1 **Identification of colonisation and virulence determinants of *Streptococcus***

2 ***pneumoniae* via experimental evolution in mouse infection models**

3 Angharad E Green[1], Deborah Howarth[1], Chrispin Chaguza[2], Haley Echlin[3], R Frèdi Langendonk[1],

4 Connor Munro[1], Thomas E Barton[1], Jay CD Hinton[1], Stephen D Bentley[2], Jason W Rosch[3], Daniel R

5 Neill[1]

6 1. Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary and Ecological
7   Sciences, University of Liverpool, UK.
8 2. Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK
9 3. St. Jude Children's Research Hospital, Department of Infectious Disease, Memphis, Tennessee, USA.
10
11 Correspondence: d.neill@liverpool.ac.uk
12

13 **Abstract:** *Streptococcus pneumoniae* is a commensal of the human nasopharynx and

14 a major cause of respiratory and invasive disease. Pneumococcus stimulates upper

15 respiratory tract inflammation that promotes shedding from mucosal surfaces and

16 transmission to new hosts. Colonisation and transmission are partially antagonistic

17 processes. Adhesion to surfaces and evasion of host responses favours the former,

18 whilst detachment, promoted by inflammation, is necessary for the latter. We sought

19 to determine how adaptation and evolution of pneumococcus within its

20 nasopharyngeal niche might progress when selective pressures associated with

21 transmission were removed. This was achieved by serial passage of pneumococci in

22 mouse models of nasopharyngeal carriage, manually transferring bacteria between

23 mice. To assess the role of host environmental factors on pneumococcal evolution,

24 we also performed analogous experimental evolution in a mouse pneumonia model,

25 passaging pneumococci through lungs. Nasopharynx-passaged pneumococci

26 became more effective colonisers, whilst those evolved within lungs showed reduced

27 virulence. We observed selection of mutations in genes associated with cell wall

28 biogenesis and metabolism in both nasopharynx and lung lineages, but identified

29 prominent examples of parallel evolution that were niche specific. We focussed on

30 *gpsA*, a gene in which the same single nucleotide polymorphism arose in two

31 independently evolved nasopharynx-passaged lineages. We identified a single

32 nucleotide change conferring resistance to oxidative stress and enhanced

33 nasopharyngeal colonisation potential. We show that *gpsA* is also a frequent target of

34 mutation during human colonisation. These findings highlight the role played by the

35 host environment in determining trajectories of bacterial evolution and the potential of

36 experimental evolution in animal infection models for identification of novel pathogen

37 virulence and colonisation factors.

38

39 **Introduction**

40 Pneumonia and invasive disease caused by *Streptococcus pneumoniae* is a major

41 contributor to global morbidity and mortality, but the primary lifestyle of this

42 opportunistic pathogen is as a commensal of the nasopharynx. Compromised defence

43 mechanisms may render a host permissive to disease, as exemplified by the rising

44 susceptibility to pneumococcal pneumonia in the elderly[1], whilst the suite of

45 pneumococcal virulence factors contributes to disease development via evasion of

46 host defences and stimulation of inflammation[2]. Several of these virulence factors are

47 likely retained by the pneumococcus primarily as a means to ensure onwards

48 transmission from nasopharynx, via the stimulation of the inflammation necessary to

49 promote detachment from host surfaces.

50

51 The positive correlation between transmission and virulence is well described for a

52 range of pathogens[3]. Experimental evolution approaches[4] have contributed to this

53 understanding, whereby infectious material is passaged from host to host, in an

54 experimental system that promotes gain of infectivity via selection of individuals

55 displaying high levels of *in vivo* growth. In studies with viral and parasite species, this

56   approach has demonstrated that increased virulence mirrors the increase in

57   transmissibility[5-8].

58

59   The functional link between transmission and virulence traits has been demonstrated

60   with pneumococcus in relation to its haemolytic toxin, pneumolysin. Pneumolysin

61   contributes to pathology via direct lysis of host cells and stimulation of inflammation[9-

62   11] and pneumolysin-deficient mutants of *S. pneumoniae* demonstrate markedly

63   attenuated virulence in animal models[9-11]. Pneumococcal lineages in which

64   pneumolysin lacks pore-forming activity typically cause non-lethal respiratory disease

65   in humans[12]. Zafar and colleagues have demonstrated that the inflammation

66   stimulated by pneumolysin in the upper respiratory tract contributes to shedding of

67   bacteria from the mucosal surfaces and thereby promotes transmission to new

68   hosts[13]. In the absence of virus-induced inflammation, which similarly promotes

69   shedding[14], pneumolysin may be the primary means by which pneumococcus ensures

70   onward transmission. However, pneumolysin is detrimental for carriage in the upper

71   airways, as its proinflammatory properties hasten clearance of colonisation[15]. The

72   trajectory of pneumococcal evolution has thus likely been determined by the need to

73   balance the processes of colonisation and transmission.

74

75   We sought to determine how pneumococcus might adapt to its natural niche in the

76   nasopharynx when selective pressures associated with transmission were removed.

77   Transmission of *S. pneumoniae* is associated with a single-cell bottleneck[16] and so

78   mutations compromising transmission would not be retained in a natural setting, even

79   if they confer advantages in colonisation. Therefore, we manually passaged

80   pneumococci between mice in a model of asymptomatic nasopharyngeal carriage. In

81   parallel, we also performed the same process during lung infection, to determine how

82   environmental differences between the upper and lower airways might shape

83   pneumococcal adaptation and evolution. Our results demonstrate rapid emergence

84   and selection of advantageous mutations in both nasopharynx and lung passaged

85   pneumococci. Some mutations were common to several independently-evolved

86   lineages of pneumococcus, but these examples of parallel evolution were found in

87   different genes in nasopharynx vs lung evolved lineages, suggesting niche-specific

88   adaptation.

89

90   **Results**

91   *In vivo experimental evolution of S. pneumoniae in murine nasopharynx and lungs*

92   *selects for increased niche colonisation potential*

93   To perform experimental evolution of *S. pneumoniae*, we chose the serotype 2 strain

94   D39. Isolated nearly a century ago by Avery, this strain has been widely propagated

95   in a laboratory environment and so may have lost many of its adaptations to the host[17].

96   Using an inoculum prepared from a single colony, we separately infected the

97   nasopharynx or the lung of CD1 mice, infecting ten mice each per niche. Each of these

98   twenty individuals was considered a founder for an independent experimental

99   evolution lineage of pneumococcus (Figure 1). Thereafter, pneumococci were

100   recovered from either nasopharynx or lungs. After minimal growth on agar, the entire

101   population was used to infect a further mouse for the next round of passage. This

102   process was repeated twenty times per lineage, with 7 days of nasopharyngeal

103   carriage or up to 48 hours of lung infection for each round of passage. This yielded

104   ten nasopharynx-passaged lineages and ten lung-passaged lineages. Each

105   nasopharynx-passaged lineage had spent 140 days in total in the nasopharynx, whilst

106   the lung-passaged lineages had spent ~26 days in lungs.

107

108   To determine whether *in vivo* experimental evolution had resulted in altered

109   pneumococcal phenotypes, the twenty times nasopharynx-passaged lineages were

110   assessed for their ability to colonise in a nasopharyngeal carriage model (Table 1,

111   Figure 2A). All ten lineages showed increased colonisation prevalence at day 14 post-

112   infection, relative to the ancestor and, in those mice that remained colonised, mean

113   colonisation density was higher in all lineages except for lineage 8. (Table 1). Over a

114   21-day carriage experiment, nasopharyngeal colonisation density was comparable in

115   the ancestor and a nasopharynx-passaged lineage over the first 7 days of infection,

116   but thereafter clearance of the ancestor proceeded more rapidly (Figure 2A).

117

118   Next, the twenty times passaged lung lineages were assessed for their ability to

119   colonise murine lungs in a pneumonia model (Table 1, Figure 2B). Disease

120   progression was slower in mice infected with lung-passaged lineages in comparison

121   to those infected with the ancestor strain. Survival proportions were increased in mice

122   infected with any of the ten lineages, relative to the ancestor, with significantly different

123   survival curves for lineages 1, 2, 4, 5 and 8 (Table 1, Figure 2B). We also assessed

124   the nasopharyngeal colonisation potential of a lung-passaged lineage and the lung

125   colonisation potential of a nasopharynx-passaged lineage (Supplementary Figure 1).

126   Both the nasopharynx- and lung-passaged lineages behaved comparably during lung

127   infection, showing attenuated virulence (Supplementary Figure 1A). However, the

128   nasopharynx-evolved lineage colonised nasopharynx at a higher density than the

129   lung-evolved lineage over the first 14 days of infection (Supplementary Figure 1B),

130   suggesting niche-specific fitness.

131

132   We determined growth in liquid culture, adhesion to airway epithelial cells *in vitro* and

133   production of pneumolysin in all lineages (Supplementary Figure 2). Growth dynamics

134   (measured as area under the logistic curve [AUC] values from 24 hours growth in

135   broth) were comparable to that of the ancestor for all lineages, with the exception of

136   lung-passaged lineages 7 and 8, and nasopharynx-passaged lineages 2 and 7, which

137   displayed modestly reduced growth rates in nutrient broth (Supplementary Figure 2A

138   and B). The levels of adhesion to A549 airway epithelial cells was highly variable in

139   lung-passaged lineages (Supplementary Figure 2C), but two of the nasopharynx-

140   passaged lineages showed evidence of increased adhesion potential (Supplementary

141   Figure 2D). Levels of pneumolysin, in lysates prepared from mid-log phase

142   pneumococci, were significantly reduced in five of ten lung-passaged lineages and

143   nine of ten nasopharynx-passaged lineages, relative to the ancestor strain

144   (Supplementary Figure 2E and F). The level of epithelial cell adhesion was negatively

145   correlated with the amount of pneumolysin produced in lung passaged lineages (R_2 =

146   0.70) (Supplementary Figure 2G).

