Pan-genome study underling the extent of genomic variation of invasive *Streptococcus pneumoniae* in Malawi

Arash Iranzadeh[1], Arghavan Alisoltani[2,3,4], Anmol M Kiran[5,6], Robert F Breiman[7], Chrispin Chaguza[8,9,10], Chikondi Peno[5,6,9], Jennifer E Cornick[5,11], Dean B Everett[12&], Nicola Mulder[1&]

[1] Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Western Cape, South Africa.

[2] Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA.

[3] Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA.

[4] Center for Pathogen Genomics and Microbial Evolution, Havey Institute for Global Health, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA.

[5] Malawi-Liverpool-Wellcome Trust, Queen Elizabeth Central Hospital, College of Medicine, Blantyre, Malawi.

[6] Centre for Inflammation Research, Queens Research Institute, University of Edinburgh, Edinburgh, United Kingdom.

[7] Rollins School of Public Health, Emory University, Atlanta, Georgia, United States of America.

[8] Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom.

[9] Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, Connecticut, USA.

[10] Yale Institute for Global Health, Yale University, New Haven, Connecticut, USA.

[11] Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, United Kingdom.

[12] Department of Pathology, College of Medicine and Health Sciences, Khalifa University, Abu Dhabi, UAE.

[&] Joint last authors

Corresponding author:

30 Nicola Mulder: nicola.mulder@uct.ac.za

31 Impact Statement

32 We applied pan-genomics principles to characterize diversity in the pneumococcus genome. The main
33 objective of the research was to highlight pneumococcal virulence genes, a crucial requirement for
34 vaccine design or drug development. We identified Serotypes 1 and 5 as the most abundant and
35 invasive pneumococcal strains in Malawi with a short period of nasopharyngeal colonization that could
36 imply their quickness in infecting the sterile sites in the human body, such as blood and the central
37 nervous system. They genetically diverged from other serotypes. Genes from an operon known as RD8a
38 that perform aerobic respiration and ATP synthesis did not exist in serotypes 1 and 5. In contrast, these
39 serotypes significantly harbored transporters that do not require ATP to function. The pattern of gene
40 presence-absence in serotypes 1 and 5 could, to some extent, explain why they are invasive with a short
41 colonization period in the upper respiratory tract.

42 Abstract

43 *Streptococcus pneumoniae* is a common cause of acute bacterial infections in Malawi. Understanding
44 the underlying molecular mechanisms of the pathogen's invasive behavior is crucial for designing new
45 therapeutic strategies. We conducted a pan-genome analysis to identify potential virulence genes in *S.*
46 *pneumoniae* by comparing the gene pool of the isolates from carriers' nasopharyngeal secretions to
47 isolates from the blood and cerebrospinal fluid of patients. We identified 56 serotypes in the cohort.
48 Most serotypes showed a similar prevalence in carriage and disease groups. However, serotypes 1, 5,
49 and 12F were disproportionately more commonly detected in specimens from patients compared to the
50 carriage group, as expected, given the known invasiveness of these serotypes. These serotypes were
51 genetically distinct from other isolates of other serotypes. The lack of genes from a genomic island
52 known as RD8a was the most pronounced difference between serotypes 1, 5, and 12F compared to
53 significantly prevalent serotypes in the nasopharynx. Genes in RD8a are involved in binding to epithelial

cells and performing aerobic respiration to synthesize ATP through oxidative phosphorylation. The absence of RD8a from serotypes 1, 5, and 12F may result in a short duration in the nasopharynx, theoretically because of decreased capacity to bind to epithelial cells and access free oxygen molecules for aerobic respiration (needed to maintain the carriage state). Available ATP may be lower in isolates without RD8a, like serotypes 1, 5, and 12F, causing them to alternatively use phosphotransferase systems to transport carbohydrates since these transporters use phosphoenolpyruvate as the energy source instead of ATP. In conclusion, serotypes 1, 5, and 12F, the most prevalent invasive pneumococcal strains in Malawi, showed a considerable genetic distinction from other strains that may explain their invasiveness.

Author summary

Despite introducing the pneumococcal conjugate vaccine in 2011, *Streptococcus pneumoniae* remains a major cause of bacterial infection in Malawi. Whilst some pneumococcal strains harmlessly colonize the nasopharynx, others find their way into normally sterile sites, such as lungs, blood, and nervous system, resulting in serious illness. We found specific pneumococcal strains the most invasive in Malawi, they had a short colonization period and were genetically distinct from other strains. Their remarkable genetic divergence included the absence of several genes involved in aerobic respiration and the presence of genes that contribute to ATP-independent carbohydrate transport. The presence or absence of these genes may be one of the underlying causes of their high invasiveness with a short colonization period. This hypothesis could make these genes potential candidate targets for future therapeutic strategies. We hypothesized that the specific gene gain and/or loss in invasive vs. other serotypes are possibly associated with invasive pneumococcal disease development.

77

78 ## Introduction

79 *Streptococcus pneumoniae,* also known as *pneumococcus*, is a Gram-positive, facultatively anaerobic
80 bacteria and is one of the leading causes of mortality worldwide. Despite reductions in the incidence of
81 pneumococcal disease in countries that introduced pneumococcal conjugate vaccines (PCV), the
82 pneumococcal mortality rate is still high. Pneumococci are estimated to be responsible for 317,300
83 deaths in children aged 1 to 59 months worldwide in 2015 [1]. In the post-PCV13 era, a high disease
84 burden has still been reported in low-income countries in Africa, such as Malawi [2].

85 Although pneumococcal nasopharyngeal colonization is asymptomatic, it is a prerequisite for
86 transmission and disease development [3][4]. Symptoms appear when isolates from the nasopharynx
87 spread to normally sterile sites such as the lung, blood, and central nervous system. Depending on the
88 infected organ, *S. pneumoniae* can cause two types of infection: (i) non-invasive (mucosal)
89 pneumococcal diseases such as otitis media and sinusitis, and (ii) invasive pneumococcal diseases (IPD)
90 such as bacteremia and meningitis. IPD incidence is highest among infants, the elderly, and
91 immunosuppressed people, most likely due to their less efficient immune systems [5].

92 Pneumococci possess several virulence factors, including the polysaccharide capsule, surface proteins,
93 and enzymes [6][7]. The polysaccharide capsule is the most important virulence factor as it aids the
94 *pneumococcus* in evading the immune response during colonization and invasion [8]. Its biosynthesis is
95 regulated by a cluster of genes in the capsular polysaccharide (*cps*) locus [8][9]. Pneumococcal serotypes
96 are defined by the type and order of monosaccharides that compose the capsule structure. To date, one
97 hundred pneumococcal serotypes have been identified [10]. Each strain has a set of capsular synthesis
98 genes in the *cps* locus that determine its serotype. Immunogenic properties of the capsular
99 polysaccharide were utilized to develop all pneumococcal vaccines in use, including PCV7, PCV10, and
100 PCV13 that cover 7, 10, and 13 serotypes, respectively. Although the introduction of PCVs has

101 significantly reduced the burden of disease caused by vaccine types (VTs), serotype replacement has
102 increased the non-vaccine types (NVTs) carriage rate and IPD incidence [11][12].

103 In November 2011, PCV13 was introduced in Malawi, which markedly decreased the health system
104 burden and rates of severe childhood pneumonia [13]. PCV13 includes serotypes 1, 3, 4, 5, 6A, 6B, 7F,
105 9V, 14, 18C, 19A, 19F, and 23F. A case-control study in Malawi showed vaccine effectiveness against VT-
106 IPD of 80·7% [14]. A nasopharyngeal carriage survey conducted in the Karonga district showed that
107 although the vaccine reduced the VT colonization rate, a moderate level of serotype replacement was
108 observed among carriers [15]. Moreover, the emergence of antibiotic-resistant pneumococci due to the
109 overuse of antibiotics is a global concern in the 21st century [16]. To develop new, more effective
110 vaccines against the vaccine-escape clones and design effective drugs against antibiotic-resistant strains,
111 it is critical to understand the functions of genes involved in pneumococcal colonization and
112 pathogenesis. During the past decade, the evolution of high-throughput sequencing technologies has
113 generated enormous amounts of genomic data that have enabled researchers to perform large-scale
114 genomic analysis. A well-known example is pan-genome studies. The pan-genome is the entire gene set
115 in a collection of closely related strains within a specie [17]. Determining the genetic drivers is an active
116 and promising area of research that can provide insights into pneumococcal disease prevention and
117 treatment to reduce mortality rates. The pan-genome is useful for analyzing recombinogenic pathogens
118 such as *pneumococcus*[18]. The recombination level is high in thirteen pneumococcal genomic loci
119 known as regions of diversity (RDs) numbered from RD1 to RD13, some of which are involved in
120 virulence [19][20].

121 In this study, we used whole-genome sequencing (WGS) of 1477 pneumococcal samples recovered from
122 people residing in Blantyre, Karonga, and Lilongwe in Malawi to address three complementary issues: (i)
123 identifying the serotype distribution, (ii) understanding the structure of the pneumococcal population,
124 and (iii) identifying potential invasive driver genes and their biological functions.

