (Keyence, Osaka, Japan).

312

313 **qPCR**

314 qPCR was performed as previously described[37,38], with minor modifications. Primers are
315 listed in Supplementary Table 4. Total RNA of pneumococcal strains grown to the mid-log
316 phase ($OD_{600} = 0.5$) was isolated with an RNeasy Mini kit (Qiagen, Valencia, CA, USA) and
317 RQ1 RNase-Free DNase (Promega, Madison, WI, USA), and cDNA was synthesised with

318 SuperScript IV VILO Master Mix (Life Technologies, Carlsbad, CA, USA). qPCR analysis

319 was performed on a StepOnePlus Real-Time PCR system using Power SYBR Green Master

320 PCR mix (Thermo Fisher Scientific, Waltham, MA, USA). 16S rRNA was used as a

321 normalising control.

322

323 **Statistical analysis**

324 Statistical analysis of *in vitro* and *in vivo* data was performed with a non-parametric

325 Kruskal-Wallis test with Dunn's multiple comparisons test. Mouse survival curves were

326 compared with the log-rank test. $P < 0.05$ was considered statistically significant. The tests

327 were carried out with Prism v.7.0d software (GraphPad Inc., La Jolla, CA, USA). All

328 experiments were repeated at least three times. In the evolutionary analyses, $P < 0.1$ was

329 regarded as a significant difference with the HyPhy default setting.

Acknowledgements

This study was supported in part by Japanese Society for the Promotion of Science KAKENHI (grant numbers 15H05012, 16H05847, 16K15787, 17H05103, 17K11666, and 17H04369); SECOM Science and Technology Foundation; Takeda Science Foundation; GSK Japan Research Grant; Asahi Glass Foundation; Kurata Memorial Hitachi Science and Technology Foundation; and Kobayashi International Scholarship Foundation. The funders had no role in study design, data collection or analysis, decision to publish, or preparation of the manuscript.

Author contributions

M.Y. and S.K. designed the study. M.Y. and Y.Y. performed bioinformatics analyses. K.G., M.Y., and Y.H. performed the experiments. M.Y., T.S., M.N., and S.K. contributed to the setup of the experimentation. M.Y. wrote the manuscript. G.K., Y.H., Y.Y., T.S., M.N., K.N., and S.K. contributed to the writing of the manuscript.

Competing financial interests statement

The authors declare that they have no competing interests.

347 **References**

348 1 O'Neill, J. Tackling Drug-Resistant Infections Globally: final report and
349 recommendations. The review on antimicrobial resistance; London: HM Government
350 and the Wellcome Trust; 2016.
351 2 Paul, S.M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand
352 challenge. *Nat Rev Drug Discov* **9**, 203–214 (2010).
353 3 CDC. Antibiotic Resistance Threats in the United States (2013).
354 4 CDC. Biggest Threats, <https://www.cdc.gov/drugresistance/biggest_threats.html>
355 (2017).
356 5 WHO. WHO Priority Pathogens List for R&D of New Antibiotics,
357 <http://www.who.int/mediacentre/news/releases/2017/bacteria-antibiotics-needed/en/>
358 (2017).
359 6 Richards, V.P. *et al.* Phylogenomics and the dynamic genome evolution of the genus
360 *Streptococcus*. *Genome Biol Evol* **6**, 741–753 (2014).
361 7 Kawamura, Y., Hou, X.G., Sultana, F., Miura, H. & Ezaki, T. Determination of 16S
362 rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic
363 relationships among members of the genus *Streptococcus*. *Int J Syst Bacteriol* **45**, 406–
364 408 (1995).
365 8 Walker, C.L. *et al.* Global burden of childhood pneumonia and diarrhoea. *Lancet* **381**,
366 1405–1416 (2013).
367 9 Castelblanco, R.L., Lee, M. & Hasbun, R. Epidemiology of bacterial meningitis in the
368 USA from 1997 to 2010: a population-based observational study. *Lancet Infect Dis* **14**,
369 813–819 (2014).
370 10 Collaborators, G.L. Estimates of the global, regional, and national morbidity, mortality,
371 and aetiologies of lower respiratory tract infections in 195 countries: a systematic
372 analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis* **17**, 1133–1161
373 (2017).
374 11 Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant
375 formed through a multi-fragment recombination event. *Nat Genet* **44**, 352–355 (2012).
376 12 Flasche, S. *et al.* Effect of pneumococcal conjugate vaccination on serotype-specific
377 carriage and invasive disease in England: a cross-sectional study. *PLoS Med* **8**, e1001017
378 (2011).
379 13 Brockhurst, M.A. *et al.* Running with the Red Queen: the role of biotic conflicts in
380 evolution. *Proc Biol Sci* **281**, doi:10.1098/rspb.2014.1382 (2014).
381 14 Sironi, M., Cagliani, R., Forni, D. & Clerici, M. Evolutionary insights into