147

148   *Genomic characterisation of niche-passaged pneumococci*

149   To uncover the genetic basis of the observed adaptations, we genome sequenced 100

150   pneumococcal populations, which included all twenty times passaged lineages and

151   passage numbers 1, 5, 10 and 15 of each lineage. By comparison with the genome

152   sequence of the ancestor strain, we identified variants that had arisen during

153   experimental evolution. We first performed *de novo* genome assembly for our ancestor

154 D39, using a combination of long- and short-read sequencing. This yielded a single

155 contig genome of 2,046,551 base pairs, a size similar to that determined in a recently

156 published assembly for this strain[18]. Genome annotation with Prokka[19] identified 1,998

157 protein coding sequences, with 664 of these being hypothetical proteins. Deep

158 sequencing of the ancestor population (300x coverage) was performed to enable us

159 to detect low frequency variants in the inoculum used to establish the 20 experimental

160 evolution lineages. Using Breseq[20], we mapped the Illumina reads from the ancestor

161 inoculum against the consensus genome sequence. After filtering out artefactual or

162 unreliable mutations and those at <2% frequency (the suggested frequency-cut off

163 limit for polymorphism mode in Breseq[20]), we observed 43 unique variants, at

164 frequencies of between 2.2 and 20.7% in the total ancestor population (Supplementary

165 Dataset 1). Most of these were subsequently lost during *in vivo* experimental evolution.

166 Those that were retained (in genes *dnaK* and *P0_01974*) are highlighted below.

167

168 Variant calling of Illumina reads from *in vivo* passaged pneumococci, with an average

169 of 147x coverage (range 105x-190x), detected a total of 735 variants across all the

170 nasopharynx lineages (passages 1, 5, 10, 15 and 20), with 206 variants identified in

171 the final populations that had been passaged twenty times through nasopharynx

172 (Supplementary Dataset 1). Within lung-passaged lineages we observed 509 variants,

173 with 128 found across the ten populations at passage twenty (Supplementary Dataset

174 1). The majority of variants detected were SNPs, although deletions of between 1 and

175 53 base pairs were identified, as well as several single nucleotide insertions. Indels

176 were predominantly located in intergenic regions. Cluster of orthologous gene (COG)

177 analysis[21] of variant data demonstrated that genes with roles in amino acid and

178 nucleotide metabolism accumulated mutations in lineages from both niches. In lung-

179 passaged lineages, 16% of variants were found in genes associated with carbohydrate

180 transport and metabolism, whilst 6% of variants in nasopharynx-passaged lineages

181 were in genes associated with energy production and conversion (Supplementary

182 Figure 3).

183

184 To identify mutations most likely to drive adaptive phenotypic change, we focused on

185 mutations that became fixed in the population during experimental evolution.

186 Considering only the final populations in each lineage (that is, those had been

187 passaged twenty times through a niche), we identified 47 different non-synonymous

188 SNPs within nasopharynx-passaged pneumococci (Figure 3A), and 31 in lung-

189 passaged pneumococci (Figure 3B), that became fixed (reached 100% frequency) in

190 at least one lineage. These included several examples of potential parallel or

191 convergent evolution, whereby the same mutation, or closely located mutations, arose

192 and reached fixation in two or more independently evolved lineages (Table 2). Two of

193 these were observed only in nasopharynx-passaged lineages, three were observed

194 only in lung-passaged lineages and three were found in lineages from both

195 nasopharynx and lungs. Despite the notable changes in pneumolysin production

196 observed, none of the lung or carriage lineages had any variants in the *ply* gene or

197 flanking regulatory regions.

198

199 We also performed a control passage process, whereby the ancestor was serially

200 passaged 20 times on blood agar, to control for any selective pressures associated

201 with sequential growth on agar plates during evolution of the niche-evolved lineages.

202 Variants identified in control passage lineages largely differed from those identified in

203 the nasopharynx and lung lineages (Supplementary Dataset 1). Of the SNPs in Table

204 2, the *P0_01974* mutation was identified in both the ancestor population and 6/8

205 control passage lineages, and one of the four identified *arnB* mutations (A96V) was

206 found in 7/8 control passage lineages. All SNPs in Table 2 that were present in lung

207 or nasopharynx lineages but were not found in the ancestor or control passage were

208 subsequently confirmed by Sanger sequencing.

209

210 *Evidence of parallel evolution in niche-passaged pneumococci*

211 Three lineages of lung-passaged pneumococci acquired the same SNP in *P0_00979*,

212 encoding the glycosyltransferase LafA, involved in synthesis of lipid anchors for cell

213 wall teichoic acid[22]. This SNP arose early in lung lineages 4, 5 and 10 and quickly

214 reached fixation within the population, as can be seen in Muller plots depicting

215 genotype frequencies and ancestries (Figure 4A, B). This SNP was not detected in

216 the ancestor or any of the nasopharynx or control passage lineages, suggesting it

217 could confer an advantage within the lung environment.

218

219 For nasopharynx-passaged lineages, two examples of potential parallel evolution

220 were found in the mutations in *P0_01029* found in lineages 1 and 7, and the co-

221 occurring *gpsA* and *yknX* SNPs in lineages 3 and 4. Lineages 1 and 7 each acquired

222 a different mutation in *P0_01029* but both resulted in an amino acid change, at

223 positions 62 and 19, respectively. This gene encodes the surface-expressed

224 pneumococcal vaccine antigen A (*pvaA*), identified in an immunising screen of

225 pneumococcal proteins as affording protection against disseminated infection[23]. Hava

226 and Camilli highlighted this gene as being essential for lung infection[24]. The mutations

227 in *pvaA* arose early in both lineages and quickly reached fixation (Figure 5A and B).

228

229   Carriage lineages 3 and 4 both acquired the same non-synonymous mutations in

230   *gpsA*, encoding a glycerol-3-phosphate dehydrogenase, and *yknX*, encoding a

231   component of an uncharacterised efflux pump. These SNPs became fixed in the

232   populations (Figure 5C). Muller plots, genotype ancestries and genotype frequency

233   plots for all lung and nasopharynx passaged lineages can be found in Supplementary

234   Figures 4-13.

235

236   Several common trends were observed amongst lineages. Between 2 and 22 distinct

237   variants were detected per lineage, per passage, in pneumococci from lungs (Figure

238   6A). Between 4 and 29 distinct variants were detected per lineage, per passage in

239   nasopharynx-passaged pneumococci (Figure 6B). Both lung and nasopharynx

240   passaged lineages had increasing variant numbers in later passages (*p<0.0001* at

241   passage 15 vs passage 1, *p=0.0057* at passage 20 vs passage 1 for lung lineages,

242   *p=0.0113* at passage 15 vs passage 1, *p<0.0001* at passage 20 vs passage 1 for

243   nasopharynx lineages) (Figure 6C). The mean number of variants detected was lower

244   at each passage in lung lineages vs nasopharynx, perhaps reflecting their reduced

245   time spent *in vivo*. Most variants identified were detected only once, that is, in a single

246   lineage at a single passage. However, there was evidence of selection for some SNPs,

247   with ten appearing ≥10 times in nasopharynx lineages and eleven appearing ≥10 times

248   in lung lineages (Figure 6D). Of the total unique variants identified, discounting those

249   present in the ancestor population, 245 were found only in nasopharynx lineages and

250   118 were found only in lung lineages (Figure 6E). Within the 22 variants shared

251   between nasopharynx and lung populations, 13 were also found in control passage

252   lineages.

253

254   *Function of gpsA, a gene associated with adaptation to the nasopharynx*

255   We observed a guanine to adenine substitution in *gpsA*, resulting in a glycine to

256   arginine transition at amino acid 208, in nasopharynx-passaged lineages 3 and 4. We

257   focussed on this gene, because in a recent longitudinal study, assessing

258   microevolution of pneumococcus over 12 months in colonised infants, SNPs in *gpsA*

259   were observed at a frequency of 35.36 per kilobase pair, placing it 23rd out of 592

260   genes in which SNPs were detected, in terms of mutation frequency[25]. The glycerol-

261   3-phosphate dehydrogenase (G3PDH) product of *gpsA* plays a role in membrane lipid

262   metabolism and maintenance of intracellular redox potential[26]. We constructed a

263   targeted deletion of *gpsA*, and individually reproduced the identified mutation that

264   generated the G208R variant, in the ancestor D39 strain, using the sweet Janus

265   system[27].

266

267   We assessed the nasopharyngeal colonisation potential of the resulting D39Δ*gpsA*

268   and D39*gpsA*G208R strains relative to a streptomycin resistant D39 control (D39Smr).

269   Both strains showed comparable levels of growth to D39Smr in nutrient broth

270   (Supplementary Figure 14). However, D39Δ*gpsA* was markedly attenuated in its ability

271   to colonise murine nasopharynx, showing reduced colonisation density relative to

272   D39Smr from day 1 post infection and evidence of clearance as early as day 3 (Figure

273   7A). D39*gpsA*G208R, by contrast, showed prolonged maintenance of carriage, with

274   100% of mice still colonised at day 14 post-infection (Figure 7A).

275

276   Because it has been proposed that GpsA is involved in the modulation of intracellular

277   redox potential[26], we sought to determine whether the G208R substitution in

278   nasopharynx lineages 3 and 4 might impact resistance to reactive oxygen species

279 (ROS). Pneumococci produce ROS as a byproduct of metabolism and are exposed to

280 oxidative stress in the respiratory tract, both from co-colonising microbes and from the

281 actions of neutrophils[28]. A hydrogen peroxide sensitivity assay revealed the

282 significantly elevated susceptibility of D39Δ*gpsA* compared to D39, whereas

283 D39*gpsA*G208R was more resistant to H_2O_2 induced killing (Figure 7B). During long-

284 term exposure to H_2O_2, both D39*gpsA*G208R and a nasopharynx-passaged lineage,

285 containing the same mutation, showed increased resistance, compared to both D39

286 and D39Δ*gpsA* (Figure 7C). Analysis of G3PDH activity confirmed a substantial

287 reduction in D39Δ*gpsA* and a significantly elevated level of activity in D39*gpsA*G208R,

288 relative to D39 (Figure 7D). These data are consistent with the G208R variant of GpsA

289 increasing the activity of G3PDH, and enhancing the nasopharyngeal colonisation

290 potential of *S. pneumoniae* via increased resistance to oxidative stress.

291

Discussion

293 Experimental evolution is a powerful tool for identification of genetic elements under

294 selection within different environments[4]. Cooper and colleagues recently showed the

295 value of this approach for *Streptococcus pneumoniae* in mouse models, using a

296 nasopharyngeal colonisation model similar to that described here, but using a 19F

297 pneumococcal strain (BHN97x)[29]. There are some interesting commonalities in the

298 findings of the two studies. Firstly, despite the bottlenecks occurring during

299 experimental passage, both studies showed evidence of strong selection operating on

300 pneumococcal populations. Parallelism and convergence were evident in

301 independently evolving lineages within both studies and when comparing lineages

302 derived from one study to those from the other, despite the different strain choice.

303

304   Cooper *et al* highlighted the acquisition of frameshift mutations in *mutS* and

305   demonstrated that loss of this gene conferred a hypermutator phenotype. We

306   observed mutations in *mutS, mutS2, mutL* and *mutX*, with those in *mutX* reaching

307   fixation in 2 lineages. Mutations in the *mut* DNA repair genes and in *recN,* which

308   collectively were found in 6 of 20 lineages, might provide an explanation for the

309   increasing numbers of variants we observed per lineage with increasing passage

310   numbers.

311

312   Acquisition of mutations in genes encoding components of ABC transport systems

313   was a further feature of both studies. We highlighted the *yknX* mutation in 2

314   nasopharynx lineages and one lung lineage, but we also observed a mutation in *yknY*

315   that reached fixation in lineage 9. Cooper and colleagues found a mutation in *yknZ* in

316   one nasopharynx-passaged lineage. Collectively, YknXYZ form an ABC transport

317   system involved in resistance to both antibiotics and host-derived antimicrobial

318   peptides [30,31]. The identification of SNPs in the all three genes of this transport system

319   suggests it may play a role in upper-airway colonisation. Further, Cooper *et al* report

320   mutations in *yheS_1, yheS_2* and *yheS_3*, whilst we identified a SNP within *yheH*.