125 # Materials and methods

126 ## Study design and sample collection

127 The study was conducted using the samples archived by Malawi-Liverpool Wellcome Trust Clinical
128 Research Programme (MLW). The study cohort included isolates obtained from asymptomatic carriers
129 and symptomatic patients that reside in three regions of Malawi including Blantyre in the south,
130 Karonga in the north, and Lilongwe in the center.

131 Carriage samples were collected from the nasopharynx of healthy individuals living in the area covered
132 by the Karonga Health and Demographic Surveillance System (HDSS) [16] and those presenting at Queen
133 Elizabeth Central Hospital (QECH) in Blantyre [21]. Invasive pneumococcal samples were collected from
134 the blood and cerebrospinal fluid (CSF) of symptomatic patients visiting QECH and Kamuzu Central
135 Hospital in Lilongwe[22]. In this study, the term "sterile sites" refers to blood and CSF, and "invasive
136 samples" are those obtained from the sterile sites.

137 There were no paired samples. From each person, only one sample was collected, either a
138 nasopharyngeal swab from a healthy individual or a blood or CSF from symptomatic patients.
139 Pneumococcal isolates were identified using the procedure previously described [16]. Briefly, isolates
140 were identified on Blood Agar Gentamicin, based on colony morphology, alpha-hemolysis, and optochin
141 susceptibility as per standard pneumococcal isolates protocols.

142 ## Whole-genome sequencing and quality control

143 To account for the true diversity of carriages, only a single isolated colony was sequenced and
144 serotyped, therefore no carriage samples included in this study had multiple serotypes. Archived
145 samples were sequenced under the Global Pneumococcal Sequencing project and Pneumococcal African
146 Genomic Consortium at Wellcome Trust Sanger Institute in the United Kingdom. The bacterial DNA was
147 extracted using QIAamp DNA mini kit and QIAgen Biorobot manufactured by QIAGEN. The whole

148 genome sequencing was accomplished using the Illumina Genome Analyzer II and HiSeq platforms,
149 generating 125 nucleotides paired-end reads. The quality of the reads was evaluated by Fastqc [23].

150 **In-silico serotyping, sequence typing, and quantification of serotype invasiveness**

151 SeroBA version 1.23.4 was used to infer the serotype of samples [24]. SeroBA applies a k-mer method to
152 determine serotypes directly from the paired-end reads in FASTQ format. Any serotype with a relative
153 frequency greater than 5% was considered as an abundant serotype. Fisher's exact test was applied to
154 identify serotypes with a significant presence in the nasopharynx and sterile sites. P-values were
155 adjusted by the Benjamini-Hochberg method, and serotypes with an adjusted p-value less than 0.01
156 were considered significant. The odds ratio (OR) was calculated as follows OR = $(ad)/(bc)$, where for
157 serotype k, a is the number of invasive serotype k, b is the number of carriage serotype k, c is the
158 number of invasive non-serotype k, and d is the number of carriage non-serotype k. For OR calculation,
159 zero values were replaced by 0.5. Any abundant serotype with a significant presence in the nasopharynx
160 was considered to have low invasiveness, and high invasiveness was assigned to abundant serotypes
161 with a significant presence in the sterile sites. Fisher's exact test was also applied to identify serotypes
162 whose frequencies changed significantly after introducing PCV13 in 2011 (adjusted p-value < 0.01).

163 **Genome assembly and annotation**

164 Genomes were assembled using Velvet Optimiser version 2.2.5 [25]. The assembler was set to produce
165 contigs longer than 500 base pairs. The starting hash value was 61, and the end hash value was 119. The
166 quality of the assembled genomes was evaluated by Quast version 5.2[26]. The assembled genomes
167 were annotated using Prokka version 13.1 [27].

168 **Pan-genome construction**

169 The pan-genome for samples was generated using Roary version 3.12.0 [28]. Roary was run to perform
170 the core gene alignments with MAFFT version 7.313 [29]. Genes in the pan-genome were divided into

171 three groups based on their abundance among samples. Genes that existed in 100% of samples were
172 considered as core genes, those that were not core but present in more than 95% of samples were
173 regarded as a soft-core genes, and the rest were considered as accessory genes.

174 **Analysis of the population structure**

175 Since core genes are conserved between all samples, diversity in the core genome exists in the form of
176 small-scale variations such as single nucleotide polymorphisms (SNPs) and short indels. SNPs and indels
177 distribution in the core-genome was analyzed to gain insight into the population structure. The genetic
178 separation between samples was illustrated by the phylogenetic tree built from SNPs and indels in the
179 alignment of core genes as the phylogeny markers. The core genome alignment was used as input into
180 the IQ-TREE version 2 [30] to draw a phylogenetic tree visualized by iTol version 3 [31].

181 Diversity in the accessory genome exists in the form of large-scale gene presence-absence variations.
182 The R package Nonnegative Matrix Factorization [32] visualized the pan-genome matrix as a gene
183 presence-absence heatmap. Furthermore, to determine which factors, including isolation sites
184 (nasopharynx, blood, and CSF), serotypes, geographical locations, and vaccination eras, have the
185 greatest impact on the gene presence-absence profile of samples, a PCA of gene distribution was
186 accomplished by the R package MixOmics version 6.20.0 [33].

187 **Gene presence-absence statistical analysis**

188 The phenotypic trait of samples was assigned based on the population structure and isolates'
189 invasiveness. The gene presence-absence statistical test was performed using Scoary version 1.6.1643
190 [34]. This tool scores the components of the pan-genome for associations with observed phenotypic
191 traits while accounting for population stratification. The test was conducted between samples from
192 different sources and serotypes to find putative virulence factors. Every gene with a Bonferroni
193 corrected p-value less than 0.05 was reported as a significant gene.

194 **Functional and gene ontology (GO) enrichment analysis**

195 The list of significant genes was submitted to STRING webtool version 11.5 [35] for the functional
196 enrichment analysis. The STRING database is a network that integrates information from all publicly
197 available protein-protein interaction databases. The STRING's computational pipeline can predict direct
198 (physical) and indirect (functional) interactions between proteins. Interactions in STRING are derived
199 from 5 sources, including (i) genomic context predictions, (ii) high-throughput lab experiments, (iii) co-
200 expression, (iv) automated text mining, and (v) previous knowledge in databases. Functional enrichment
201 analysis in STRING uses information from classification systems such as the Kyoto Encyclopedia of Genes
202 and Genomes (KEGG) [36] and the Protein families database (Pfam) [37]. The tool entitled "Multiple
203 Sequences" was selected, and the *S. pneumoniae TIGR4* was chosen as the reference organism. STRING
204 reports the associated enriched pathways with a false discovery rate of less than 0.05.

205 Results

206 In total, 825 isolates from the nasopharynx of healthy carriers, 368 isolates from the blood of
207 bacteremia patients, and 284 isolates from the CSF of meningitis patients were sequenced. The
208 demographics of the samples are shown in Table 1.

Table 1. Demographics of 1477 pneumococcal isolates collected in Malawi.

Characteristics	Categories	Nasopharynx	Blood	CSF
	< 5	538	165	141
Age (in years)	5-19	109	42	60
	20-40	60	67	50
	> 40	7	24	7
	Unknown	111	70	26
	Female	401	131	111
Sex	Male	313	122	122

	Unknown	111	115	51
	Blantyre	169	357	259
	Karonga	656	0	0
City	Lilongwe	0	0	23
	Unknown	0	11	2
Sampling period		2009-2014	1997-2015	2000-2015

209

210 **Serotypes 1, 5, and 12F had the highest invasiveness, likely with a short period of**