382 host-pathogen interactions from mammalian sequence data. *Nat Rev Genet* **16**, 224–236
383 (2015).
384 15 Maestro, B. & Sanz, J.M. Choline binding proteins from *Streptococcus pneumoniae*: a
385 dual role as enzybiotics and targets for the design of new antimicrobials. *Antibiotics*
386 *(Basel)* **5**, doi:10.3390/antibiotics5020021 (2016).
387 16 Fumagalli, M. & Sironi, M. Human genome variability, natural selection and infectious
388 diseases. *Curr Opin Immunol* **30**, 9–16 (2014).
389 17 Karlsson, E.K., Kwiatkowski, D.P. & Sabeti, P.C. Natural selection and infectious
390 disease in human populations. *Nat Rev Genet* **15**, 379–393 (2014).
391 18 Siddle, K.J. & Quintana-Murci, L. The Red Queen's long race: human adaptation to
392 pathogen pressure. *Curr Opin Genet Dev* **29**, 31–38 (2014).
393 19 Terao, Y. *et al.* Group A streptococcal cysteine protease degrades C3 (C3b) and
394 contributes to evasion of innate immunity. *J Biol Chem* **283**, 6253–6260 (2008).
395 20 Yamaguchi, M., Terao, Y., Mori, Y., Hamada, S. & Kawabata, S. PfbA, a novel plasmin-
396 and fibronectin-binding protein of *Streptococcus pneumoniae*, contributes to
397 fibronectin-dependent adhesion and antiphagocytosis. *J Biol Chem* **283**, 36272–36279
398 (2008).
399 21 Sumitomo, T. *et al.* Streptolysin S contributes to group A streptococcal translocation
400 across an epithelial barrier. *J Biol Chem* **286**, 2750–2761 (2011).
401 22 Mori, Y. *et al.* alpha-Enolase of *Streptococcus pneumoniae* induces formation of
402 neutrophil extracellular traps. *J Biol Chem* **287**, 10472–10481 (2012).
403 23 Yamaguchi, M. *et al. Streptococcus pneumoniae* invades erythrocytes and utilizes them
404 to evade human innate immunity. *PLoS One* **8**, e77282, (2013).
405 24 Gallagher, M.D. & Chen-Plotkin, A.S. The post-GWAS era: from association to function.
406 *Am J Hum Genet* **102**, 717–730 (2018).
407 25 Marigorta, U.M., Rodriguez, J.A., Gibson, G. & Navarro, A. Replicability and
408 prediction: lessons and challenges from GWAS. *Trends Genet*
409 doi:10.1016/j.tig.2018.03.005 (2018).
410 26 Hollingshead, S.K. *et al.* Pneumococcal surface protein A (PspA) family distribution
411 among clinical isolates from adults over 50 years of age collected in seven countries. *J*
412 *Med Microbiol* **55**, 215–221 (2006).
413 27 Brooks-Walter, A., Briles, D.E. & Hollingshead, S.K. The *pspC* gene of *Streptococcus*
414 *pneumoniae* encodes a polymorphic protein, PspC, which elicits cross-reactive
415 antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infect Immun*
416 **67**, 6533–6542 (1999).

417 28 Cornick, J.E. *et al.* The global distribution and diversity of protein vaccine candidate
418 antigens in the highly virulent *Streptococcus pneumoniae* serotype 1. *Vaccine* **35**, 972–
419 980 (2017).
420 29 Mosser, J.L. & Tomasz, A. Choline-containing teichoic acid as a structural component of
421 pneumococcal cell wall and its role in sensitivity to lysis by an autolytic enzyme. *J Biol*
422 *Chem* **245**, 287–298 (1970).
423 30 Berry, A.M., Lock, R.A., Hansman, D. & Paton, J.C. Contribution of autolysin to
424 virulence of *Streptococcus pneumoniae*. *Infect Immun* **57**, 2324–2330 (1989).
425 31 Orihuela, C.J., Gao, G., Francis, K.P., Yu, J. & Tuomanen, E.I. Tissue-specific
426 contributions of pneumococcal virulence factors to pathogenesis. *J Infect Dis* **190**, 1661–
427 1669 (2004).
428 32 Carvalho Mda, G. *et al.* Evaluation and improvement of real-time PCR assays targeting
429 *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. *J Clin Microbiol* **45**,
430 2460–2466 (2007).
431 33 Saukkoriipi, A. *et al. lytA* quantitative PCR on sputum and nasopharyngeal swab
432 samples for detection of pneumococcal pneumonia among the elderly. *J Clin Microbiol*
433 **56**, doi:10.1128/JCM.01231-17 (2018).
434 34 Hakenbeck, R., Madhour, A., Denapaite, D. & Bruckner, R. Versatility of choline
435 metabolism and choline-binding proteins in *Streptococcus pneumoniae* and commensal
436 streptococci. *FEMS Microbiol Rev* **33**, 572–586 (2009).
437 35 Gutierrez-Fernandez, J. *et al.* Modular architecture and unique teichoic acid recognition
438 features of choline-binding protein L (CbpL) contributing to pneumococcal pathogenesis.
439 *Sci Rep* **6**, 38094 (2016).
440 36 Dickey, S.W., Cheung, G.Y.C. & Otto, M. Different drugs for bad bugs: antivirulence
441 strategies in the age of antibiotic resistance. *Nat Rev Drug Discov* **16**, 457–471 (2017).
442 37 Yamaguchi, M. *et al.* Evolutionary inactivation of a sialidase in group B *Streptococcus*.
443 *Sci Rep* **6**, 28852 (2016).
444 38 Yamaguchi, M. *et al.* Zinc metalloproteinase ZmpC suppresses experimental
445 pneumococcal meningitis by inhibiting bacterial invasion of central nervous systems.
446 *Virulence* **8**, 1516–1524 (2017).
447 39 Solovyev, V. & Salamov, A. *Metagenomics and its Applications in Agriculture,*
448 *Biomedicine and Environmental Studies* (ed. Li W.R.) Ch. 4, 61–78 (Nova Science
449 Publishers, 2011).
450 40 Tanabe, A.S. Phylogears2 ver. 2.0, <http://www.fifthdimension.jp/> (2008).
451 41 Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny

452 reconstruction. *Syst Biol* **55**, 637–643 (2006).
453 42 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
454 improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
455 43 Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview
456 Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*
457 **25**, 1189–1191 (2009).
458 44 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and
459 ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577
460 (2007).
461 45 Tanabe, A.S. Kakusan4 and Aminosan: two programs for comparing nonpartitioned,
462 proportional and separate models for combined molecular phylogenetic analyses of
463 multilocus sequence data. *Mol Ecol Resour* **11**, 914–921 (2011).
464 46 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model
465 choice across a large model space. *Syst Biol* **61**, 539–542 (2012).
466 47 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
467 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
468 48 Rambaut, A. FigTree ver.1.4.2, <http://tree.bio.ed.ac.uk/software/figtree/> (2014).
469 49 Pond, S.L., Frost, S.D. & Muse, S.V. HyPhy: hypothesis testing using phylogenies.
470 *Bioinformatics* **21**, 676–679 (2005).
471 50 Bricker, A.L. & Camilli, A. Transformation of a type 4 encapsulated strain of
472 *Streptococcus pneumoniae*. *FEMS Microbiol Lett* **172**, 131–135 (1999).
473 51 Yamaguchi, M. *et al.* Role of *Streptococcus sanguinis* sortase A in bacterial colonization.
474 *Microbes Infect* **8**, 2791–2796 (2006).
475