321   These genes also encode ABC transporter components, although the Yhe system is

322   poorly characterised in pneumococcus. Non-synonymous mutations in *yheS_2* were

323   previously identified when comparing paired carriage and disease isolates of serotype

324   22F [32], suggesting that this gene may be involved in niche adaptation. Finally, both

325   studies identified mutations in genes with roles in thiamine, copper and pyruvate

326   uptake or metabolism, highlighting the key requirement for trace metal scavenging and

327   regulation of glycolysis during colonisation [33].

328

329   Although experiments were established with D39 stocks that originated from a single

330   colony, we identified some potentially important genetic variation within the population

331   used to found the twenty lineages. Comparison of genetic ancestries of the lineages

332   (Figures 4, 5 and Supplementary Figure 4-13) suggests that the sub-populations of

333   the ancestor bearing the *P0_01974* and *dnaK* mutations were distinct, with each

334   subsequently giving rise to separate genotypes. One of these often drove the other to

335   extinction and, beyond passage 10, most lineages had lost either the *dnaK*-derived

336   genotypes or those arising from the *P0_01974* genotype. In lung lineages, the *dnaK*

337   mutation present in the ancestor became fixed in 8/10 lineages but was lost from the

338   other two, where instead the *P0_01974* mutation became fixed. In nasopharynx

339   lineages, both mutations were eventually lost in 5/10 lineages. Examples of all three

340   outcomes are shown in Supplementary Figure 15. It is possible that, within the

341   nasopharynx, the *dnaK* or *P0_01974* mutations themselves did not significantly affect

342   fitness and that the success of one or the other genotype came down to which was

343   first to acquire a beneficial mutation, or to survive a population bottleneck. Intriguing,

344   however, is the relative success of the *dnaK* mutation in lung lineages, where it

345   became fixed in eight of ten lineages.

346

347   Although the lung is not the primary niche of pneumococcus, we opted to

348   experimentally passage D39 through the lung environment, as colonisation of this

349   niche is associated with severe disease in humans. The nasopharynx and lungs differ

350   in several important ways, including temperature, pH, availability of nutrients and

351   degree of immune surveillance [34-36]. Pneumococcal lineages differ in their abilities to

352   cause pneumonia and invasive pneumococcal disease [37], consistent with some being

353   better adapted to thrive in the lung environment than others. By bypassing the carriage

354 step that would ordinarily precede lung infection we hoped to select for traits

355 specifically advantageous in the lung environment that might, therefore, provide

356 information about how the pneumococcus is able to cause disease. We observed only

357 ten variants that were shared by nasopharynx and lung lineages but were not present

358 in the control passage or ancestor, as compared to 245 variants unique to

359 nasopharynx lineages and 118 unique to lung. Whilst it is likely that only a small

360 fraction of identified variants could be considered adaptive mutations under positive

361 selection, it was notable that the examples of parallelism or convergence we identified

362 tended to be niche-specific. Mutations in *gpsA* and *pvaA* were only observed in

363 nasopharynx lineages, whilst those in *mutX1, lafA* and *sdhB* were unique to lung

364 lineages. The finding that nasopharynx-evolved lineages, but not lung-evolved

365 lineages, showed enhanced carriage potential further suggests that adaptations are

366 at least partially niche-specific. Only *yknX* showed evidence of selection in both

367 niches, with a mutation causing a leucine to valine change at position 333 becoming

368 fixed in one lung lineage, and a SNP causing a serine to asparagine substitution at

369 position 331 becoming fixed in two nasopharynx lineages.

370

371 The same non-synonymous mutation in *gpsA* was identified in two nasopharynx

372 lineages and a recent study of pneumococcal carriage in infants in the Gambia also

373 identified a high frequency of mutation in this gene [25] (Supplementary Table 2).

374 Deletion of *gpsA* in the ancestor strain compromised nasopharyngeal carriage, whilst

375 the identified SNP conferred increased carriage potential and resistance to oxidative

376 stress, when reproduced individually in D39. Glycerol 3-phosphate dehydrogenase

377 reduces $NADP_+$ to NADPH during the conversion of sn-glycerol 3-phopshate to

378 glycerine phosphate, thereby helping maintain intracellular redox potential. Increased

379 redox control during carriage may be beneficial in the context of competition with co-

380 colonising microbes or in resistance to neutrophil responses. Mutations conferring

381 such phenotypes may also have consequences for antimicrobial resistance in

382 pneumococcus, as antibiotics including penicillin and ciprofloxacin induce production

383 of ROS[26,38].

384

385 We identified increased adhesion to airway epithelial cells in three of ten nasopharynx-

386 passaged lineages and reduced pneumolysin production in nine of ten lineages. Both

387 these traits could compromise transmission whilst promoting stable carriage. An

388 assessment of transmission potential in niche-adapted lineages will be an important

389 next step for this work. If nasopharynx-passaged lineages have increased potential for

390 non-invasive carriage together with loss of transmission potential, then they may be

391 suitable starting points for the development of live-attenuated pneumococcal vaccines,

392 although requiring further laboratory attenuation to prevent reversion to virulence or

393 acquisition of new traits *in vivo* via recombination.

394

395 An open question is the extent to which the genetic changes we observed might be

396 constrained in a natural infection setting. In our model, the process of transmission –

397 associated with tight bottlenecks[16] – was bypassed by manual transfer of

398 pneumococci between animals. The need to balance attachment and colonisation with

399 shedding and transmission is likely a major determining factor in the trajectory of

400 pneumococcal evolution and the trade-off between transmissibility and within-host

401 fitness has been recently highlighted[39]. The authors outline the role played by

402 competitive exclusion in restricting population diversity, whereby the first variant to

403 colonise prevents later arrivals from doing so. The extent to which this phenomenon

404   might impact evolution in our experimental system, which utilised manual

405   administration of the bacterial population as opposed to natural exposure, is unclear.

406   We did observe loss of some genotypes between passages but also instances where

407   multiple, co-existing genotypes were maintained for several passages.

408

409   One important caveat to the findings presented relates to the anatomical, biochemical

410   and immunological differences between the mouse and human respiratory tracts. In

411   addition to adapting pneumococci to the nasopharynx or to the lungs, the passage

412   process we describe might select for variants that give an advantage specific to

413   colonisation of murine respiratory tissue. We are, however, encouraged by a recent

414   study in humans, analysing microevolution of pneumococci during natural colonisation

415   of infants[25]. Twenty (43%) of the genes in which we observed fixed non-synonymous

416   SNPs in nasopharynx-passaged lineages were also targets of mutation during

417   prolonged colonisation in infants (Supplementary Table 1). These included *gpsA* and

418   *yknX*, in which we observed parallel evolution. Considering all the genes in which we

419   identified gene variants (synonymous or non-synonymous SNPs or indels) in

420   nasopharynx passaged lineages, 39% (68/174) were also identified in the dataset of

421   Chaguza and colleagues (Supplementary Dataset 2).

422

423   Collectively, our findings suggest that *in vivo* experimental evolution can be a valuable

424   tool for the study of host-pathogen interactions and disease mechanisms. The local

425   environment shapes pneumococcal evolution, and differences between the niches of

426   the upper and lower airway are sufficient to alter the trajectory of evolution in

427   populations of bacteria exposed to one or the other. We identified several examples

428   of parallel evolution in independently passaged populations and the genes under

429   selection highlight the importance of cell surface modification and metabolic versatility

430   in pneumococcal colonisation and disease processes. *In vivo* experimental evolution

431   approaches can be used to identify pathogen genes with roles in within-host

432   processes, as exemplified by *gpsA*, that appeared to be under strong selection in

433   nasopharynx and in which we identified a single nucleotide change conferring

434   phenotypic change and increased within-host fitness.

435

436   **Methods**

437   **Media and growth conditions**

438   *S. pneumoniae* D39 (serotype 2, NCTC 7466) isolates were grown at 37°C in 5% CO_2

439   on blood agar base (BAB) (Oxoid) supplemented with 5% (vol/vol) defibrinated horse

440   blood (Oxoid) and 1 µg/ml of gentamycin and overnight cultures were prepared in 10

441   ml of brain heart infusion (BHI) broth (Sigma). Bacteria were identified as pneumococci

442   by α-hemolysis on blood agar and by optochin sensitivity. The D39 used in this study

443   was generously provided by Prof Tim Mitchell (University of Birmingham, UK).

444

445   **Ethics statement**

446   All mouse infection work was performed at the University of Liverpool with prior

447   approval by the UK Home Office (project licence PB6DE83DA) and the University of

448   Liverpool Ethics Committee. For all experiments, 6 to 8 week old female CD1 mice

449   were purchased from Charles River (Margate, UK). Mice were randomly assigned to

450   a cage of 1-5 mice, depending on the experiment, on arrival at the unit, by staff with

451   no role in study design. Mice were housed in individually ventilated cages and allowed

452   to acclimatise for 7 days prior to infection.

453

Experimental evolution in mouse infection models

A standardised suspension of the D39 ancestor strain was prepared for inoculation by overnight growth in BHI from a single colony. The next day, sub-culture was performed and the bacteria allowed to grow to mid-exponential phase (~6 hours) before storage at -80°C. For the initial infection, this suspension was thawed at room temperature and bacteria were harvested by centrifugation and suspended in PBS. For the carriage model, 10 mice were intranasally infected under light anaesthesia, using a mix of oxygen and isoflurane, with 1×10^5 colony forming units (cfu) in 10 µl saline. To induce pneumonia, 10 mice were infected intranasally with 1.5×10^6 cfu, in 50 µl saline, to ensure pneumococci reached and colonised the lung. The 20 infected mice were treated as founders of 20 independent pneumococcal lineages, 10 of which were passaged through nasopharynx and 10 which were passaged through lung.

In the carriage model, the bacteria remained in the nasopharynx for 7 days for each passage and, for the pneumonia model, bacteria were in the lung environment for 48hrs, or less if the mice reached the severity endpoint before this time. The time the lineages spent in the lungs ranged from 22-48 hours per passage, with an average time of 31.91 hours. After each passage, pneumococci were recovered from the nasopharynx, in the carriage model, or lungs, in the pneumonia model. After minimal passage on gentamicin BAB agar with an optochin disk to confirm identity as *S. pneumoniae*, the entire population of recovered pneumococci for each lineage was resuspended into 1 ml of BHI/10% glycerol, which was split 500:250:250 between three tubes and stored at -80°C for future use in infection experiments, phenotypic analysis or genome sequencing. One of the 250 µl frozen aliquot tubes was used to determine the cfu by Miles and Misra and to re-infect a new mouse for the next round

479   of passage. This passage process was repeated a total of 20 times for each of the

480   separately evolving lineages in the carriage and pneumonia models of infection.

481   Figure 1 gives an overview of the *in vivo* experimental evolution process undertaken

482   during this study.