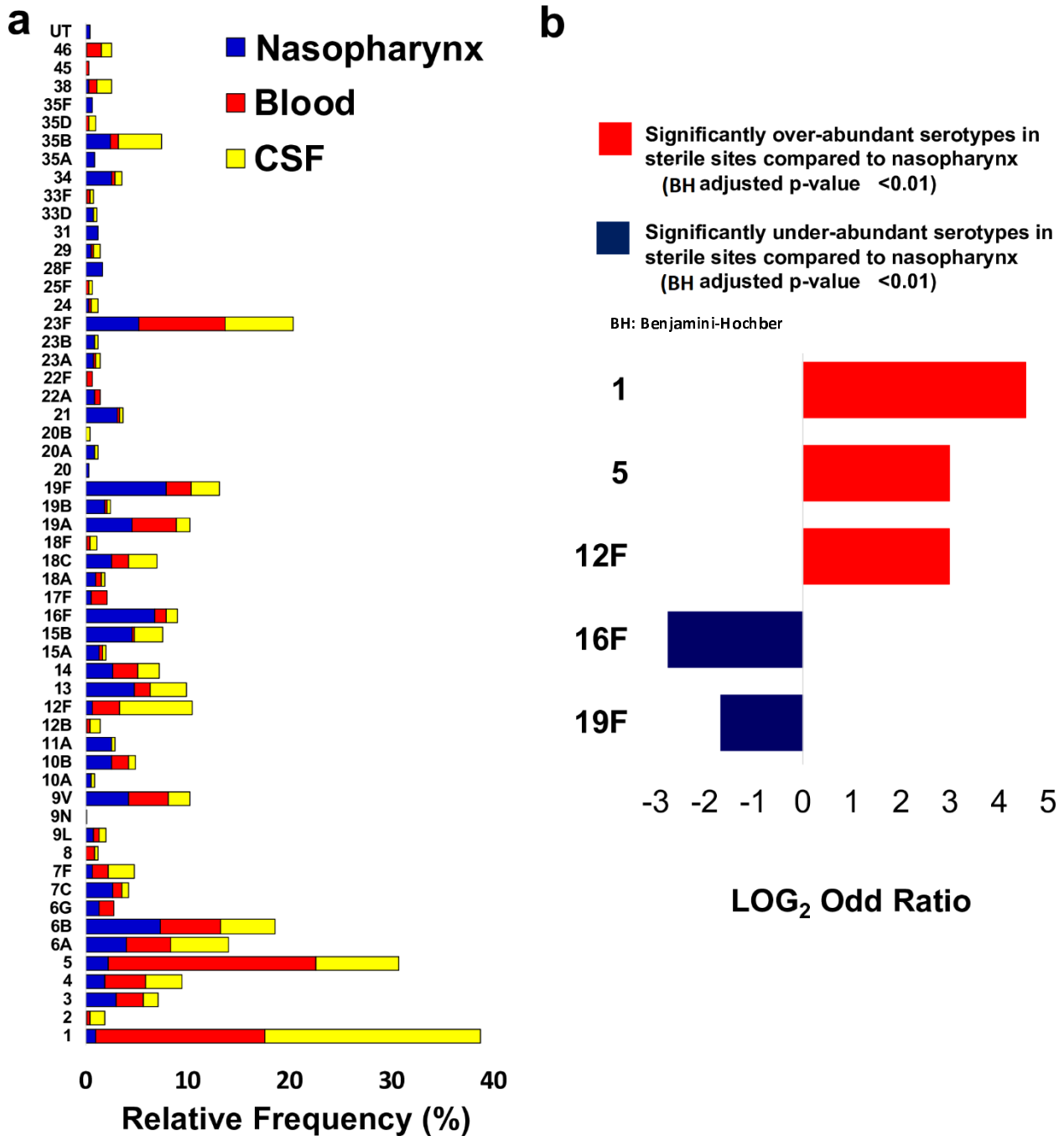
211 **nasopharyngeal colonization**

212 In total, isolates were assigned to 56 serotypes. Regardless of the isolation sites, serotypes 1 (8.7%), 5
213 (7.8%), 6B (6.6%), 23F (6.3%), and 19F (5.5%) were abundant, with a relative frequency greater than 5%
214 of the entire cohort. 66% of samples were obtained before the PCV13 rollout in 2011, and 27% were
215 from the post-PCV13 era (S1 Fig).

216 Amongst the carriage isolates, serotypes 19F (7.88%), 6B (7.27%), 16F (6.79%), and 23F (5.21%) were
217 abundant. Serotype 5 (20.38%) followed by 1 (16.58%), 23F (8.42%), and 6B (5.98%) dominated the
218 blood samples, while serotypes 1 (21.13%), 5 (8.1%), 12F (7.04%), 23F (6.69%), 6A (5.63%), and 6B
219 (5.28%) were predominant in the CSF. When aggregating blood and CSF groups, the most common
220 serotypes were 1 (18.56%), 5 (15.03%), 23F (7.67%), 6B (5.67%), and 6A (5%).

221 As illustrated in Fig 1 and S1 Table, of the abundant serotypes in the blood or CSF, serotypes 1 ($p =$
222 1.96E-34), 5 ($p = 3.97E-19$), and 12F ($p=5.29E-06$) had a significant presence among patients with a low
223 frequency in the nasopharynx including serotype 1 (1%), serotype 5 (2%) and serotype 12F (0.6%). Given
224 that the colonization phase is a pre-requisite for infection, serotypes 1, 5, and 12F may have a short
225 period of nasopharyngeal colonization before infecting the sterile sites. Since serotypes 1 and 5 were
226 also the most frequent in the entire cohort, they could be considered as the most prevalent serotypes

227 with the highest invasiveness. In this study, the terms “significant invasive serotypes” or "hyper-invasive
228 serotypes" were assigned to serotypes 1, 5, and 12F. Comparatively, abundant serotypes 16F and 19F in
229 the nasopharynx had a significant presence among carriers suggesting that they have a lower potential
230 to cause invasive disease. Other frequently detected serotypes, such as 6A, 6B, and 23F, were abundant
231 and equally distributed among carriers and patients. They may need to colonize the nasopharynx longer
232 than the hyper-invasive serotypes before infecting the sterile sites. Serotypes 6A, 6B, and 23F were
233 previously reported as abundant and common serotypes between non-invasive carriers and invasive
234 infection [38][39].



235

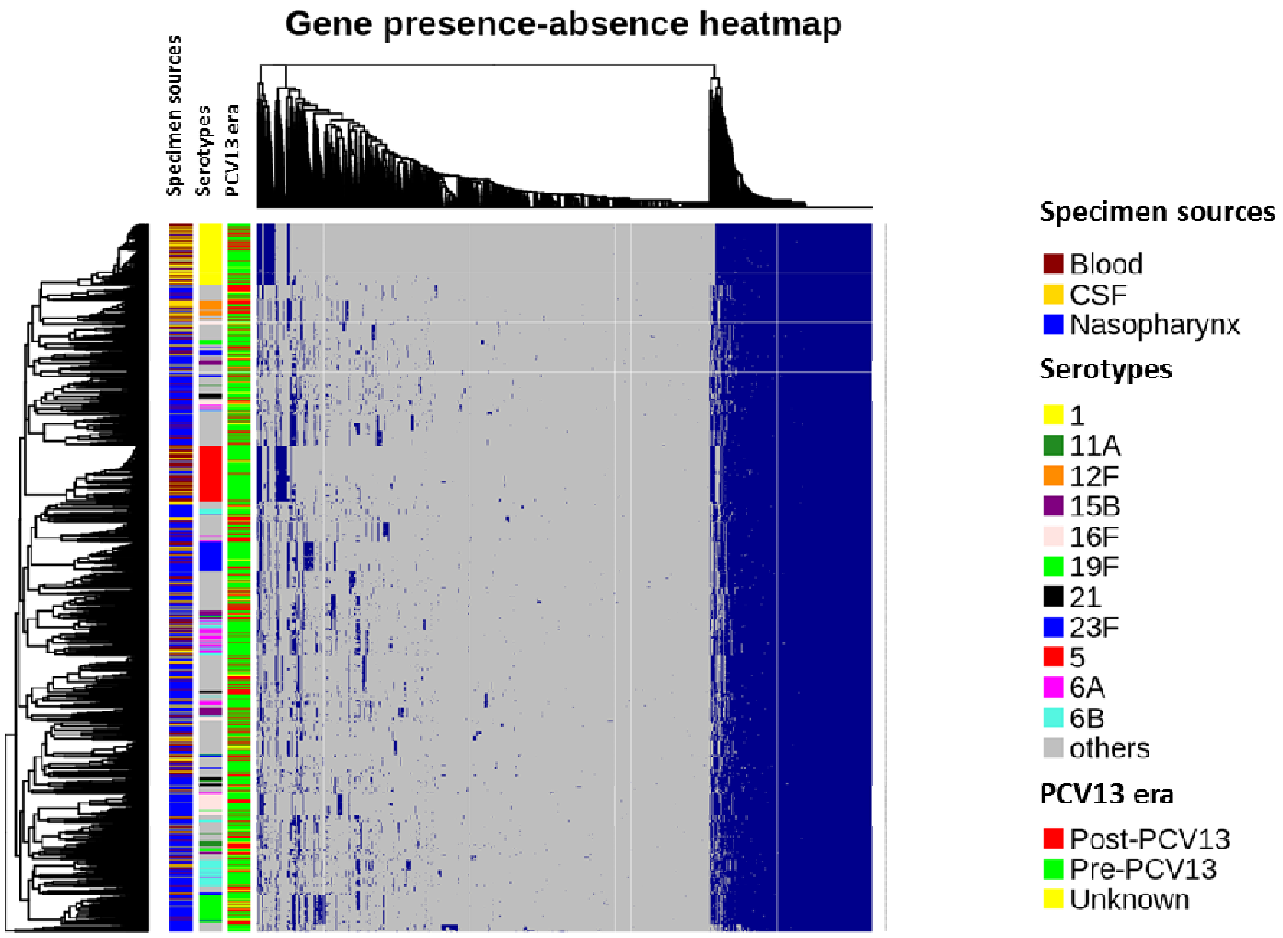
Fig 1. The distribution of the 56 pneumococcal serotypes assigned to 1477 samples from Malawi. (a) The relative frequency of each serotype in the nasopharynx of carriers, the blood of bacteremia patients, and the CSF of meningitis patients is shown in blue, red, and yellow, respectively (UT: Un-Typeable). (b) The log-transformed odds ratio of the significantly over- and under-abundant serotypes in the sterile sites (blood and CSF). Fisher's exact test was applied to identify serotypes with a significant differential abundance among carriers and patients (nasopharynx and sterile sites) at the significance level of the Benjamini-Hochberg adjusted p-value < 0.01 (BH: Benjamini-Hochberg).