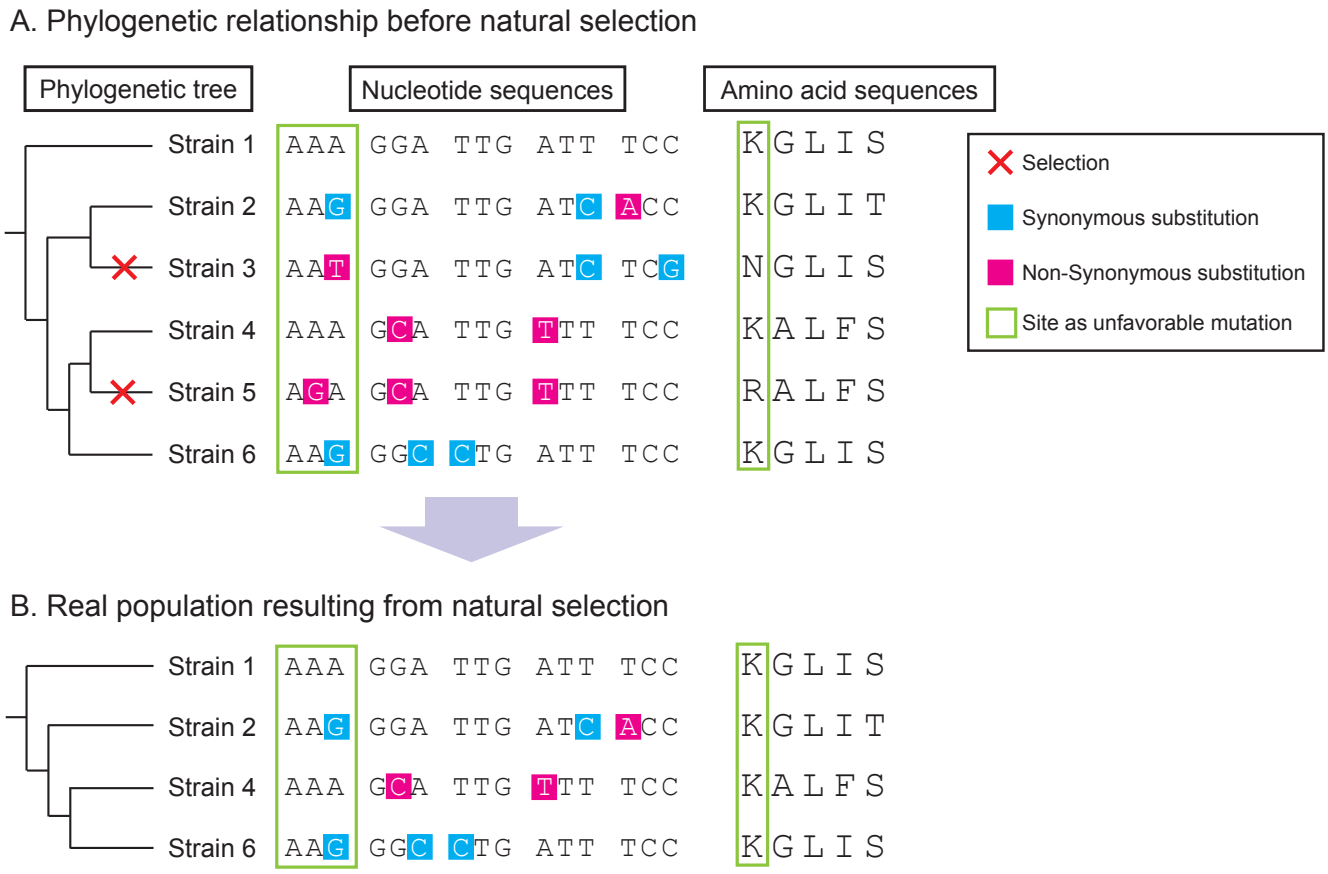


Figure 1. Scheme for intra-species molecular evolutionary analysis. A. Random genetic drift induces synonymous and non-synonymous mutations with equal probability. However, non-synonymous mutations in the essential region cause host selection. **B.** As a result of natural selection, synonymous substitutions are concentrated in important genes. Phylogenetic and molecular evolutionary analyses can detect significant accumulation of synonymous substitutions in codons of host proteins. Codon-based analysis yields much more information than nucleotide- or amino acid-based analyses.

Figure 1. Yamaguchi *et al.*

A

Serotype	4	1	1	1	1	2	N.T.	3	3	3	3	3	3	3	5	6B	11A	14	14	19A	19A	19F	19F	19F	19F	19F	23F	N.T.