483

484   **Control laboratory passage on BAB agar**

485   The D39 ancestor freezer stock used for the initial infection in the *in vivo* evolution

486   experiments was thawed and used to streak 10 gentamicin BAB plates with an

487   optochin disc and incubated for 20-24 hours at 37°C in 5%. For the next passage for

488   each lineage, the population was re-streaked onto a fresh gentamicin BAB plate and

489   this process was repeated 20 times for the 10 separately evolving lineages. Freezer

490   stocks were made from passages 1, 5, 10, 15 and 20 and stored for future analysis.

491   In the final analysis, only 8 of the 10 lineages were included due to contamination of

492   the DNA used for sequencing.

493

494   **Illumina sequencing and read processing**

495   DNA extraction protocols are detailed in supplementary methods. Short read

496   sequencing was performed to produce 151-bp, paired-end reads on an Illumina

497   HiSeq2500 at the Wellcome Trust Sanger Institute, UK. This was carried out for the

498   ancestor strain and for passages 1, 5, 10, 15 and 20, for both the nasopharynx and

499   lung evolved lineages. Short read sequencing was also carried out for the passage 20

500   control lineages. To obtain a deep sequence coverage of the ancestor D39 strain,

501   DNA was paired-end sequenced on an Illumina NextSeq 500 system using 150-bp

502   read lengths, by Vertis Biotechnologie AG, Germany. All bioinformatic analysis of

503   sequencing data was carried out using a virtual machine hosted by the Cloud

504   Infrastructure for Microbial Bioinformatics (CLIMB) consortium[40]. Raw Illumina reads

505   were processed using Trim Galore (v0.4.4) with Cutadapt (v1.9.1) for paired-end

506   reads, using default settings, to remove short low quality reads (< 20 bp), trim Illumina

507   adaptor sequences, and eliminate poor-quality bases from the sequences (< Q20).

508   FastQC (v0.11.5) confirmed that the resulting trimmed reads were of sufficient quality

509   for subsequent analysis.

510

511   **PacBio long read sequencing and genome assembly**

512   For the ancestor strain, single molecule real-time (SMRT) sequencing (Pacific

513   Biosciences, PacBio, CA, USA) of the ancestor D39 isolate was performed at the

514   Wellcome Trust Sanger Institute (Cambridge, UK). The resulting BAM file was

515   converted to FASTQ with bamToFastq (v2.28.0). PacBio long reads were filtered by

516   quality using Filtlong (v0.2.0- https://github.com/rrwick/Filtlong) and assembled with

517   the Flye Assembler (v2.4.2) [41]; default settings were applied for both programmes. The

518   resulting contig, with 469x coverage depth, was polished using the PacBio SMRT

519   minimap2   tool   (v2.15-r905)   and   Racon   Polisher   (v1.3.3   -

520   https://github.com/isovic/racon), with default settings applied. A consensus sequence

521   of the ancestor was generated using PacBio SMRT tools Blasr alignment (v5.1) and

522   Arrow (v2.3.3). This was then corrected by aligning the sequence with the ancestor

523   Illumina reads using bwa mem (v0.7.17-r1188) and re-assembling with Pilon (v1.22)

524   [42]. The resulting corrected consensus sequence was annotated using Prokka (v1.13)[19]

525   and run through the Quality Assessment Tool for Genome Assemblies (QUAST,

526   v4.6.3) to obtain the assembly statistics. Comparison of Prokka annotation with the

527   recent D39 genome assembly of Slager *et al.*[18] is detailed in supplementary methods.

528

529   **Variant calling**

530   Variants were detected from the processed Illumina reads for the nasopharynx and

531   lung evolved lineages at passages 1, 5, 10, 15 and 20, as well as the control lineages

532   at passage 20. The short sequence reads were mapped to the D39 ancestor corrected

533   consensus sequence to identify mutations, using the Breseq programme (v0.31.1)[43].

534   Default settings were applied with the predict-polymorphisms function. This process

535   identified the percentage frequency of variants within a lineage population at each

536   passage. Synonymous and non-synonymous single nucleotide polymorphisms

537   (SNPs) as well as insertions/deletions and intergenic mutations were detected. The

538   deeper sequenced Illumina reads for the ancestor strain (x300 coverage) were also

539   used to detect variants within the original D39 ancestor inoculum, with the

540   polymorphism-frequency-cut off parameter set to >2%.

541

542   **Comparison with clinical isolate data**

543   We analysed pneumococcal genomes from a recently published dataset of isolates

544   sampled longitudinally in newborn infants, to assess whether the mutations identified

545   during experimental evolution occur frequently within hosts during natural carriage

546   episodes[25]. We mapped the reads of each isolate against the ancestral D39 reference

547   genome and generated a consensus alignment for each episode using Snippy (v4.3.6)

548   (https://github.com/tseemann/snippy). The genomic positions with SNPs during each

549   carriage episode were identified, annotated, and compared between episodes to

550   identify episodes with changes occurring in parallel at the same location using

551   BioPython[44].

552

553   **Survival and time point pneumonia experiments**

554   Mice were infected with standardised pneumococcal isolate stocks at a dose of 1 x

555   10^6 cfu in 50 μl saline via intranasal administration under light anaesthesia, to induce

556   acute pneumococcal disease. Mice were monitored for signs of disease and pain

557   score was determined using the scheme of Morton[45]. For survival experiments, mice

558   were monitored for 7 days and culled once lethargic. For time point experiments, mice

559   were culled at 0, 12, 24 and 36 hours post-infection or if they reached the experimental

560   endpoint (lethargy). Lungs and blood samples were taken and pneumococcal cfu

561   determined via Miles and Misra dilution onto gentamycin BAB plates. Construction of

562   SNP and deletion mutants for use in these experiments is described in Supplementary

563   Methods.

564

565   **Carriage time point experiments**

566   Mice were infected as described above but with a dose of 1 x 10^5 cfu in 10 μl saline to

567   induce asymptomatic nasopharyngeal colonisation. At 1, 3, 7, 14 and 21 days post

568   infection, mice were culled, nasopharynx samples taken and pneumococcal cfu

569   determined.

570

571   **Analysing pneumococcal growth dynamics in the presence of hydrogen**

572   **peroxide**

573   Growth assessment was performed as detailed in Supplementary Methods, with the

574   following modifications. Pneumococcal isolates were grown in triplicate in BHI in 96-

575   well plates in the presence of hydrogen peroxide (H_2O_2, Sigma) at concentrations 0,

576   2.5, 5.0 and 7.5 mM. Growth dynamics were analysed with the R package

577   GrowthCurver and the percentage reduction in the AUC values for each isolate in the

578   presence of H_2O_2 was calculated in comparison to 0 mM.

579

Hydrogen peroxide sensitivity assays

580

Mid-log phase pneumococci were added in duplicate to microtiter plate wells and
incubated at 37°C for 30 minutes in BHI with 12 mM of H_2O_2. The CFU/mL of each
isolate following the incubation was determined via Miles and Misra dilution onto
gentamycin BAB plates and the percentage killing was calculated for each test isolate
in comparison to the untreated (BHI only) control.

586

Glycerol-3-phosphate dehydrogenase assay

587

G3PDH activity was measured using a colorimetric assay kit (Abcam). Overnight
cultures of isolates were centrifuged for 15 mins at 3000 rpm and the pellet
resuspended in 200 ul of G3PDH assay buffer and lysed with the addition of sodium
deoxycholate at 10%. This was incubated at room temperature for 10 minutes,
centrifuged for 15 minutes at 3000 rpm and the supernatant used for the assay
following the manufacturer's instructions. Absorbance readings at 450 nm were taken
at time zero and following a 60 minute incubation at 37°C and the activity of G3PDH
for each isolate was calculated by comparison to a NADH standard curve.

596

Statistical analysis

597

GraphPad Prism version 8.2.1 was used for statistical analysis. Statistical tests
undertaken for individual experiments are detailed in the respective figure legends.
$p < 0.05$ was considered to be statistically significant. Data were tested for normality
and to define the variance of each group tested. All multi-parameter analyses included
corrections for multiple comparisons and data are presented as mean ± standard
deviation (SD) unless otherwise stated.

604

Author contributions

605

606   DN conceived the study. AG, JH, SB, JR and DN designed experiments. AG, FL, DH,

607   CM, HE and DN performed experiments. AG, CC, FL, TB, JH, JR and DN analysed

608   data. SB, JR and DN provided resources and reagents and supervised staff. AG and

609   DN wrote the manuscript, with input from all authors.

610

622

623   **References**

624   1   Drijkoningen, J. J. & Rohde, G. G. Pneumococcal infection in adults: burden of disease.
625   *Clin Microbiol Infect* **20 Suppl 5**, 45-51, doi:10.1111/1469-0691.12461 (2014).
626   2   Brooks, L. R. K. & Mias, G. I. Streptococcus pneumoniae's Virulence and Host
627   Immunity: Aging, Diagnostics, and Prevention. *Front Immunol* **9**, 1366,
628   doi:10.3389/fimmu.2018.01366 (2018).
629   3   Lipsitch, M. & Moxon, E. R. Virulence and transmissibility of pathogens: what is the
630   relationship? *Trends Microbiol* **5**, 31-37, doi:10.1016/S0966-842X(97)81772-6 (1997).