236 The temporal distribution of the hyper-invasive serotypes with respect to the vaccine rollout time was of
237 note. The relative frequency of serotype 1 significantly increased (pre-PVC13=8.6% and post-
238 PCV13=11.3%), serotype 5 significantly decreased (pre-PVC13=11% and post-PCV13=1.5%), and serotype
239 12F significantly increased (pre-PVC13=1.1% and post-PCV13=5.7%) after the vaccination program in
240 Malawi. Vaccination might be effective against serotype 5 but did not reduce the burden of IPDs caused
241 by serotype 1. Serotype 12F, not included in PCV13, may be an invasive strain emerging in the post-
242 PCV13 era. However, a more extensive dataset with more recent samples is needed to characterize the
243 long-term PCV13 effects.

244 **High diversity in the pneumococcal pan-genome**

245 For the genome assembly, the average optimized assembly hash value and the average N50 were 96 and
246 113,986, respectively. The mean assembled genome size was estimated at 2,116,779 nucleotides with a
247 standard deviation of 106,481. This length is in the range of previously reported *S. pneumoniae* genome
248 size [40]. The length of the pan-genome was 5,178,167 base pairs, composed of 6,803 genes, including
249 729 core-(10.7%), 820 soft-core- (12.1%), and 5,254 accessory genes (77.2%). The pan-genome was open
250 as it did not reach sampling saturation as the number of genes increased as the sample size grew (S2
251 Fig). In Fig 2, the pan-genome is demonstrated as a gene presence-absence heatmap that shows
252 serotypes as the strongest determinant of the gene distribution in the pan-genome. The distinction was
253 evident for the hyper-invasive serotypes 1, 5, and 12F since they clustered in unique clades. The sets of
254 genes specifically present in different serotypes appear as small blue blocks in the heatmap.

255



256

Fig 2. The pan-genome matrix of 1477 pneumococcal isolates from Malawi. The pan-genome is visualized as a gene presence-absence heatmap representing the hierarchical unsupervised clustering of samples based on the distribution of genes in the pan-genome. Each row is a sample, and each column is a gene. A blue dot denotes the presence of each gene. On the right side of the heatmap, the large blue block shows core genes present in all samples. The left side of the heatmap represents the accessory genome along with the clustering bands. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with source-based p-value < 0.05, including 21, 11A, and 15B, are also highlighted on the heatmap.

257 **The significant invasive serotypes showed the highest distinction in the core- and accessory-**

258 **genome**

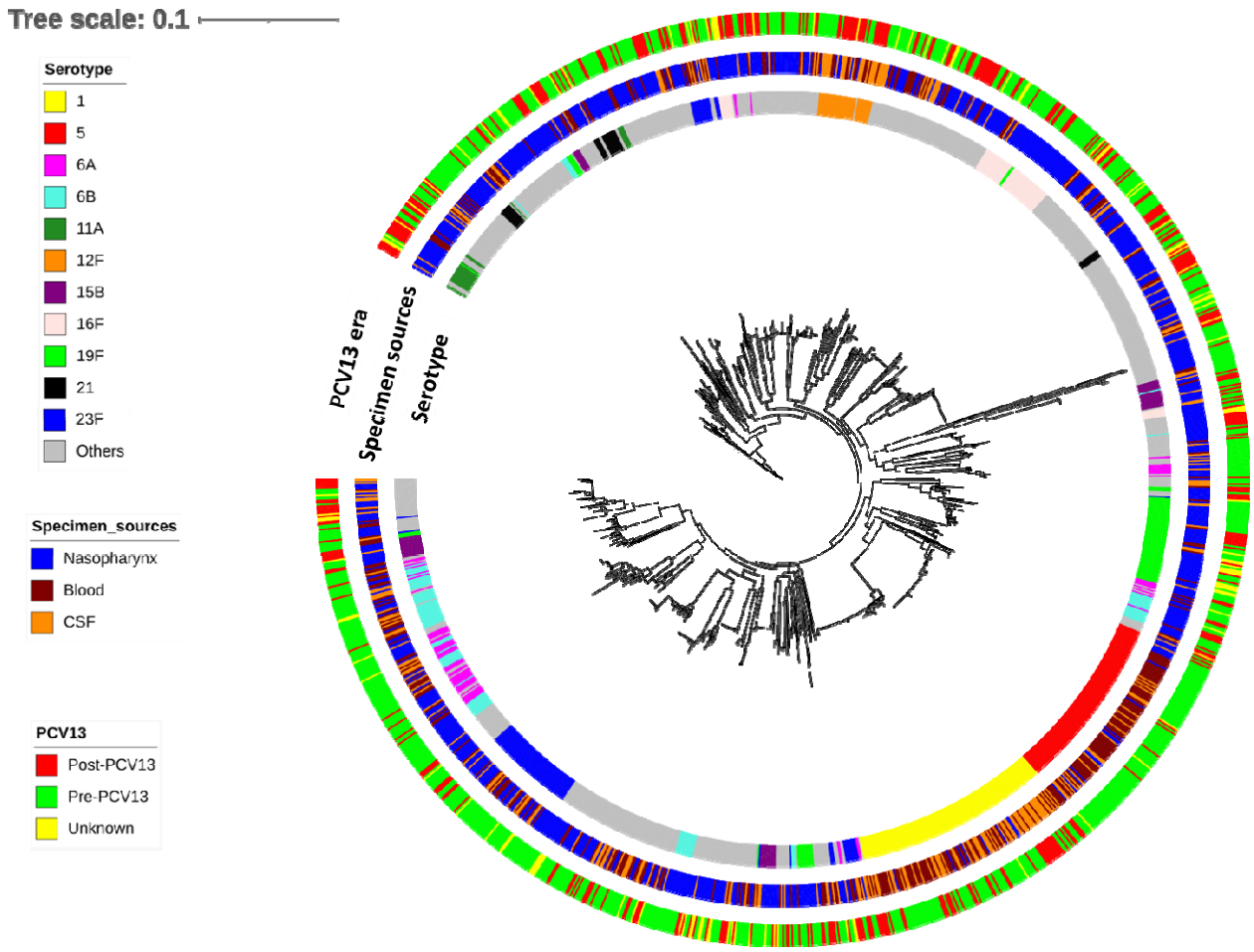
259 The maximum-likelihood tree characterized by the distribution of the small-scale variants (SNPs and

260 indels) in the core-genome identified the hyper-invasive serotypes 1, 5, and 12F as monophyletic

261 clusters (Fig 3). The distinction of hyper-invasive serotypes that were separated as unique clusters was

262 higher than other abundant serotypes, such as 6B, 19F, and 23F, which appeared as multiple clusters on

263 the phylogenetic tree (Fig 3).



264

Fig 3. The phylogenetic population structure of 1477 pneumococcal samples from Malawi. The phylogenetic tree was built based on the multiple sequence alignment of the core genome using the maximum likelihood method. Colors on the loops show the serotypes, specimen sources (isolation sites), and PCV13 eras. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with source-based p-value < 0.05, including 21, 11A, and 15B, are also highlighted on the tree.

265 The PCA of large-scale variants distribution in the accessory-genome (gene presence/absence) showed

266 serotypes 1 and 5 clustered distantly from other strains (Fig 4). Meanwhile, a moderate level of

267 separation was observed for serotypes 12F, 19F, and 23F. This highlighted the distinction of serotypes 1

268 and 5 in their gene content as observed in their SNPs profiles demonstrated by the phylogenetic tree.