Gene \ Strain	TIGR4	P1031	INV104	gamPNI0373	NCTC7465	D39	R6	SPNA45	OXC141	SPN034156	SPN034183	SPN994038	SPN994039	A66	70585	670-6B	AP200	CGSP14	INV200	Hungary19A-6	TCH8431/19A	JJA	Taiwan19F-14	G54	ST556	A026	ATCC 700669	NT_110_58
cbpI																												
pspA																												
cbpF								*																				
cbpJ								*																				
cbpC																												
cbpG																												
cbpK																												
cbpL																												
cbpE																												
lytB																												
cbpM																												
lytC																												
lytA																												
cbpA																												
pcpA																												
cbpD																												

B

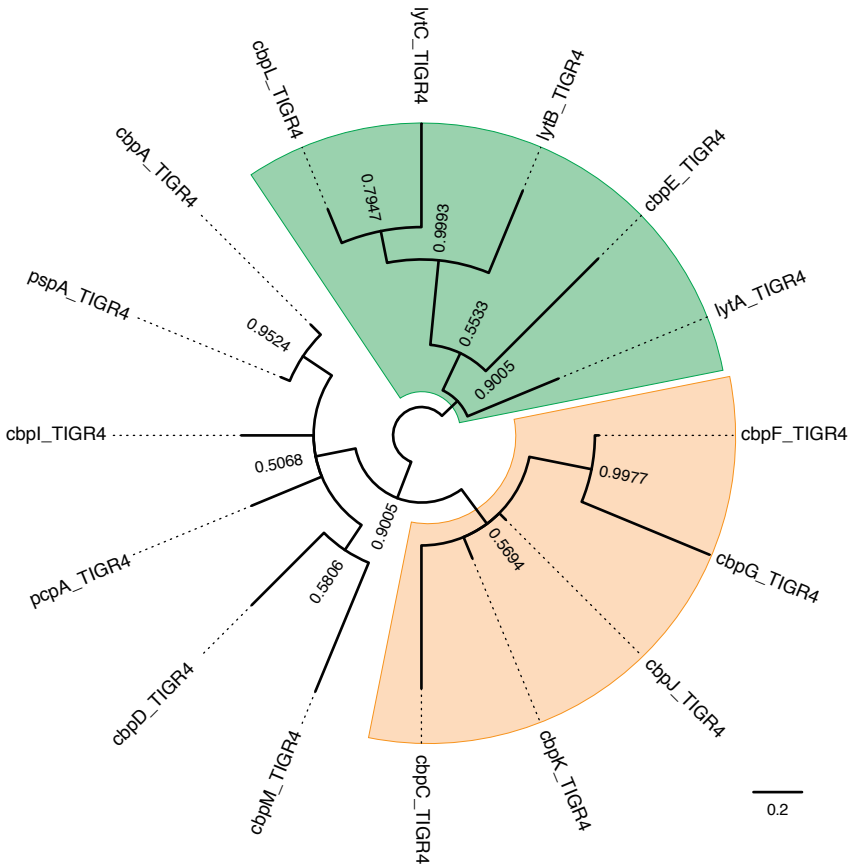
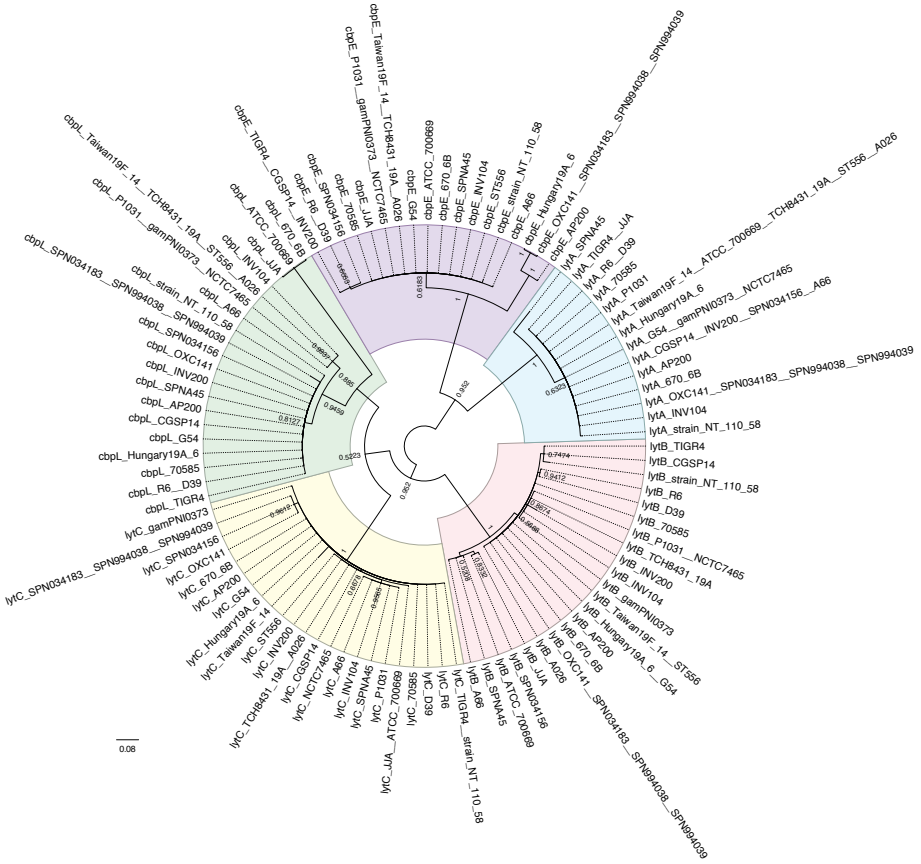