631 4   McDonald, M. J. Microbial Experimental Evolution - a proving ground for evolutionary
632   theory and a tool for discovery. *EMBO Rep* **20**, e46992,
633   doi:10.15252/embr.201846992 (2019).
634 5   Scheiblauer, H., Kendal, A. P. & Rott, R. Pathogenicity of influenza A/Seal/Mass/1/80
635   virus mutants for mammalian species. *Arch Virol* **140**, 341-348,
636   doi:10.1007/BF01309867 (1995).
637 6   Woolcock, P. R. & Crighton, G. W. Duck virus hepatitis: serial passage of attenuated
638   virus in ducklings. *Vet Rec* **105**, 30-32, doi:10.1136/vr.105.2.30 (1979).
639 7   Diffley, P., Scott, J. O., Mama, K. & Tsen, T. N. The rate of proliferation among African
640   trypanosomes is a stable trait that is directly related to virulence. *Am J Trop Med Hyg*
641   **36**, 533-540, doi:10.4269/ajtmh.1987.36.533 (1987).
642 8   Dearsly, A. L., Sinden, R. E. & Self, I. A. Sexual development in malarial parasites:
643   gametocyte production, fertility and infectivity to the mosquito vector. *Parasitology*
644   **100 Pt 3**, 359-368, doi:10.1017/s0031182000078628 (1990).
645 9   Berry, A. M., Yother, J., Briles, D. E., Hansman, D. & Paton, J. C. Reduced virulence of a
646   defined pneumolysin-negative mutant of Streptococcus pneumoniae. *Infect Immun*
647   **57**, 2037-2042, doi:https://www.ncbi.nlm.nih.gov/pubmed/2731982 (1989).
648 10   Hirst, R. A., Kadioglu, A., O'Callaghan, C. & Andrew, P. W. The role of pneumolysin in
649   pneumococcal pneumonia and meningitis. *Clin Exp Immunol* **138**, 195-201,
650   doi:10.1111/j.1365-2249.2004.02611.x (2004).
651 11   McNeela, E. A. *et al.* Pneumolysin activates the NLRP3 inflammasome and promotes
652   proinflammatory cytokines independently of TLR4. *PLoS Pathog* **6**, e1001191,
653   doi:10.1371/journal.ppat.1001191 (2010).
654 12   Jefferies, J. M. *et al.* Presence of nonhemolytic pneumolysin in serotypes of
655   Streptococcus pneumoniae associated with disease outbreaks. *J Infect Dis* **196**, 936-
656   944, doi:10.1086/520091 (2007).
657 13   Zafar, M. A., Wang, Y., Hamaguchi, S. & Weiser, J. N. Host-to-Host Transmission of
658   Streptococcus pneumoniae Is Driven by Its Inflammatory Toxin, Pneumolysin. *Cell
659   Host Microbe* **21**, 73-83, doi:10.1016/j.chom.2016.12.005 (2017).
660 14   Zafar, M. A., Kono, M., Wang, Y., Zangari, T. & Weiser, J. N. Infant Mouse Model for
661   the Study of Shedding and Transmission during Streptococcus pneumoniae
662   Monoinfection. *Infect Immun* **84**, 2714-2722, doi:10.1128/IAI.00416-16 (2016).
663 15   Matthias, K. A., Roche, A. M., Standish, A. J., Shchepetov, M. & Weiser, J. N.
664   Neutrophil-toxin interactions promote antigen delivery and mucosal clearance of
665   Streptococcus pneumoniae. *J Immunol* **180**, 6246-6254,
666   doi:10.4049/jimmunol.180.9.6246 (2008).
667 16   Kono, M. *et al.* Single Cell Bottlenecks in the Pathogenesis of Streptococcus
668   pneumoniae. *PLoS Pathog* **12**, e1005887, doi:10.1371/journal.ppat.1005887 (2016).
669 17   Lanie, J. A. *et al.* Genome sequence of Avery's virulent serotype 2 strain D39 of
670   Streptococcus pneumoniae and comparison with that of unencapsulated laboratory
671   strain R6. *J Bacteriol* **189**, 38-51, doi:10.1128/JB.01148-06 (2007).
672 18   Slager, J., Aprianto, R. & Veening, J. W. Deep genome annotation of the opportunistic
673   human pathogen Streptococcus pneumoniae D39. *Nucleic Acids Res* **46**, 9971-9989,
674   doi:10.1093/nar/gky725 (2018).
675 19   Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-
676   2069, doi:10.1093/bioinformatics/btu153 (2014).

677 20   Deatherage, D. E. & Barrick, J. E. Identification of mutations in laboratory-evolved
678   microbes from next-generation sequencing data using breseq. *Methods Mol Biol* **1151**,
679   165-188, doi:10.1007/978-1-4939-0554-6_12 (2014).
680 21   Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool
681   for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-
682   36, doi:10.1093/nar/28.1.33 (2000).
683 22   Engholm, D. H., Kilian, M., Goodsell, D. S., Andersen, E. S. & Kjaergaard, R. S. A visual
684   review of the human pathogen Streptococcus pneumoniae. *FEMS Microbiol Rev* **41**,
685   854-879, doi:10.1093/femsre/fux037 (2017).
686 23   Wizemann, T. M. *et al.* Use of a whole genome approach to identify vaccine molecules
687   affording protection against Streptococcus pneumoniae infection. *Infect Immun* **69**,
688   1593-1598, doi:10.1128/IAI.69.3.1593-1598.2001 (2001).
689 24   Hava, D. L. & Camilli, A. Large-scale identification of serotype 4 Streptococcus
690   pneumoniae virulence factors. *Mol Microbiol* **45**, 1389-1406,
691   doi:https://www.ncbi.nlm.nih.gov/pubmed/12207705 (2002).
692 25   Chaguza, C. *et al.* Within-host microevolution of Streptococcus pneumoniae is rapid
693   and adaptive during natural colonisation. *Nat Commun* **11**, 3442, doi:10.1038/s41467-
694   020-17327-w (2020).
695 26   Lupien, A. *et al.* Genomic characterization of ciprofloxacin resistance in a laboratory-
696   derived mutant and a clinical isolate of Streptococcus pneumoniae. *Antimicrob Agents*
697   *Chemother* **57**, 4911-4919, doi:10.1128/AAC.00418-13 (2013).
698 27   Li, Y., Thompson, C. M. & Lipsitch, M. A modified Janus cassette (Sweet Janus) to
699   improve allelic replacement efficiency by high-stringency negative selection in
700   Streptococcus pneumoniae. *PLoS One* **9**, e100510, doi:10.1371/journal.pone.0100510
701   (2014).
702 28   Yesilkaya, H., Andisi, V. F., Andrew, P. W. & Bijlsma, J. J. Streptococcus pneumoniae
703   and reactive oxygen species: an unusual approach to living with radicals. *Trends*
704   *Microbiol* **21**, 187-195, doi:10.1016/j.tim.2013.01.004 (2013).
705 29   Cooper, V. S. *et al.* Experimental Evolution In Vivo To Identify Selective Pressures
706   during Pneumococcal Colonization. *mSystems* **5**, doi:10.1128/mSystems.00352-20
707   (2020).
708 30   Majchrzykiewicz, J. A., Kuipers, O. P. & Bijlsma, J. J. Generic and specific adaptive
709   responses of Streptococcus pneumoniae to challenge with three distinct antimicrobial
710   peptides, bacitracin, LL-37, and nisin. *Antimicrob Agents Chemother* **54**, 440-451,
711   doi:10.1128/AAC.00769-09 (2010).
712 31   Yang, H. B. *et al.* Structure of a MacAB-like efflux pump from Streptococcus
713   pneumoniae. *Nat Commun* **9**, 196, doi:10.1038/s41467-017-02741-4 (2018).
714 32   Cleary, D. W. *et al.* Comparative Genomics of Carriage and Disease Isolates of
715   Streptococcus pneumoniae Serotype 22F Reveals Lineage-Specific Divergence and
716   Niche Adaptation. *Genome Biol Evol* **8**, 1243-1251, doi:10.1093/gbe/evw066 (2016).
717 33   McAllister, L. J. *et al.* Molecular analysis of the psa permease complex of Streptococcus
718   pneumoniae. *Mol Microbiol* **53**, 889-901, doi:10.1111/j.1365-2958.2004.04164.x
719   (2004).
720 34   Man, W. H., de Steenhuijsen Piters, W. A. & Bogaert, D. The microbiota of the
721   respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol* **15**, 259-270,
722   doi:10.1038/nrmicro2017.14 (2017).

723 35 Walsh, R. L. & Camilli, A. Streptococcus pneumoniae is desiccation tolerant and
724   infectious upon rehydration. *mBio* **2**, e00092-00011, doi:10.1128/mBio.00092-11
725   (2011).
726 36 Aprianto, R., Slager, J., Holsappel, S. & Veening, J. W. High-resolution analysis of the
727   pneumococcal transcriptome under a wide range of infection-relevant conditions.
728   *Nucleic Acids Res* **46**, 9990-10006, doi:10.1093/nar/gky750 (2018).
729 37 Sleeman, K. L. *et al.* Capsular serotype-specific attack rates and duration of carriage of
730   Streptococcus pneumoniae in a population of children. *J Infect Dis* **194**, 682-688,
731   doi:10.1086/505710 (2006).
732 38 Fani, F., Leprohon, P., Legare, D. & Ouellette, M. Whole genome sequencing of
733   penicillin-resistant Streptococcus pneumoniae reveals mutations in penicillin-binding
734   proteins and in a putative iron permease. *Genome Biol* **12**, R115, doi:10.1186/gb-
735   2011-12-11-r115 (2011).
736 39 Jacobs, N. T. & Weiser, J. N. Competitive exclusion strengthens selection for
737   transmissibility and increases the benefit of recombination for within-host adaptation.
738   *bioRxiv* (2020).
739 40 Connor, T. R. *et al.* CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an
740   online resource for the medical microbiology community. *Microb Genom* **2**, e000086,
741   doi:10.1099/mgen.0.000086 (2016).
742 41 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
743   using repeat graphs. *Nat Biotechnol* **37**, 540-546, doi:10.1038/s41587-019-0072-8
744   (2019).
745 42 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
746   detection and genome assembly improvement. *PLoS One* **9**, e112963,
747   doi:10.1371/journal.pone.0112963 (2014).
748 43 Barrick, J. E. *et al.* Identifying structural variation in haploid microbial genomes from
749   short-read resequencing data using breseq. *BMC Genomics* **15**, 1039,
750   doi:10.1186/1471-2164-15-1039 (2014).
751 44 Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular
752   biology and bioinformatics. *Bioinformatics* **25**, 1422-1423,
753   doi:10.1093/bioinformatics/btp163 (2009).
754 45 Morton, D. B. Pain and laboratory animals. *Nature* **317**, 106,
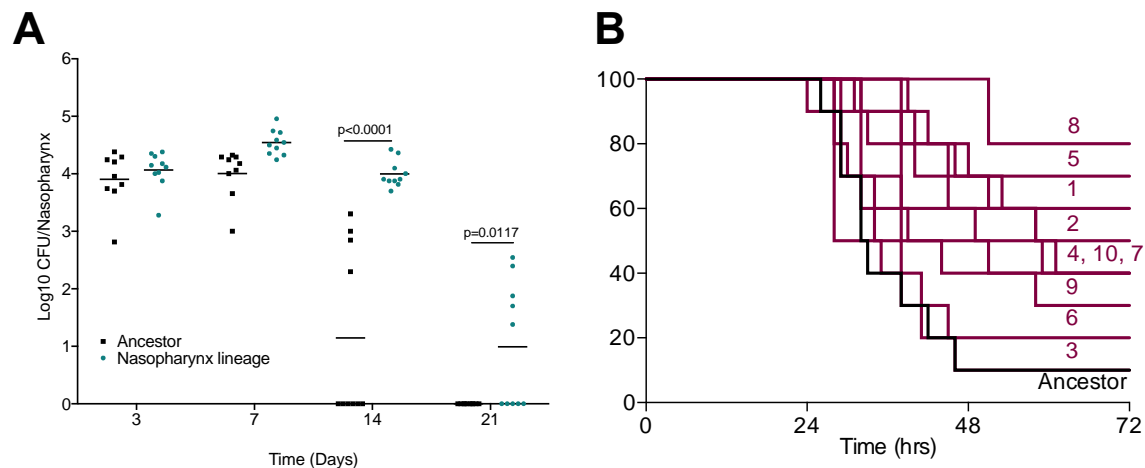755   doi:https://www.ncbi.nlm.nih.gov/pubmed/4033793 (1985).
756
757
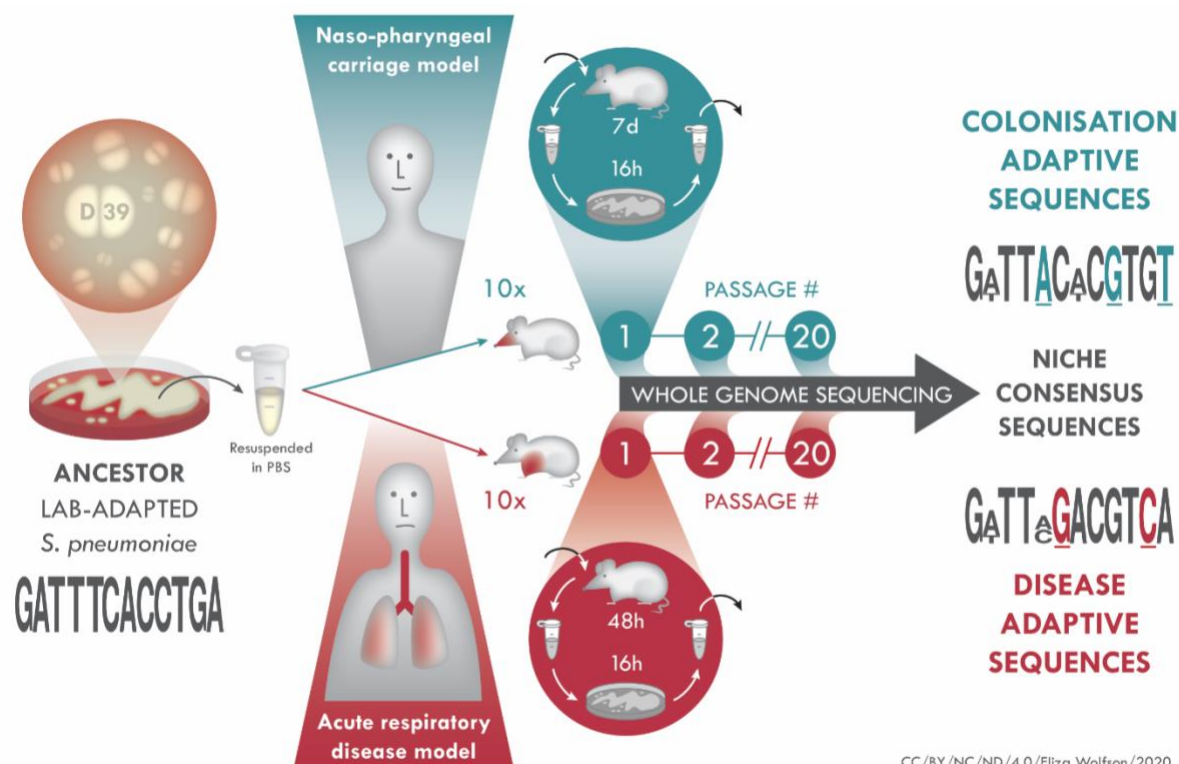
Figure 1. *In vivo* experimental evolution of *S. pneumoniae*. D39 (ancestor) was grown on agar, colonies resuspended in PBS and mice intranasally infected with 10 µl (carriage model, 10 mice) or 50 µl (pneumonia model, 10 mice) inoculum. The 20 infected mice were treated as founders of independent pneumococcal lineages. After each passage, which was 7 days for carriage, 48 hrs for pneumonia, pneumococci were recovered from nasopharynx, in the carriage model, or lungs, in the pneumonia model and, after minimal passage on agar, introduced into a new mouse for the next round of passage. At every passage and at the final (20x) passage, pneumococci were stored for future analysis, including whole genome sequencing for identification of adaptive mutations associated with colonisation of nasopharynx or lungs.