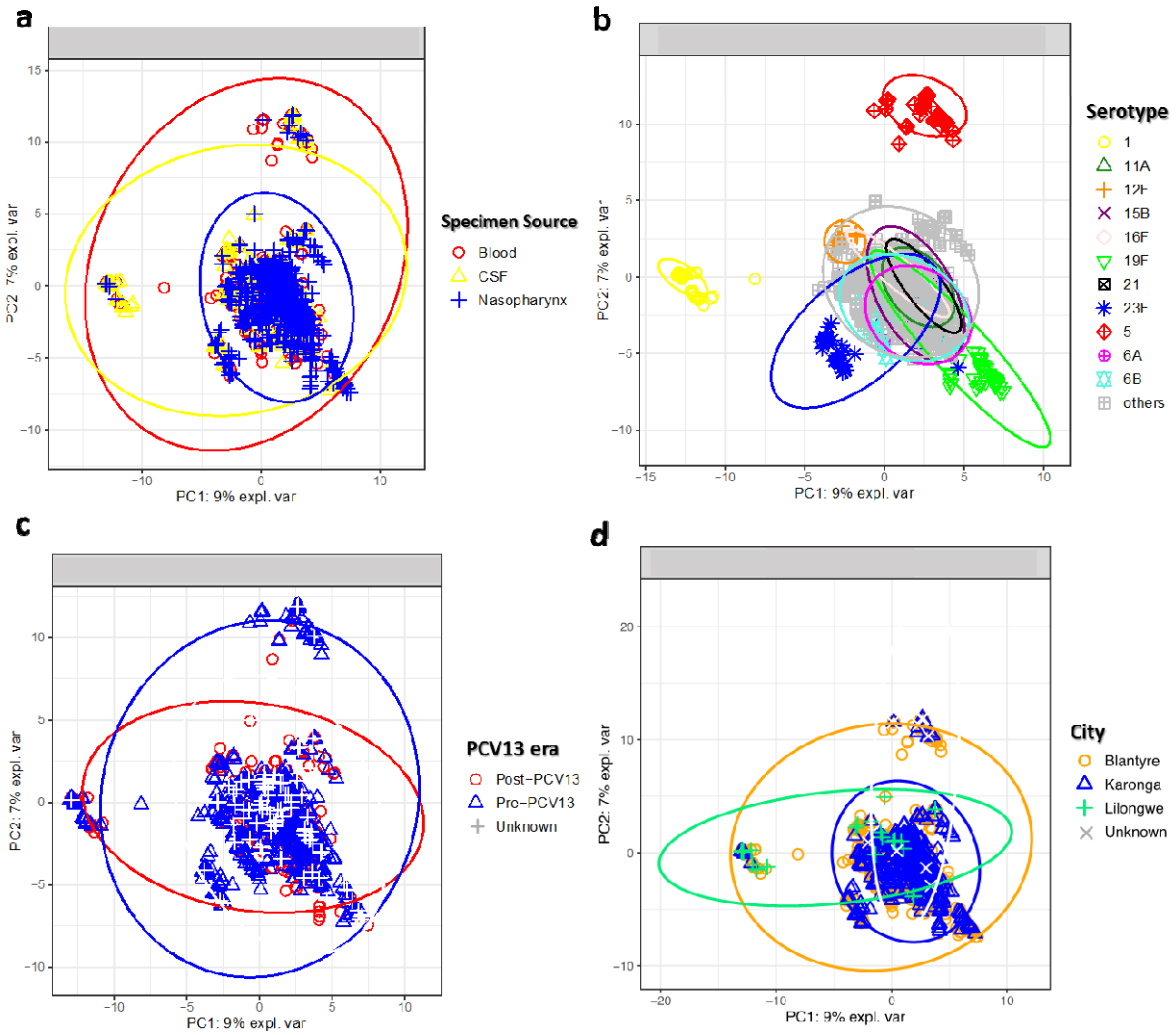


Fig 4. The PCA of the gene distribution in the pan-genome of pneumococcal isolates from 1477 Malawians. The PCA of variants (gene presence-absence) in the accessory-genome indicates the influence of (a) specimen sources (isolation sites), (b) serotypes, (c) PCV13 (vaccination) era, and (d) geographical locations on the gene presence-absence profile of pneumococcal isolates in Malawi. Serotypes 1 and 5 were clearly separated from other samples.

Fig 3 and Fig 4 show that factors such as time, locations, and isolation sites (specimen sources) could not explain the small- and large-scale genetic variants in the pan-genome. Indeed, the serotype of samples was the best driver of the population structure. Of importance was that the hyper-invasive serotypes had the maximum core and accessory distances from other strains, which means they are genetically different from other serotypes. Since these serotypes were associated with disease, their high genetic distinction may be associated with their invasive potential. Nonetheless, the analysis must rigorously

account for population structure and carefully curate any hits to show statistical association with the disease across the population. The first thing to point out is that the genetic distinction of serotype 1 and 5 is not observed in 12F. This may be due to the lower number of 12F (n=35) compared to 1 (n=129) and 5 (n=116). Moreover, the near-perfect separation between serotypes 1 and 5 and the other serotypes may be just because these strains were hardly found in the carriage group. Therefore, these patterns may primarily reflect biases in sampling rather than genuine genetic phenomena. To address these issues, ten samples of each hyper-invasive serotype and the vaccine types involved in PCV13 were randomly selected from the nasopharynx, blood, and CSF. The PCA of the gene distribution was repeated for the downsampled dataset and identified the stratification so that each of serotypes 1, 5, and 12F formed a distinct group clustered distantly from other strains (S3 Fig).

Typically, serotypes are separated from each other since they have different serotype-defining capsule genes. The issue that must be addressed is why hyper-invasive serotypes cluster more distantly in contrast with other serotypes. The hypothesis we explored is that serotypes 1, 5, and 12F gained or lost a list of genes that caused their quick invasiveness. There are other abundant serotypes in the blood and CSF (such as 6B and 23F), however, they have a similar prevalence in the nasopharynx, which means they may stay in the nasopharynx longer than the hyper-invasive serotypes.

The gene presence-absence statistical analysis

The following issues could skew the gene presence-absence analysis:

a) The batch effect introduced by geographical locations: 85% of carriage samples were from Karonga, whereas 95% of disease samples were from Blantyre (Table 1). A comparison between carriage and disease groups may only identify the difference between pneumococcal genomes from two geographical locations rather than between the non-invasive and invasive groups.
b) Study limitation: The likely presence of invasive serotypes in the carriage group made it unclear which nasopharyngeal samples progressed to disease after collection. Indeed, the carriage

300 population likely contained invasive serotypes that could bias the test between carriage and
301 disease samples to identify potential virulence genes.
302 c) Population structure: The significant abundance of the hyper-invasive serotypes in the patient
303 group and their highest genetic distinction would skew the test between the carriage and
304 patient groups. The difference between carriage and disease groups would actually be the
305 difference between the carriage and hyper-invasive serotypes (and not all strains in the blood
306 and CSF).

307 To measure the geographical batch effect, the carriage isolates from Karonga were compared to
308 carriage isolates from Blantyre. The test did not detect any significant gene in either Blantyre or Karonga
309 groups, implying that the gene content of the carriage samples from Blantyre and Karonga was similar.
310 The serotype distributions in the carriage group from Karonga and Blantyre were also similar. Therefore,
311 the effect of the geographical locations on the pneumococcal genomes was not remarkable. This was
312 expected as the distance between Karonga and Blantyre is about 830 kilometers, and the race of people
313 in these cities is similar.

314 Despite the possible presence of the invasive serotypes in the carriage group (e.g., 6A, 6B, and 23F that
315 were prevalent among the carriers and patients), comparing the whole carriage and patient groups
316 could still make sense since an invasive serotype in the nasopharynx may change its genome before
317 entering the sterile sites. Indeed, the potential invasive serotypes in the nasopharynx were isolated
318 during their colonization phase when they could exchange genes with other pneumococci or bacterial
319 species through recombination and horizontal transfer of genes. Thus, the genome of an invasive
320 serotype in the nasopharynx is not necessarily identical to its genome in the blood and CSF. The
321 association test was applied to compare the whole carriage and disease group, while the hyper-invasive
322 serotypes were excluded (location-based analysis). The exclusion of the hyper-invasive serotypes
323 prevented them from skewing the difference between the carriage and patient groups. The location-

324 based analysis identified 27 significant genes including 11 genes significantly present in the blood and

325 CSF and 16 genes in the nasopharynx (Table2, Table 3, and S2 Table).

Table 2. Significant genes (p-value < 0.05) in the blood and CSF (hyper-invasive serotypes were excluded).

ID	Annotation	P-value	Odds ratio
SP_0360	Capsular polysaccharide biosynthesis protein (from RD3)	8.17E-05	4.897059
SP_0358	Capsular polysaccharide biosynthesis protein (from RD3)	8.17E-05	4.897059
SP_0351	Capsular polysaccharide biosynthesis protein (from RD3)	8.17E-05	4.897059
SP_0359	Capsular polysaccharide biosynthesis protein (from RD3)	8.17E-05	4.897059
SP_0357	Capsular polysaccharide biosynthesis protein	0.000175	4.758421
SP_1037	Type II restriction endonuclease BcgI	0.001074	2.878307
SP_1953	Bacteriocin/lantibiotic secretion ABC transporter permease protein	0.001864	4.631892
SP_0535	Putative immunity protein	0.004674	1.988523
SP_1056	Relaxase/Mobilisation nuclease domain (From RD6)	0.012604	2.932153
SP_1656	Hypothetical protein	0.014050	1.963212
SP_0347	Capsular polysaccharide biosynthesis protein (from RD3)	0.020560	1.808997

326

327

328

329

330

331

332

333

334

335

Table 3. Significant genes (p-value < 0.05) in the nasopharynx (hyper-invasive serotypes were excluded).