Figure 2. Yamaguchi *et al.*

Figure 2. Distribution of *cbp* genes and phylogenetic relationship in TIGR4. A. Distribution of genes encoding CBPs among pneumococcal strains. Blue, yellow, and gray show the presence, pseudogenisation, and absence of genes, respectively. *These genes are annotated as one gene, but our bioinformatic analysis indicates that they are independent genes. **B.** Nucleotide-based Bayesian phylogenetic tree of *cbp* genes of *S. pneumoniae* strain TIGR4. The tree is unrooted and posterior probabilities are shown near the nodes. The scale bar indicates nucleotide substitutions per site.

A



B

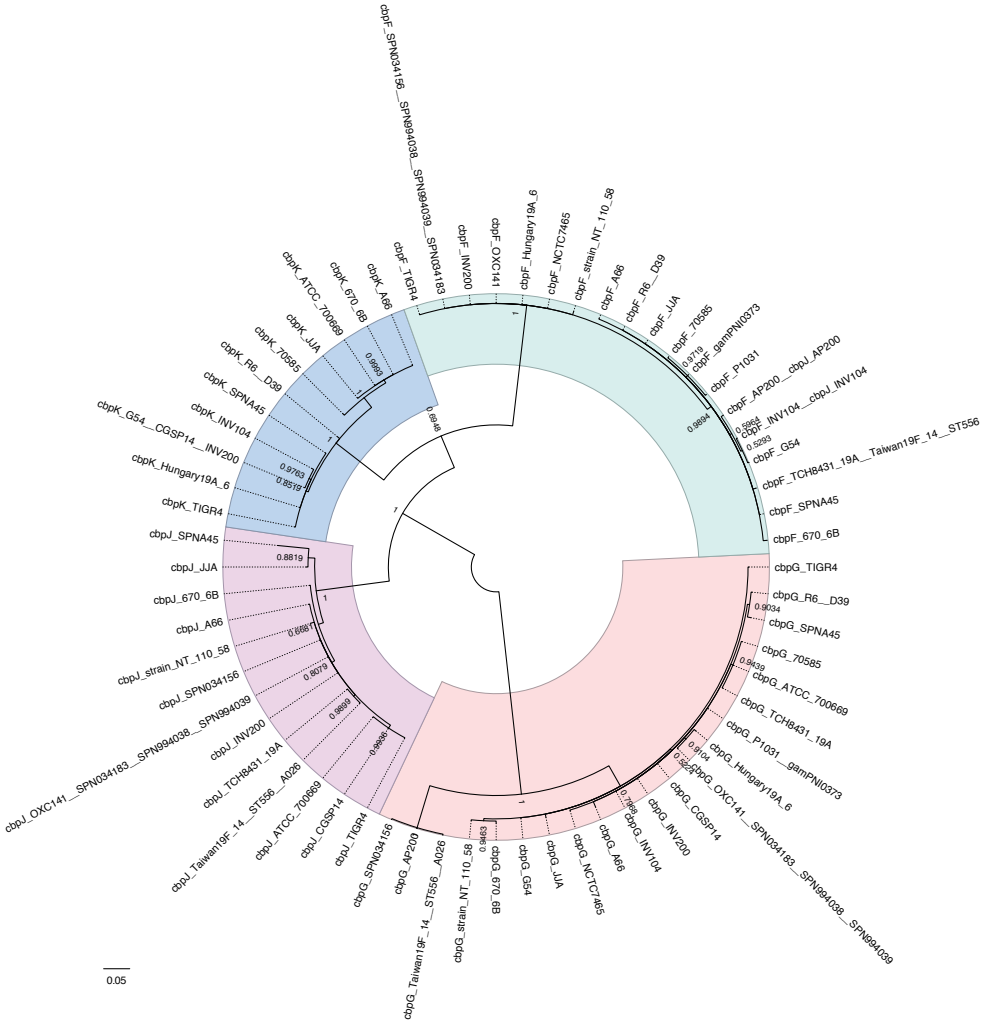


Figure 3. Yamaguchi *et al.*

Figure 3: Phylogenetic analyses of *cbp* genes with high similarity. A, B. Nucleotide-based Bayesian phylogenetic tree of the *lytA*, *lytB*, *lytC*, *cbpE*, and *cbpL* genes (A) and the *cbpF*, *cbpG*, *cbpJ*, and *cbpK* genes (B) in *S. pneumoniae*. The trees are unrooted although they are presented as midpoint-rooted for clarity. Strains with identical sequences are listed on the same branch. Posterior probabilities are shown near the nodes. The scale bar indicates nucleotide substitutions per site.

A

TIGR4 CbpL

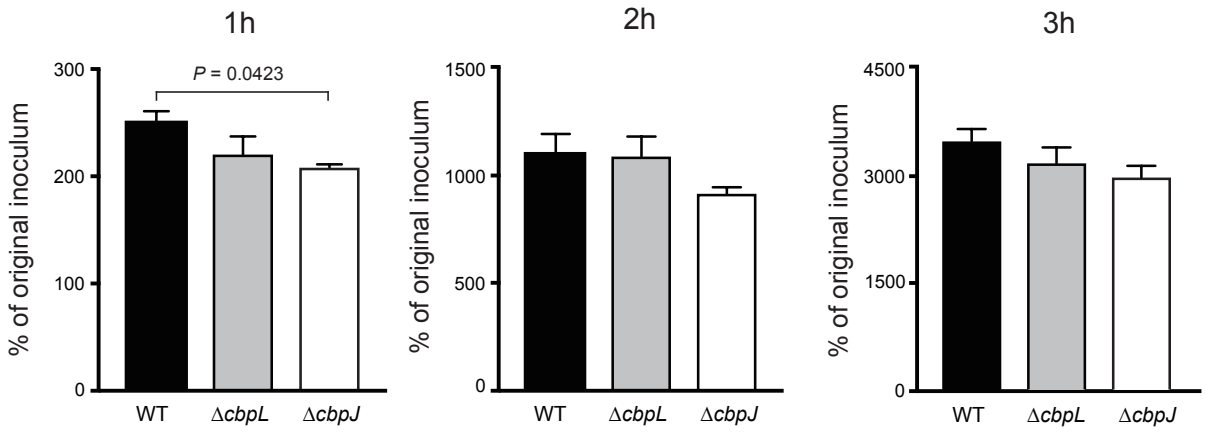
MNKRLFSKMSLVTLPILALFSQSVLAEENI
HFSSCKEAWANGYSDIHEGEPGYSAKLDRD
HDGVACELKNAPKGAFKAKQSTAIQINTSS
ATTSGWVKQDGAWYYFDGNGNLVKNAWQGS
YYLKADGKMAQSEWIYDSSYQAWYYLKSDG
SYAKNAWQGAYYLKSNGKMAQGEWVYDSSY
QAWYYLKSDGSYARNAWQGNYYLKSDGKMA
KGEWVYDATYQAWYYLTSDGSYAYSTWQGN
YYLKSDGKMAVNEWVDGGRYYVGADGVWKE
VQASTASSSNDSNSEYSAALGKAKSYNSLF
HMSKKRMYRQLTSDFDKFSNDAAQYAIDHL
DD*

TIGR4 CbpJ

MKILKKTMQVGLTVFFFGLLGTSTVFADDS
EGWQFVQENGRTYYKKGDLKETYWRVIDGK
YYYFDSLSGEMVVGWQYIPFPSKGSTIGPY
PNGIRLEGFPKSEWYYFDKNGVLQEFVGWK
TLEIKTKDSVGRKYGEKREDSEDKEEKRYY
TNYYFNQNHSLETGWLYDQSNWYYLAKTEI
NGENYLGGERRAGWINDDSTWYYLDPTTGI
MQTGWQYLGNKWYYLRSSGAMATGWYQDGS
TWYYLNAGNGDMKTGWFQVNGNWYYAYSSG
ALAVNTTVDGYSVNYNGEWVR*