770 ***Table 1. Phenotypes of nasopharynx-passaged and lung-passaged S. pneumoniae lineages.***
771 Female MF1 mice were infected with the ancestor D39 or one of the nasopharynx- or lung-passaged
772 lineages. Nasopharynx-passaged lineages were compared in a carriage model, infecting with 1×10^5
773 colony forming units (cfu) in 10 µl saline. Lung-passaged lineages were compared in a pneumonia
774 model, infecting with 1×10^6 cfu in 50 µl saline. N=10 per group for all. Colonisation was confirmed at
775 day 14 (carriage) or day 3 (pneumonia) by plating nasopharynx or lung homogenates on blood agar.
776 S.d. is standard deviation. P-values are from N-1 Chi squared test (carriage proportions) or Mantel-Cox
777 log-rank test (pneumonia survival) vs the ancestor. Significant differences are shown in red.
778

| Nasopharynx-adapted lineage	Percentage mice colonised at 14 days (p-value vs ancestor)	Nasopharynx colonisation density at 14 days, colonised mice only (log10 mean cfu (s.d.))
Ancestor (D39)	40	2.860 (0.421)
1	100 (*0.004)	3.995 (0.234)
2	90 (*0.022)	3.549 (0.651)
3	80 (0.075)	3.484 (0.890)
4	90 (*0.022)	3.086 (0.804)
5	80 (0.075)	2.928 (0.632)
6	90 (*0.022)	3.418 (0.485)
7	80 (0.075)	4.138 (0.882)
8	50 (0.661)	2.818 (0.596)
9	100 (*0.004)	3.134 (0.653)
10	70 (0.189)	3.360 (0.880)
Lung-adapted lineage	**Percentage mice surviving to 72 hours (p-value vs ancestor)**	**Percentage surviving mice still colonised (log10 mean CFU)**
Ancestor (D39)	10	0 (0)
1	60 (**0.0033)	100 (5.124)
2	50 (**0.0037)	100 (3.622)
3	10 (0.8706)	100 (5.115)
4	40 (*0.0185)	25 (0.781)
5	70 (**0.0027)	57 (1.830)
6	20 (0.5989)	100 (3.975)
7	40 (0.0669)	75 (3.325)
8	80 (****0.0001)	75 (2.275)
9	30 (0.1213)	33 (0.722)
10	40 (0.0966)	50 (2.144)

779

780

781

782
Figure 2. Tissue-passaged pneumococci show evidence of niche adaptation. Mice were infected with the ancestral D39 population or twenty-times passaged lineages from **(A)** nasopharyngeal carriage (lineage 1) or **(B)** pneumonia models. **(A)** Mice were infected with 1×10^5 colony forming units (cfu) of the ancestor or nasopharynx lineage 1 to induce asymptomatic carriage. Bacterial numbers in nasopharynx were determined by serial dilution of tissue homogenates onto blood agar at days 3, 7, 14 and 21 post infection. P-values are from two-way ANOVA with Sidak's multiple comparison test. **(B)** Mice were infected with 1×10^6 cfu of the ancestor or one of the ten lung-passaged lineages to induce pneumonia (n=10 per group). Mice were monitored for signs of disease and culled once lethargic. Survival curves were compared by Mantel-Cox log-rank test (Table 1). Labels show lineage numbers.

792
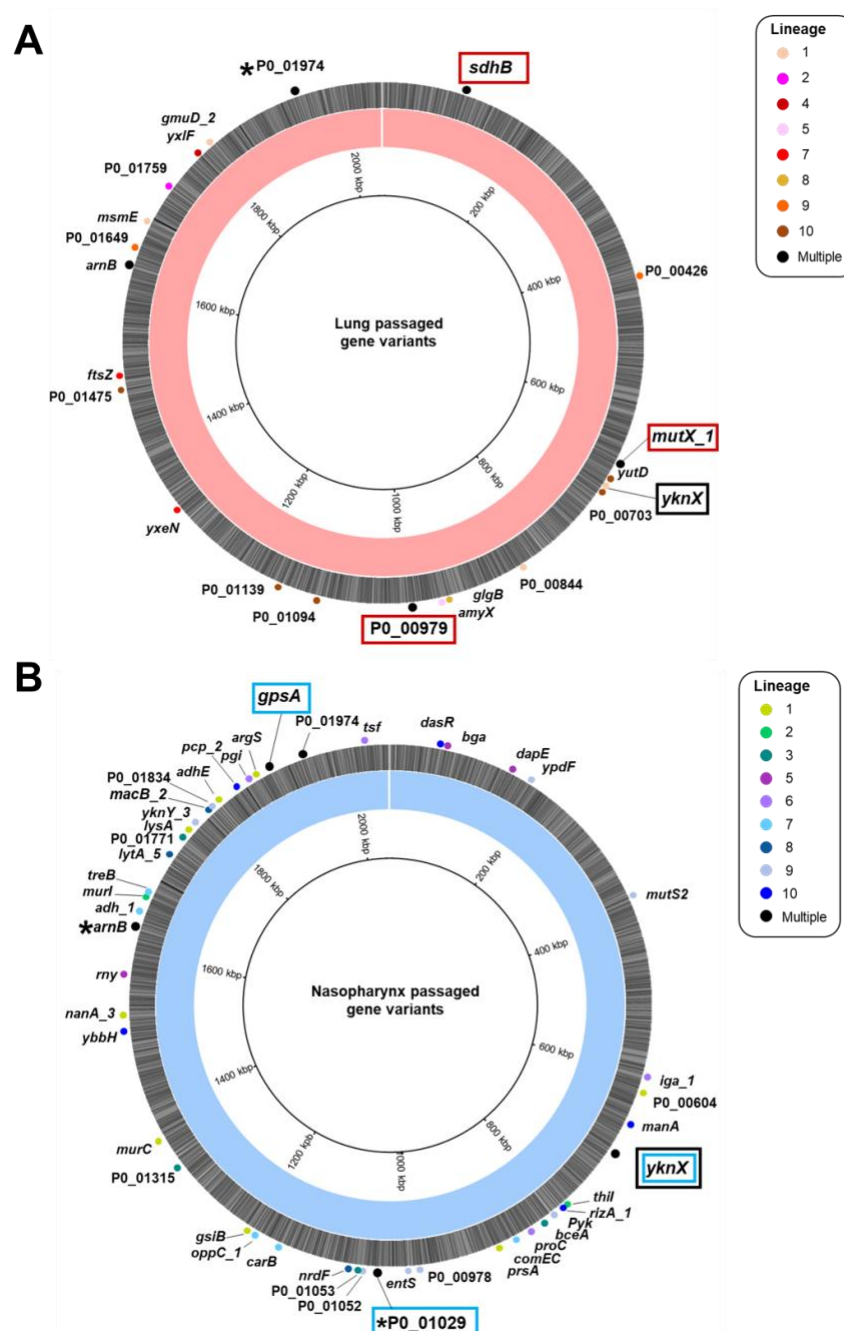
Figure 3. Fixed non-synonymous mutations identified in twenty-times passaged *S. pneumoniae* lineages evolved in mouse lungs or nasopharynx. Genome maps of the D39 ancestor strain assembly, highlighting the non-synonymous single nucleotide polymorphisms identified in twenty-times passaged pneumococci from the lung **(A)** or nasopharynx **(B)**. Labels show the name and position of the gene in which the SNP occurs, with the circles coloured according to the lineage where the variant was identified. Black indicates SNPs identified in multiple lineages. Potential examples of parallel evolution are highlighted in red boxes for the lung and blue boxes for the nasopharynx evolved lineages. The same SNP in the *yknX* gene was identified in two nasopharynx lineages, with a separate SNP identified in one lung lineage. This gene is highlighted with a black box. The mutations for all genes are defined in Supplementary Dataset 1 and *'s indicate those genes where different mutations were observed in separate lineages. Variants in *arnB* and *P0_01974* were also identified in the ancestor and control passage.