ID	Annotation	P-value	Odds ratio
SP_1770	Glycosyl transferase, glyB (from RD10)	0.000132	0.502513
SP_1771	Glycosyl transferase, family 2/family 8 (from RD10)	0.012770	0.560236
SP_1763	Preprotein translocase secY family protein (from RD10)	0.017254	0.563256
SP_1765	Glycosyl transferase, glyF (from RD10)	0.018150	0.564542
SP_0939	Hypothetical protein	0.020740	0.483266
SP_1766	Glycosyl transferase, glyE (from RD10)	0.020740	0.483266
SP_1767	Glycosyl transferase, glyD (from RD10)	0.023196	0.568879
SP_1762	Accessory secretory protein asp1 (from RD10)	0.023196	0.568879
SP_1761	Accessory secretory protein asp2 (from RD10)	0.023196	0.568879
SP_1760	Accessory secretory protein asp3 (from RD10)	0.023196	0.568879
SP_1755	Hypothetical protein	0.031054	0.558554
SP_1757	Glycosyl transferase, glyB (from RD10)	0.031054	0.558554
SP_1764	Glycosyl transferase, glyG (from RD10)	0.023196	0.568879
SP_1768	Conserved hypothetical protein (from RD10)	0.031054	0.558553
SP_1758	Poly(glycerol-phosphate) alpha-glucosyltransferase, tagE (from RD10)	0.031122	0.571941
SP_1759	Preprotein translocase, secA subunit (from RD10)	0.031321	0.574557

The most significant genes in the blood and CSF were from *cps* locus (RD3) may be due to a higher level of encapsulation during the disease. Other significant genes, such as SP_1953, SP_0535, SP_1037, and SP_1056, contribute to toxic secretion and recombination. SP_1056 is from a region known as pneumococcal pathogenicity island 1 (PPI1), located in RD6. This gene encodes a mobilization protein required for the horizontal transfer of genes and plasmids through bacterial conjugation. SP_1056 forms the relaxation complex or relaxosome with the help of other enzymes [41].

343 Significant genes in the nasopharynx (absent from samples in the blood and CSF) were from RD10, which
344 is known as SecY2A2 island and is responsible for the secretion of pneumococcal serine-rich repeat
345 protein or PsrP (S4 Fig)[42]. RD10 contains several glycosyltransferases and secretory components that
346 were significantly absent from the genome of samples collected from blood and CSF.

347 To explore the divergence of the hyper-invasive serotypes, they were compared to serotypes 16F and
348 19F, which were significantly present in the nasopharynx (serotype-based analysis). Serotypes 16F and
349 19F could represent non-invasive strains better than the whole nasopharyngeal population that
350 potentially contained some invasive serotypes. Indeed, it was a test between serotypes with the highest
351 and lowest invasiveness to characterize the genomes of serotypes 1, 5, and 12F. The gene gain and loss
352 profiles in the hyper-invasive serotypes may include components contributing to their virulence and
353 short colonization period. The serotype-based analysis identified 184, 157, and 186 significant genes in
354 serotypes 1, 5, and 12F, respectively (S3 Table, S4 Table, and S5 Table) that were much larger than the
355 number of significant genes identified by the location-based analysis. The functional enrichment analysis
356 identified the phosphotransferase system (PTS, KEGG ID: spn02060) as over-represented and oxidative
357 phosphorylation (KEGG ID: spn00190) as under-represented pathways in the hyper-invasive serotypes
358 (p-value < 0.05).

359 In total, there were 18 significant genes jointly present in the hyper-invasive serotypes (Fig 5.a),
360 including elements of the PTSs that transport sucrose and lactose across the membrane (SP_0302,
361 SP_0303, SP_0304, SP_0305, SP_0306, SP_0308, SP_0309, and SP_0310), bacteriocins (SP_0544 and
362 SP_1051) and a permease protein (SP_1527). The over-represented pathway (spn02060) was associated
363 with significant genes that code for PTS transporters involved in carbohydrate metabolism. Seven genes
364 were unannotated. The PTS transporters genes were also present in a high proportion of abundant
365 serotypes in sterile sites such as 6B (67%) and 23F (86%).

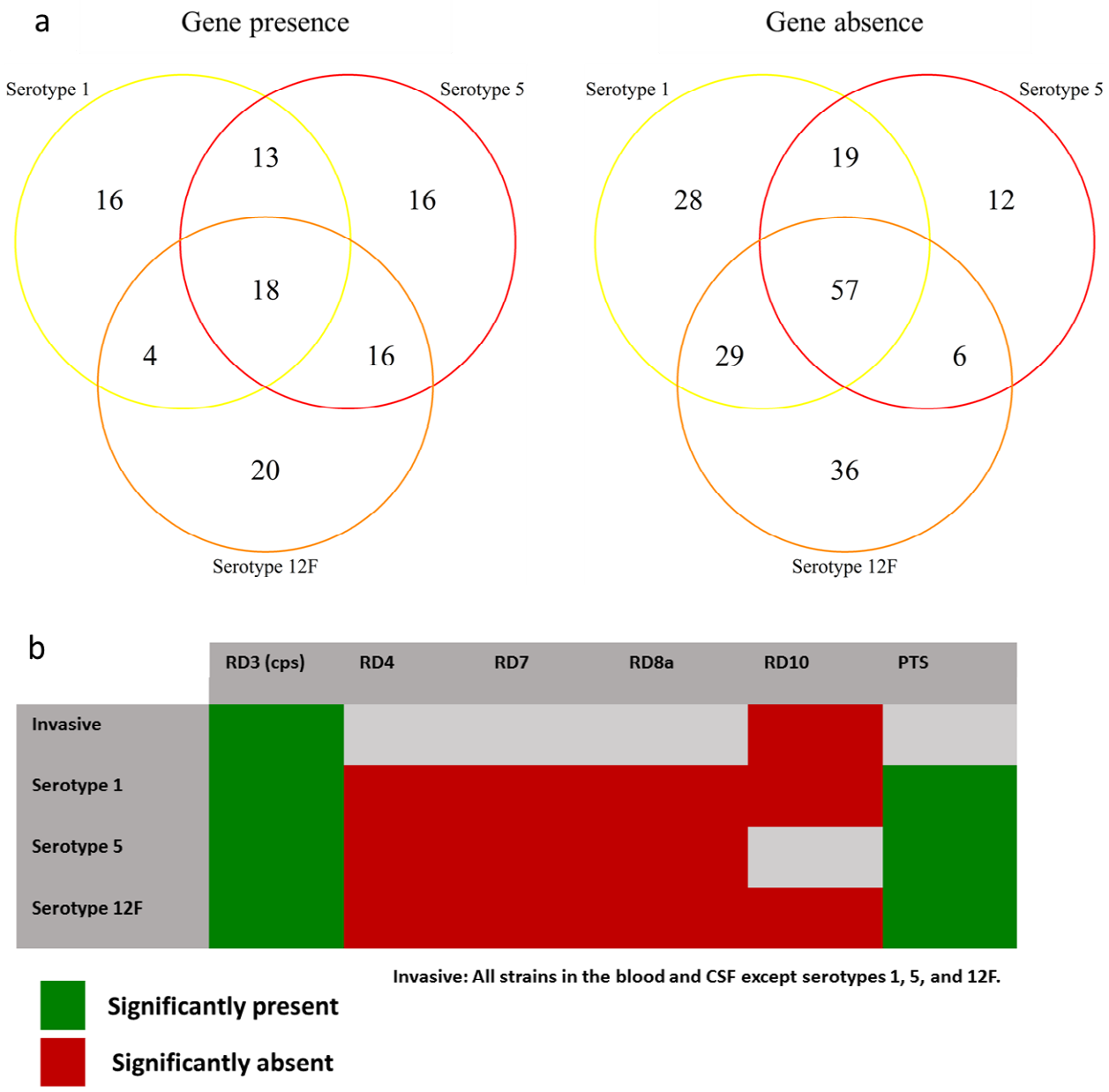
366 There were 57 significantly absent genes from serotypes 1, 5, and 12F (Fig 5.a). The most significant
367 ones were from RD8a, a region composed of two operons, RD8a1 (SP_1315-1324) and RD8a2 (SP_1325-
368 SP_1331) (S5 Fig). Genes in RD8a1 include eight *ntp* genes that code for the multiunit *V-type*
369 *proton/sodium ATP synthase* complex that produces ATP through oxidative phosphorylation in the
370 presence of a Na^+ gradient across the membrane [43]. RD8a2 includes *neuraminidase*, *N-*
371 *acetylneuraminate lyase (nanA),* and *N-acetylmannosamine-6-phosphate epimerase (nanE).* These genes
372 cleave carbohydrates from the glycoproteins on the surface of epithelial cells. Other genes in RD8a2
373 encode the Sodium/solute symporter subunits that use Na^+ gradient to import the carbohydrates [44].
374 Symporter means the channel transports the solute (carbohydrates) and co-solute (Na^+) in the same
375 direction using the energy stored in an inwardly directed sodium gradient. In summary, genes in RD8a
376 operons work together to produce ATP, cleave carbohydrates from the host epithelial cells, and import
377 them into the bacterial cell. RD8a was not present in any samples assigned to serotypes 1, 5, and 12F,
378 whereas it was present in other abundant serotypes such as 6A, 6B, 16F, and 19F. The under-
379 represented pathway (sp00190) in the hyper-invasive serotypes was linked to genes in RD8a. The lack of
380 RD8a in the hyper-invasive serotypes may relate to their speed at entering the blood and CSF, where
381 they do not have access to free oxygen molecules that are required for oxidative phosphorylation [45].