Signal Sequence | Excalibur domain | Choline binding domain | Lipoprotein domain

B



C

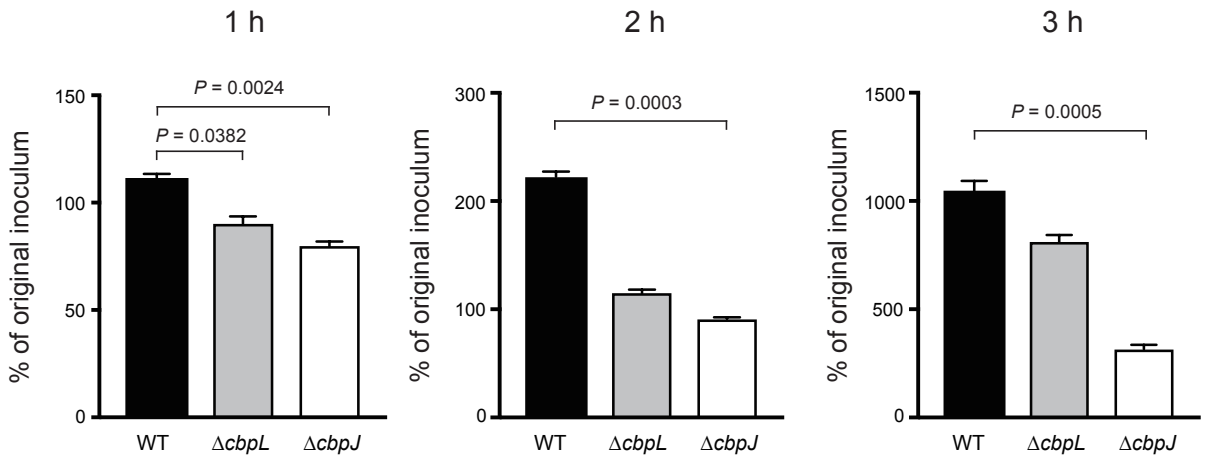


Figure 4. Yamaguchi *et al.*

Figure 4: Blood and neutrophil bactericidal assays using TIGR4 WT, *ΔcbpJ*, and *ΔcbpL* strains. A. Amino acid sequences and domain structures of CbpL and CbpJ in strain TIGR4. Bold, black underlined, and magenta underlined characters represent comparable codons and those under purifying or positive selection, respectively. **B, C.** Growth of pneumococcal strains in mouse blood (B) and in the presence of human neutrophils (C). Bacterial cells were incubated in blood or with neutrophils for 1, 2, and 3 h at 37°C and 5% CO_2, then serially diluted and plated on TS blood agar. The number of CFUs was determined following incubation. Growth index was calculated by dividing CFU after incubation by the CFU of the original inoculum. Data are presented as the mean of six samples with standard error.