807 **Table 2. Parallel or convergent evolution in niche-passaged pneumococcal lineages.** Table
808 shows all genes in which non-synonymous single nucleotide polymorphisms were identified in more
809 than one independently evolved lineage. Columns, from the left, show; 1. the gene affected and the
810 resulting amino acid change(s), 2. the corresponding gene in the D39 genome recently published by
811 Slager and colleagues [18], 3. the number of twenty-times nasopharynx passaged lineages in which the
812 variant is found at 100% frequency, 4. the number of twenty-times lung passaged lineages in which the
813 variants is found at 100% frequency, 5. whether the same variant was identified in either the ancestor
814 population or the laboratory passaged (control) lineages, and 6. The COG classification [21] of the
815 affected gene.
816

| Gene (amino acid change caused by mutation)	D39V annotation	In passage 20 nasopharynx lineages? (n/10)	In passage 20 lung lineages? (n/10)	In control passage lineages or ancestor?	Clusters of Orthologous Groups (COG) category
mutX1 (P29S)	*SPV_0644* MutT/Nudix family protein		2	N	Nucleotide transport and metabolism
P0_00979 (Y331C)	*SPV_0961 lafA*		3	N	Cell wall/membrane biogenesis
sdhB (G14R)	*SPV_0103 sdhB*		2	N	Amino acid transport and metabolism
arnB (S192F, A197G, A96V, P11S)	*SPV_1619 aatB*	3	2	Y (A96V)	Amino acid transport and metabolism
P0_01974 (A608S, L170M)	*SPV_1969* Glycosyl hydrolase-related protein	3	3	Y	Carbohydrate transport and metabolism
yknX (S331N) (L333V)	*SPV_0686* Periplasmic component of efflux system	2	1	N	Cell wall/membrane biogenesis
P0_01029 (T62I, Y19H)	*SPV_0912 pvaA*	2		N	Cell wall/membrane biogenesis
gpsA (G208R)	*SPV_1918 gpsA*	2		N	Lipid transport and metabolism

817
818
819

820
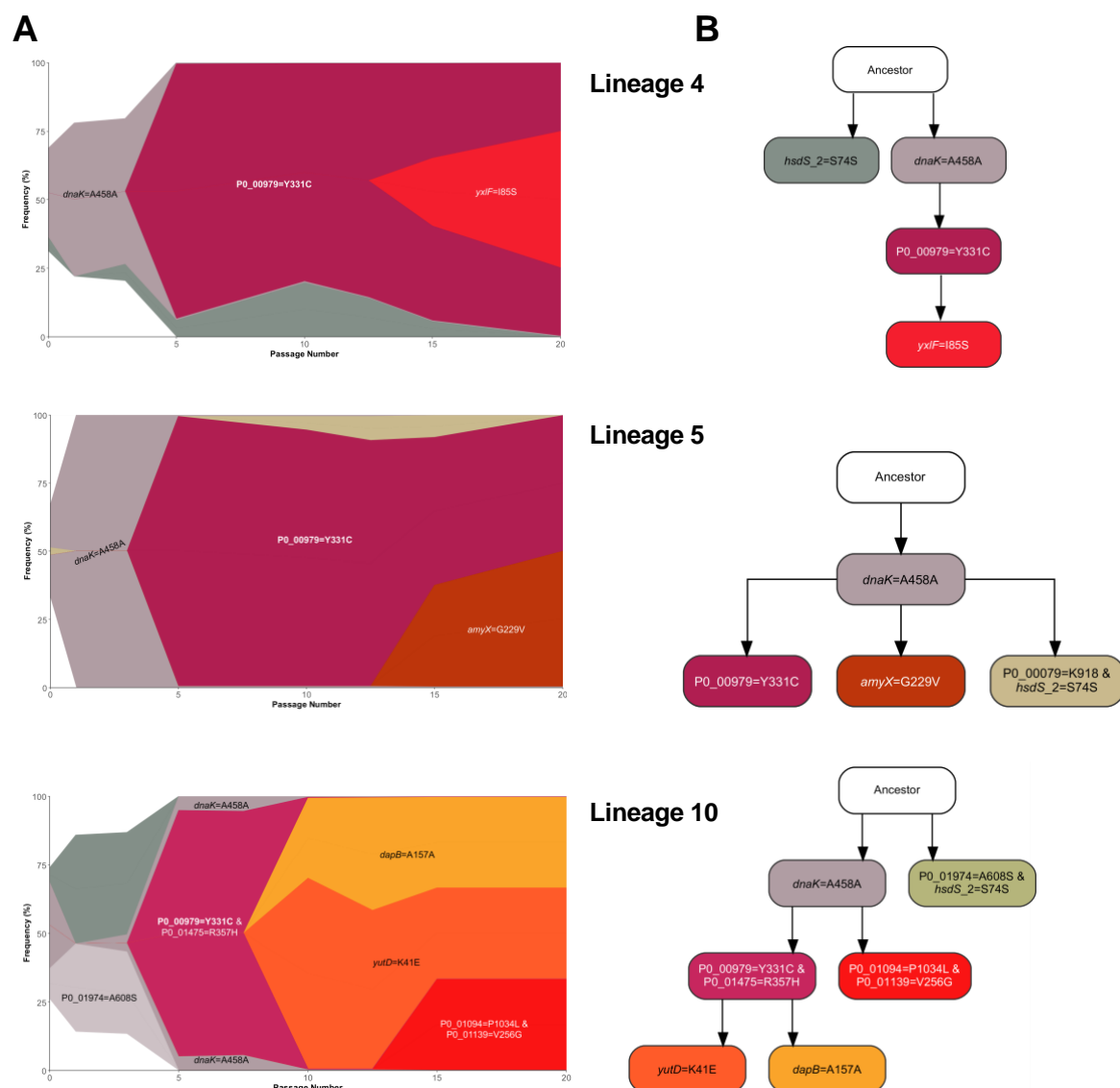
Figure 4. Success of *lafA* polymorphisms in lung-passaged pneumococcal lineages. (A) Muller plots showing genotype ancestry and frequency for lineages 4, 5 and 10 of lung-passaged pneumococci. The genotype associated with the *P0_00979* (*lafA*) SNP is shown in maroon. For lineages 4 and 5, this genotype is uniquely associated with this SNP, whilst in lineage 10 the genotype is also defined by a non-synonymous mutation in *P0_01475*. Additional colours are arbitrary but other successful genotypes containing non-synonymous SNPs are labelled. The ancestral genotype is in white. **(B)** Ancestry diagrams inferring relationships between genotypes within the population. Where multiple variants define a genotype, order of acquisition is uncertain.