382 Other significant genes absent from the hyper-invasive serotypes were from RD4 and RD7. RD4 is
383 composed of a group of sortase enzymes that catalyze the assembly of pilins into pili and the anchoring
384 of pili and other surface proteins to the cell wall [46][47]. The pilus is a hair-like structure associated
385 with bacterial adhesion and colonization [48]. Due to the possible short colonization period of the
386 hyper-invasive serotypes, they may not benefit from genes in RD4 that assemble the pilus. Genes in RD7
387 have remained uncharacterized to date.

388 There were similarities between hyper-invasive serotypes and other strains in the sterile sites. As
389 observed for samples in the blood and CSF, capsule genes were significantly present in the hyper-
390 invasive serotypes, and RD10 was absent from serotypes 1 and 12F, however, this region was fully

391 conserved in serotype 5. RD10 was also conserved in 100% of serotypes 16F and 19F. The summary of

392 the gene present-absent analysis is illustrated in Fig 5.b.



393

Fig 5. The summary of the gene presence-absence analysis. (a) The number of significant genes present-absent in the hyper-invasive serotypes. The gene presence-absence analysis was applied using Scoary to compare the gene pools of the hyper-invasive serotypes and serotypes 16F and 19F. P-values were corrected by the Bonferroni method, and significant genes had an adjusted p-value of less than 0.05. (b) The significant presence and absence of RDs in samples from blood and CSF is shown as a presence-absence heatmap.

394 As mentioned, serotypes 6B and 23F were abundant in both carrier and patient groups, the test
395 between serotype 6B from the nasopharynx (n=60) and sterile sites (n=37) and serotype 23F from the
396 nasopharynx (n=43) and sterile sites (n=50) did not identify any significant gene. The gene content of
397 these two serotypes in the nasopharynx and sterile sites was similar.

398 To highlight genes that might assist pneumococci in crossing the blood-brain barrier, a test was
399 conducted between the samples from blood and CSF. The analysis between whole blood (n=368) and
400 CSF (n=284), serotype 1 samples from blood (n=60) and CSF (n=61), and serotype 5 samples from blood
401 (n=75) and CSF (n=23) did not identify any significant genes.

402 Discussion

403 The *S. pneumoniae* genome is highly diverse, and only a small fraction of genes is conserved between all
404 strains. In this species, the pan-genome is open and provides an unlimited gene repertoire for the highly
405 recombinogenic pneumococci. Changing the *S. pneumoniae* habitat could promote using different
406 combinations of genes by the organisms to diversify their genome and efficiently respond to
407 environmental stresses. In this study, the genetic diversity was identified to be very high, with only
408 10.7% of genes being core. These genes have been conserved in all samples over a long time, at least
409 from 1997 to 2015, and their presence may be vital for cell survival, making them potential targets for
410 drug design and vaccine development. In particular, core genes with no SNP in their structure would of
411 interest. Examples of the conserved core genes identified in this study included SP_1961 (*rpoB*, DNA-
412 dependent RNA polymerase), SP_0251 (formate acetyltransferase), SP_1891 (*amiA*, Oligopeptide
413 binding protein), and SP_0855 (*parC,* Topoisomerase IV). The conserved core genes are involved in DNA
414 transcription and translation.

415 Serotypes 1, 5, and 12F were significantly prevalent among patients and scarce in the carrier group. This
416 characteristic was interpreted as their high invasiveness likely with a short period of nasopharyngeal
417 colonization. On the other hand, serotypes 16F and 19F were significantly found among carriers,

418 implying that they dominate the nasopharynx with low invasiveness. Most of the other serotypes were
419 common between carriers and patients. Knowing that pneumococcal virulence strongly depends on the
420 serotype of isolates, we sought to address why several serotypes were shared across both nasopharynx
421 and sterile sites. Here, we discuss two possible scenarios that could justify the ubiquitous presence of
422 some serotypes in both nasopharynx and sterile sites.

423 The first scenario is related to the colonization of *S. pneumoniae*, which is known as a prerequisite for
424 virulence[4]. Several samples from the carriage group may be actually the invasive serotypes collected
425 during their colonization phase. Abundant serotypes such as 6B and 23F that had a similar frequency
426 amongst carriers and patients need to colonize the upper respiratory tract longer than the hyper-
427 invasive serotypes before entering the sterile organs. In contrast, the hyper-invasive serotypes 1, 5, and
428 12F colonize the nasopharynx for a short period and quickly enter the sterile sites.

429 The second possible scenario relates to the differential gene expression pattern of shared genes in the
430 ubiquitous serotypes. Although the type-specific *cps* genes were identified in the isolates of both
431 nasopharynx and sterile sites, the expression pattern of these genes could vary within each strain, which
432 would contribute to the invasiveness of some strains. Several studies described a cycle of encapsulation
433 and un-encapsulation amongst pneumococcal strains. Isolates benefit from mutations in the *cps* locus to
434 either cease or re-start the capsule expression [49]. The lack of a capsule at the epithelial surface
435 enables the bacterium to expose its surface proteins on the cell wall underneath the capsule and
436 promote adherence to the host epithelial cells. It has been estimated that 15% of isolates in the upper
437 respiratory tract are unencapsulated and adhere to the respiratory epithelial cells more efficiently than
438 encapsulated isolates [50][51]. Lack of capsule also facilitates acquiring virulence and resistance genes
439 from other isolates. The thick capsule prevents immunoglobulins from interacting with the pathogen
440 surface proteins during disease. Meanwhile, the negatively charged CPS interferes with the function of
441 the host phagocytes [52][53]. Taken together, the presence of the *cps* locus in the genome of isolates
442 assigned to the same serotype does not necessarily reflect the encapsulation of all cells.

443 Serotypes 1 and 5 can infect all age groups and cause severe IPDs [54]. Serotype 1 is genetically distinct
444 between different geographical regions [55] and is known as the leading cause of pneumococcal
445 meningitis in Africa [56][57]. Our findings support previous research showing that serotypes 1, 5, and
446 12F are the major cause of IPDs in Malawi [58]. Serotype 1 was persistently dominant in pre- and post-
447 PCV13 eras, serotype 5 was only predominant in pre-PCV3, and serotype 12F emerged after vaccination.
448 The study also characterized the high genetic distinction of the hyper-invasive serotypes in Malawi by
449 identifying significant genes that are present or absent in their genome structure compared to other
450 serotypes. Many of the significant genes present in the nasopharynx and sterile sites were homologous
451 and had the same function, and there were significant genes with an unknown function (hypothetical
452 proteins). However, of greatest interest, we did find several significant genes with specific functions that
453 could explain the difference between the biology of the hyper-invasive serotypes and nasopharyngeal
454 samples.

455 Genes from RD8a were not present in the genome of serotypes 1, 5, and 12F. The presence of this
456 region has been reported to be associated with the virulence of serotypes 6B and 14 in the United States
457 [59]. We found this region conserved in serotypes 13, 14, 16F, and 19F in Malawi. One may hypothesize
458 that RD8a may be required by serotypes that colonize the nasopharynx longer than serotypes 1, 5, and
459 12F. The functions of genes in RD8a strengthen the assumption to some extent that RD8a may be
460 essential for long nasopharyngeal colonization. RD8a harbors genes such as *neuraminidase*, *nanA,* and
461 *nanE* that cleave terminal sialic acid residues from mucoglycans and epithelial glycoconjugates and assist
462 the pathogen in penetrating the mucus layer and adhering to the epithelial cells in the nasopharynx. In
463 the meantime, because free carbohydrates are scarce in the upper respiratory tract [60], cleaved sialic
464 acid can be used as the carbon source for metabolism. During colonization, secretion of the
465 pneumococcal toxins elevates the level of sodium ions (Na^+) in the nasopharynx [61], the sodium-solute
466 symporter in RD8a imports a wide variety of substrates with the sodium ions into the cell [62].

467 Most importantly, the *ntp* gene cluster in RD8a encodes the *V-type sodium ATP synthase* that pumps the
468 extra sodium ions out of the cell [43] and uses the sodium-motive force for oxidative phosphorylation
469 and ATP synthesis [63]. Oxidative phosphorylation is the final step of aerobic respiration that requires
470 free oxygen molecules for ATP synthesis. Pneumococci are facultative anaerobes that can either
471 perform aerobic or anaerobic respiration with or without oxygen. In the upper respiratory tract, they
472 access atmospheric oxygen molecules that can be used by *ntp* genes to perform oxidative
473 phosphorylation. However, genes in RD8a may not be beneficial for hyper-invasive serotypes that
474 supposedly do not stay in the nasopharynx for long. Thus, the level of aerobic ATP synthesis is
475 presumably higher in serotypes 13, 14, 16F, and 19F in contrast with serotypes 1, 5, and 12F, which lack
476 RD8a.