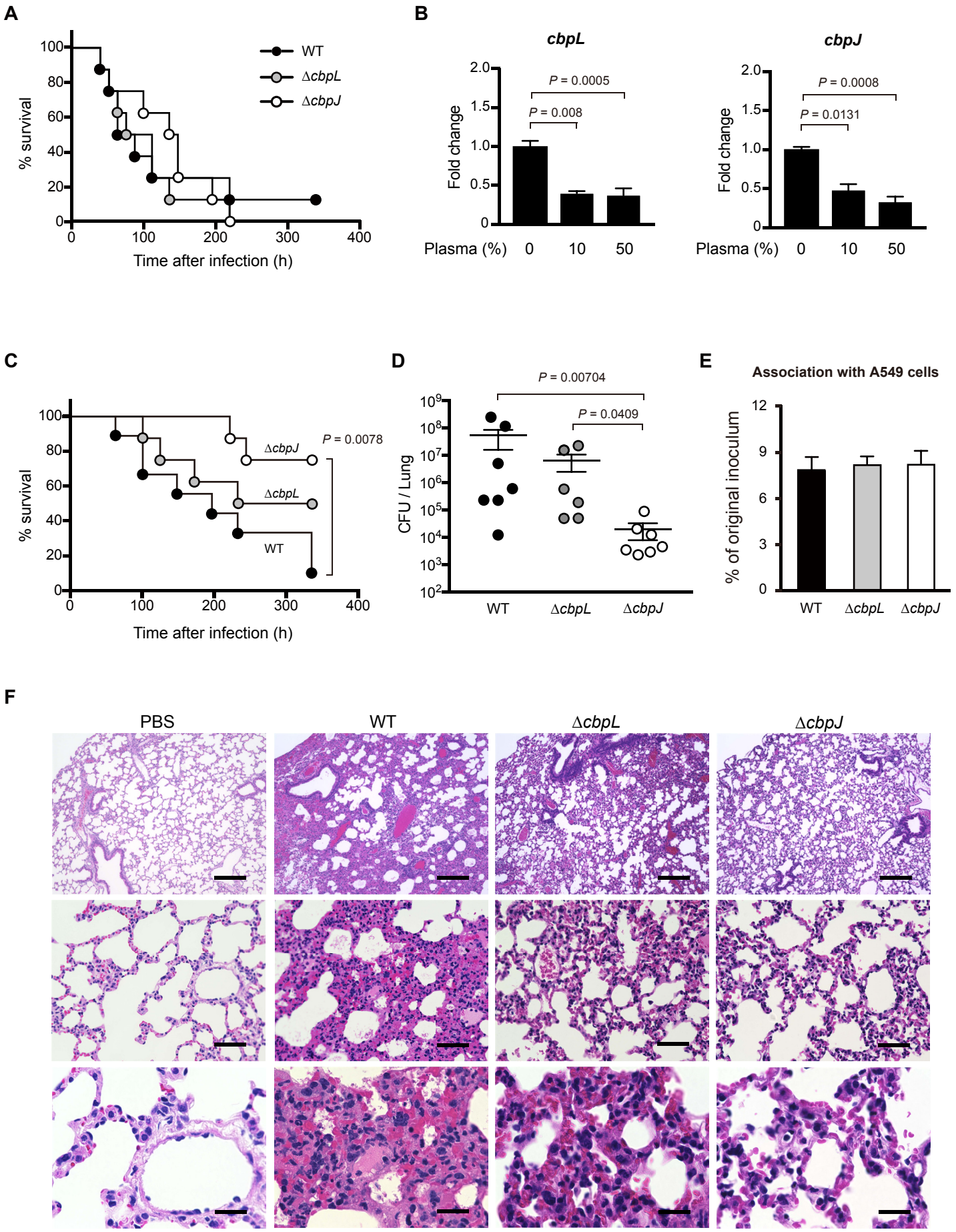


Figure 5. Yamaguchi *et al.*

Figure 5: Intranasal and intravenous infection with *S. pneumoniae* TIGR4 WT, *ΔcbpJ*, and *ΔcbpL* strains. A. Mouse sepsis model. Mice were intravenously infected with 2×10^6 CFU of *S. pneumoniae* TIGR4 WT, *ΔcbpL*, or *ΔcbpJ*, and survival was monitored for 14 days. Differences between infected mouse groups were analysed with the log-rank test. **B.** Fold transcript levels of *cbpL* and *cbpJ* in TIGR4 WT *S. pneumoniae* cells in the presence or absence of human plasma. 16S rRNA was used as an internal standard. Data were pooled and normalised from three independent experiments, each performed in quadruplicate. **C.** Mouse pneumonia model. Mice were intranasally infected with 5×10^7 CFU of *S. pneumoniae* TIGR4 WT, *ΔcbpL*, or *ΔcbpJ* strains, and survival was monitored for 14 days. **D.** Pneumococcal CFU in BALF collected 24 h after intranasal infection. The difference between groups was analysed with the Kruskal-Wallis test with Dunn's multiple comparisons test. **E.** *S. pneumoniae* TIGR4 WT, *ΔcbpL*, and *ΔcbpJ* strains were examined for their ability to associate with A549 cells. Data are presented as the mean of six samples. Differences between groups were analysed with the Kruskal-Wallis test with Dunn's multiple comparisons test. **F.** Haematoxylin and eosin staining of infected mouse lung tissue collected 24 h after intranasal infection. Scale bars, 200 µm (upper panels), 50 µm (middle panels), and 20 µm (lower panels). Data are shown as mean with standard error (B, D, and E).

476 **Table 1.** Evolutionary analyses of genes encoding choline-binding proteins*

Genes	Number of sequences[1]	dN/dS	Coverage of comparable codons relative to whole protein in TIGR4	Codons evolving under positive selection	Codons evolving under purifying selection	% Of codons under purifying selection relative to total codons
pspA	24	0.857	19.060% (142/745)	6.338% (9/142)	12.676% (18/142)	2.416%
cbpF	19	0.395	60.411% (206/341)	0.485% (1/206)	3.398% (7/206)	2.053%
cbpJ	15	0.346	84.084% (280/333)	1.429% (4/280)	18.571% (52/280)	15.616%
cbpC	13	–	0% (0/93)	–	–	0.000%
cbpG	21	–	0% (0/286)	–	–	0.000%
cbpK	11	0.353	85.630% (292/341)	0.342% (1/292)	3.082% (9/292)	2.639%
cbpL	20	–	0% (0/333)	–	–	0.000%
cbpE	18	0.325	99.363% (624/628)	0.160% (1/624)	4.968% (31/624)	4.936%
lytB	22	0.185	92.868% (612/659)	0% (0/612)	4.739% (29/612)	4.401%
cbpM	10	0.642	98.462% (128/130)[2]	0% (0/128)	0% (0/128)	0.000%
lytC	23	0.400	19.348% (95/491)	0% (0/95)	5.263% (5/95)	1.018%
lytA	14	0.141	80.564% (257/319)	0% (0/257)	17.121% (44/257)	13.793%
cbpA	19	0.864	22.334% (155/694)	3.226% (5/155)	7.742% (12/155)	1.729%
pcpA	18	0.261	77.010% (479/622)	0% (0/479)	0.418% (2/479)	0.322%
cbpD	19	0.359	75.278% (338/449)	0.296% (1/338)	3.550% (12/338)	2.672%
cbpI	2	–	–	–	–	–

477 [1]Sequences with 100% identity were treated as the same sequence; [2]compared to D39.

478 *Evolutionary analysis was performed by Bayesian inference of aligned *cbp* sequences from complete genomes of *S. pneumoniae* with the two-rate fixed-effects
479 likelihood function in HyPhy software package. dN/dS is the ratio of non-synonymous to synonymous changes in overall analysed genes. Individual codons with a
480 statistically significant signature were also calculated and are expressed as a percentage of the total number of codons included in the analysis.