831
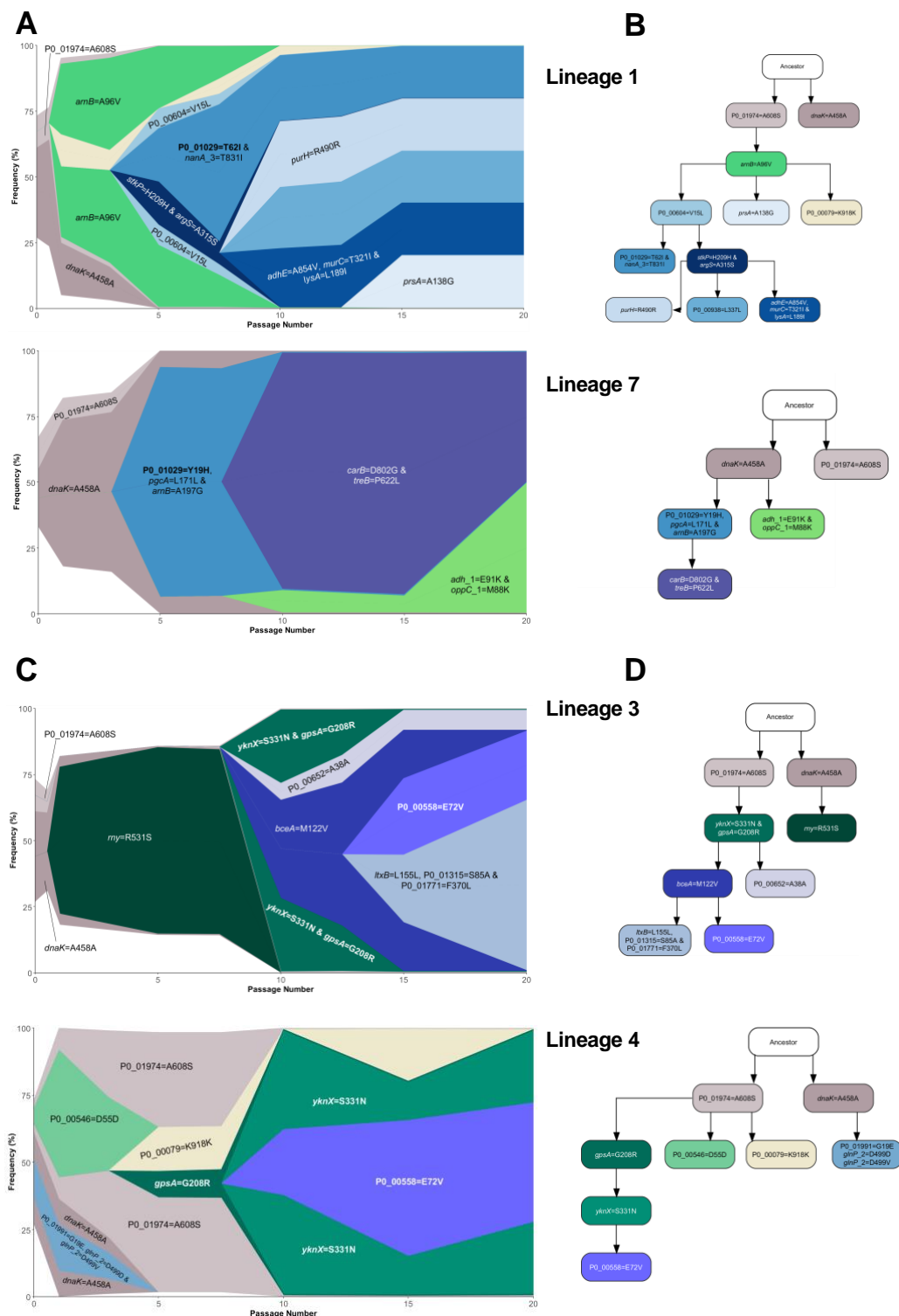
Figure 5. Parallel evolution in nasopharynx-passaged pneumococci. (A) Muller plots showing the emergence and success of *P0_01029* (*pvaA*) variants in nasopharynx-passaged lineages 1 and 7. The *pvaA*-associated genotype is shown in teal. Additional colours are arbitrary but other successful genotypes or those originating from the ancestor population are labelled. **(B)** Genotype ancestry for lineages 1 and 7. **(C)** Muller plots showing the emergence and success of *gpsA* and *yknX* variants in nasopharynx-passaged lineages 3 and 4. The *gpsA* and *yknX* genotypes are shown in shades of green. The two variants are associated with a single genotype in lineage 3. A shared non-synonymous variant in *P0_00558* is shown in purple. Additional colours are arbitrary. **(D)** Genotype ancestry for lineages 3 and 4.
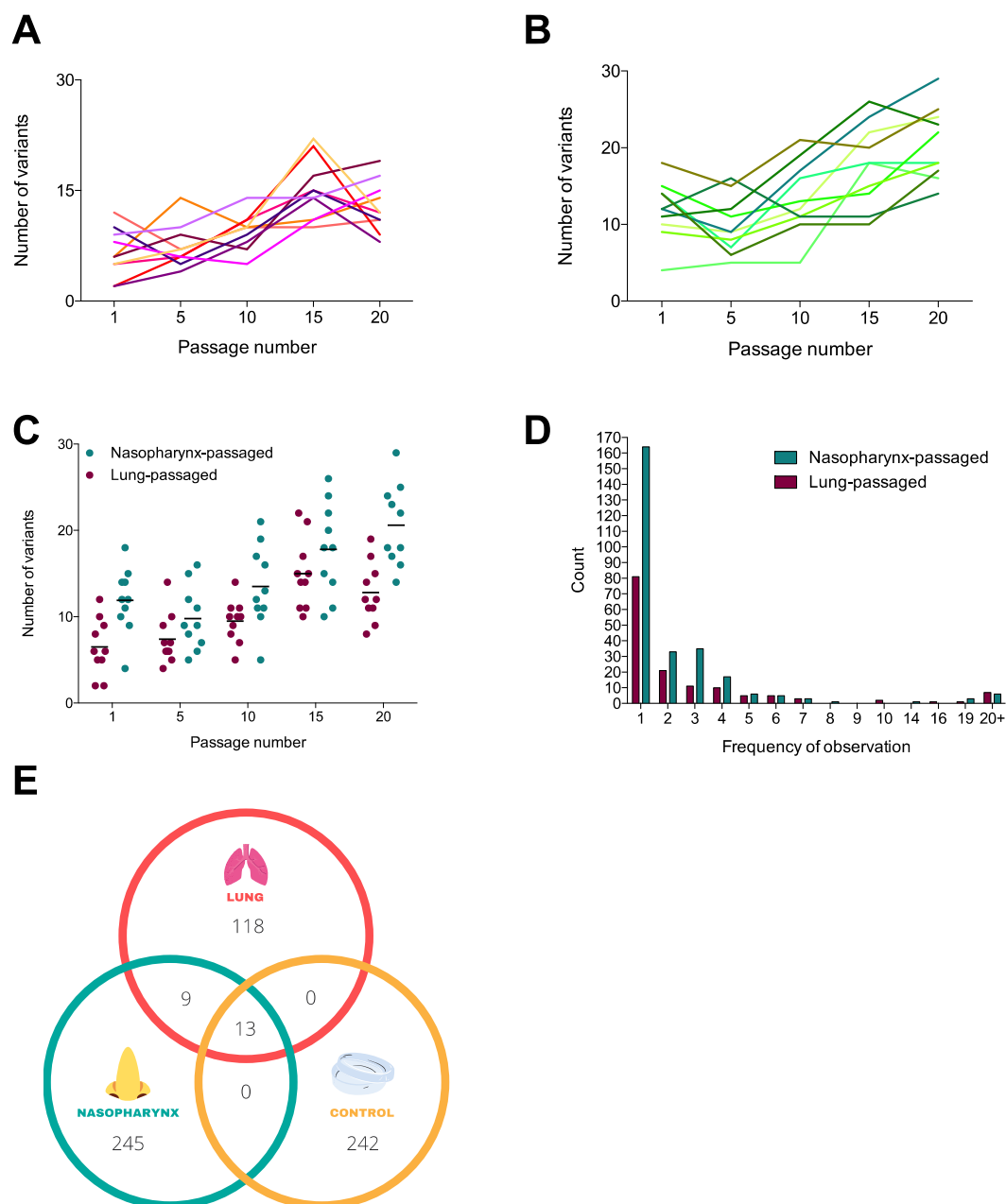
841
842 **Figure 6. Variant dynamics in niche-passaged pneumococci. (A)** Number of variants identified per
843 passage in ten *S. pneumoniae* lineages during passage through murine lungs in a pneumonia model.
844 Each line represents an independently evolving lineage (n=1 per lineage, per timepoint). **(B)** Variants
845 per passage for ten independently evolving nasopharynx-passaged lineages in a murine carriage model
846 (n=1 per lineage, per timepoint). **(C)** Comparison of variant numbers per passage in nasopharynx and
847 lung passaged pneumococci. Each dot represents data from a single lineage. Data are a composite of
848 those in **(A)** and **(B)** and are included for clarity. **(D)** Number of times each unique variant was detected.
849 Data are from a composite of all passages of all lineages. For example, if the same variant is detected
850 at passage 1 and 5 in lineage 2 and at passage 20 in lineage 3, but is not found in other lineages, it
851 would be recorded as a count of 3. **(E)** Venn diagram showing unique variants in nasopharynx-, lung-,
852 or lab (control) passaged lineages. Here, each variant is counted only once, regardless of the number
853 of times it was detected.
854
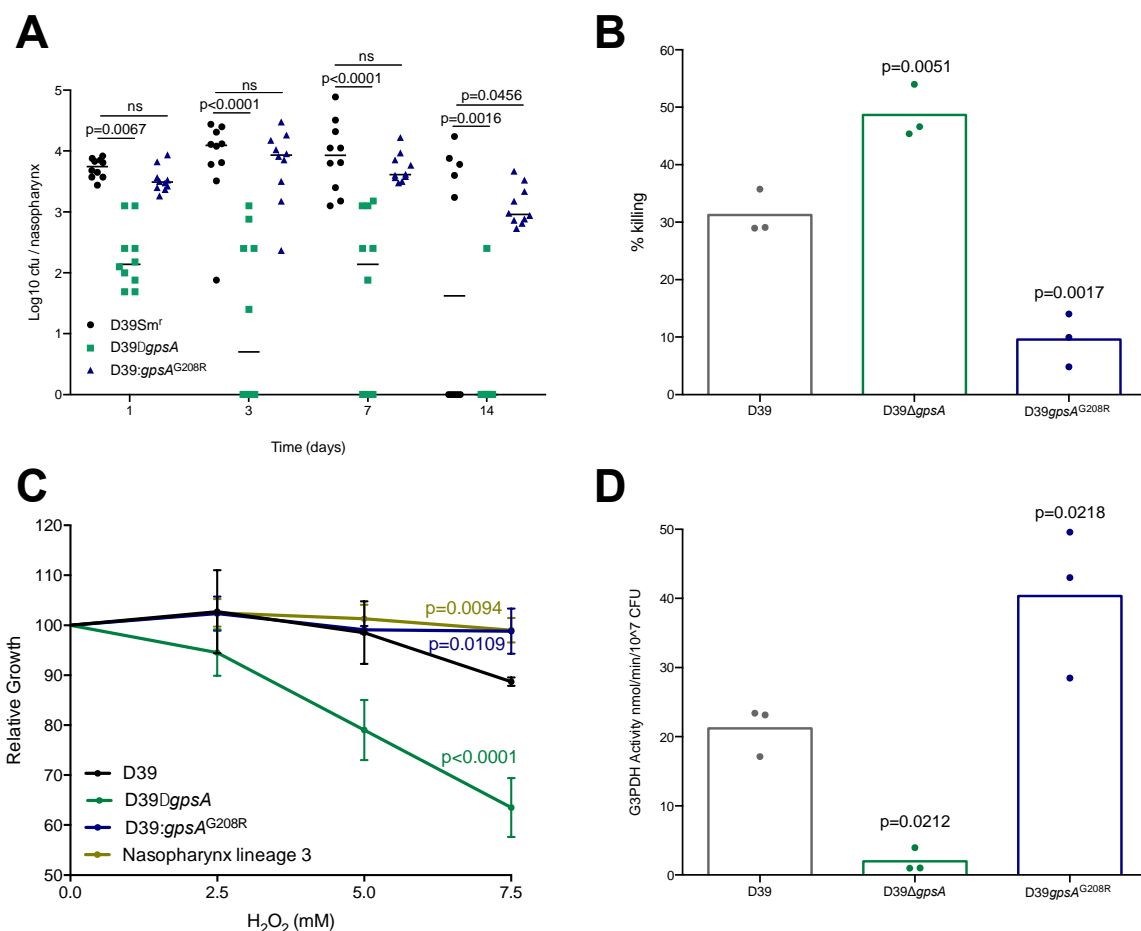
855
856 **Figure 7. Glycerol-6-phosphate dehydrogenase is required for effective colonisation of**
857 **nasopharynx. (A)** Mice were infected with a streptomycin resistant version of the ancestor D39 in which
858 either a clean deletion of *gpsA* (D39Δ*gpsA*) or the SNP causing the G208R substitution (D39:*gpsA*G208R)
859 had been generated using the sweet Janus system. Control animals were infected with the streptomycin
860 resistant D39 ancestor expressing wild type *gpsA* (D39Smr – streptomycin resistant). Infection dose
861 was 1×10^5 colony forming units (cfu) to induce asymptomatic carriage. Bacterial numbers were
862 determined by serial dilution of nasopharynx homogenates onto selective agar at days 1, 3, 7 or 14
863 post-infection. N = 10 per group. P-values are from two-way ANOVA with Sidak's multiple comparison
864 test. **(B)** Cfu counts for D39, D39Δ*gpsA* and D39*gpsA*G208R following 30 minutes incubation in a 12 mM
865 hydrogen peroxide (H_2O_2) solution. Data are presented as percentage killing vs the no H_2O_2 control for
866 each strain. P-values are from a one-way ANOVA with Dunnett's multiple comparison test vs D39. **(C)**
867 Percentage reduction in growth (measured as area under the logistic curve [AUC]) for D39**,** D39Δ*gpsA*,
868 D39*gpsA*G208R and nasopharynx lineage 3 in the presence of increasing concentrations of H_2O_2. P-
869 values are from a two-way ANOVA with Dunnett's multiple comparison test vs D39. Each point is the
870 mean of three biological replicates, each containing three technical replicates. Error bars are standard
871 deviation. **(D)** Glycerol-3-Phosphate dehydrogenase (G3PDH) enzyme activity measured in the
872 ancestor, D39Δ*gpsA* and D39*gpsA*G208R using a colorimetric assay. Activity was determined by
873 comparison of absorbance readings at 450 nm to a NADH standard curve and are presented as nmol
874 NADH per minute per 10^7 cfu. P-values are from a one-way ANOVA with Dunnett's multiple comparison
875 test vs the D39 ancestor. In **(B and D)** each data point represents a single biological replicate, each of
876 which is the mean of three **(B)** or two **(D)** technical replicates.
877