477 Pneumococci can ferment up to 30 types of carbohydrates, imported mainly by two types of membrane
478 transporters, including ATP-binding cassette (ABC) transporters and PTS transporters [64]. The major
479 differences between ABC and PTS transporters are: (i) ABC transporters use energy from ATP, but PTS
480 transporters use energy from phosphoenolpyruvate, and (ii) ABC transporters do not modify the
481 imported substrate, but PTS transporters phosphorylate the incoming sugar upon transport. Generally,
482 ABC transporters require more energy than PTS transporters, albeit they can transport longer and more
483 complicated carbohydrates [65]. Unlike isolates in the nasopharynx, serotypes 1, 5, and 12F have access
484 to more simple and free host dietary carbohydrates in the blood and the central nervous system. Due to
485 a potential lower ATP synthesis level in serotypes 1, 5, and 12F (due to the lack of RD8a), they may
486 prefer to use PTS transporters to uptake sugars such as fructose and lactose, and that is why genes that
487 encode PTS transporters are significantly more present in the hyper-invasive serotypes. In addition to
488 the sugar uptake, PTS transporters regulate several pathways in bacteria, such as gene expression and
489 communication between cells. Thus, the phenotypic effects of the PTS transporters should not be
490 limited just to their ability to import carbohydrates [66].

491 Genes from RD10 were absent from serotypes 1 and 12F. However, they were conserved in serotypes 5,
492 16F, and 19F. Moreover, the location-based gene presence-absence analysis showed that RD10 was
493 significantly present in the nasopharyngeal samples compared to samples from sterile sites. These genes
494 have homology to the general secretion pathway protein B sceA2/Y2 system components in
495 *Streptococcus gordonii* that contribute to the secretion of the general secretion pathway protein B
496 involved in infective endocarditis [67]. In the *S. pneumoniae* genome, the homolog of general secretion
497 pathway protein B is the PsrP that is transported to the bacterial cell surface by the SecA2/Y2 system
498 encoded by genes in RD10. Studies conducted on *Streptococcus gordonii* showed that the presence of
499 SecA2/Y2 promotes the adhesion to epithelial cells in the nasopharynx and erythrocytes in the blood
500 [68][69], which is why SecA/Y2 is significantly present in nasopharyngeal samples and serotype 5
501 (abundant in the blood). The presence of the secA2/Y2-like component should also facilitate the export
502 of pneumolysin that enhances adhesion to the host cell and survival in the blood [70][71].

503 In conclusion, some of the genes specifically present (absent) in the hyper-invasive serotypes (1, 5, and
504 12F) may contribute to the invasiveness and low colonization rate. However, experimental verification is
505 required to confirm the outcome of computational work in this study. The serotype is the best driver of
506 the pneumococcal population structure, however, the study characterized the genetic divergence of
507 serotypes 1, 5, and 12F to be much greater than other serotypes. Their significant presence in the blood
508 and CSF accounted for the most pronounced genomic and functional differences observed between the
509 nasopharynx and sterile sites. The low frequency of serotypes 1, 5, and 12F among carriers may be due
510 to their short-term colonization before entering the sterile sites. These invasive serotypes harbor the
511 elements of PTS transporters while they lack genes from RD8a. In serotype 5, RD10 is highly conserved,
512 but this region was absent from serotypes 1 and 12F. The study demonstrated that isolation sites do not
513 significantly influence the genomic structure of pneumococcal isolates. Although a few genes were
514 associated with the virulence of commonly present serotypes in both nasopharynx and sterile sites, we
515 speculated that the differences between them could be captured by other high throughput techniques

such as gene expression analysis. In summary, the results shed light on the pneumococcal population structure and serotypes in Malawi. The unique cluster of significant genes in the hyper-invasive serotypes and highly conserved core genes could be potential therapeutic targets.

Author contributions

Sample collection, metadata curation, and genome sequencing: Jennifer Cornick, Dean Everett, Anmol Kiran, Chrispin Chaguza, and Chikondi Peno.

Methodology and data analysis: Arash Iranzadeh, Anmol Kiran, and Arghavan Alisoltani.

Result interpretation: Arash Iranzadeh, Arghavan Alisoltani, Nicola Mulder, and Dean Everett.

Initial manuscript writing: Arash Iranzadeh.

Review of the manuscript: Arash Iranzadeh, Arghavan Alisoltani, Anmol Kiran, Robert F Breiman, Chrispin Chaguza, Chikondi Peno, Dean B Everett, Nicola Mulder.

All authors have given consent to participate in the study.

Acknowledgment

Computations were performed using facilities provided by the University of Cape Town's ICTS High-Performance Computing team: hpc.uct.ac.za. The authors also acknowledge the Centre for High-Performance Computing (CHPC), South Africa, for providing computational resources to this research project. We thank the study participants and all involved staff at the Karonga Prevention Study and the Malawi-Liverpool-Wellcome Trust Clinical Research Programme. We thank Olivier Koole and Naor Bar-Zeev for their scientific input.

Funding information

We are grateful for the financial support received from the Wellcome Trust and Institute of Infection and Global Health. The study was supported by Wellcome Trust grant number 079828. We would also like to acknowledge the contribution of Global Pneumococcal Sequencing project

539 (https://www.pneumogen.net/gps/), funded by the Bill and Melinda Gates Foundation, and
540 Pneumococcal African Genomic Consortium (http://www.pagegenomes.org) in the generation of data
541 used in this publication.

542 Conflicts of interest

543 The author(s) declare that there are no conflicts of interest.

544 Ethical approval

545 Not required as the research delas with bacterial samples.

546 References

547 [1] T. D. Swarthout *et al.*, 'High residual carriage of vaccine-serotype Streptococcus pneumoniae
548 after introduction of pneumococcal conjugate vaccine in Malawi', *Nature Communications 2020*
549 *11:1*, vol. 11, no. 1, pp. 1–12, May 2020, doi: 10.1038/s41467-020-15786-9.

550 [2] P. Kamthunzi, 'Impact of PCV13 vaccination in Blantyre, Malawi', *Lancet Glob Health*, vol. 9, no. 7,
551 pp. e893–e894, Jul. 2021, doi: 10.1016/S2214-109X(21)00258-8.

552 [3] L. Paixão *et al.*, 'Host glycan sugar-specific pathways in streptococcus pneumonia: Galactose as a
553 key sugar in colonisation and infection', *PLoS One*, 2015, doi: 10.1371/journal.pone.0121042.

554 [4] D. Bogaert, R. de Groot, and P. W. M. Hermans, 'Streptococcus pneumoniae colonisation: The key
555 to pneumococcal disease', *Lancet Infectious Diseases*. 2004. doi: 10.1016/S1473-3099(04)00938-
556 7.

557 [5] E. Backhaus *et al.*, 'Epidemiology of invasive pneumococcal infections: Manifestations, incidence
558 and case fatality rate correlated to age, gender and risk factors', *BMC Infect Dis*, 2016, doi:
559 10.1186/s12879-016-1648-2.

560 [6] A. Kadioglu, J. N. Weiser, J. C. Paton, and P. W. Andrew, 'The role of Streptococcus pneumoniae
561 virulence factors in host respiratory colonization and disease', *Nature Reviews Microbiology*.
562 2008. doi: 10.1038/nrmicro1871.

563 [7] J. N. Weiser, D. M. Ferreira, and J. C. Paton, 'Streptococcus pneumoniae: Transmission,
564 colonization and invasion', *Nature Reviews Microbiology*. 2018. doi: 10.1038/s41579-018-0001-8.

565 [8] K. A. Geno *et al.*, 'Pneumococcal capsules and their types: Past, present, and future', *Clin*
566 *Microbiol Rev*, 2015, doi: 10.1128/CMR.00024-15.

567 [9] A. L. Nelson, A. M. Roche, J. M. Gould, K. Chim, A. J. Ratner, and J. N. Weiser, 'Capsule enhances
568 pneumococcal colonization by limiting mucus-mediated clearance', *Infect Immun*, 2007, doi:
569 10.1128/IAI.01475-06.

570 [10] J. Brown, S. Hammerschmidt, and C. Orihuela, *Streptococcus Pneumoniae: Molecular*
571 *Mechanisms of Host-Pathogen Interactions*. 2015. doi: 10.1016/C2012-0-00722-3.

572 [11] F. Ganaie *et al.*, 'A new pneumococcal capsule type, 10D, is the 100th serotype and has a large
573 cps fragment from an oral streptococcus', *mBio*, 2020, doi: 10.1128/mBio.00937-20.

574 [12] D. M. Weinberger, R. Malley, and M. Lipsitch, 'Serotype replacement in disease after
575 pneumococcal vaccination', *The Lancet*. 2011. doi: 10.5455/apd.239006.

576 [13] D. R. Feikin *et al.*, 'Serotype-Specific Changes in Invasive Pneumococcal Disease after
577 Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites',
578 *PLoS Med*, 2013, doi: 10.1371/journal.pmed.1001517.

579 [14] E. D. McCollum *et al.*, 'Impact of the 13-valent pneumococcal conjugate vaccine on clinical and
580 hypoxemic childhood pneumonia over three years in central Malawi: An observational study',
581 *PLoS One*, 2017, doi: 10.1371/journal.pone.0168209.

582 [15] N. Bar-Zeev *et al.*, 'Impact and Effectiveness of 13-Valent Pneumococcal Conjugate Vaccine on
583 Population Incidence of Vaccine and Non-Vaccine Serotype Invasive Pneumococcal Disease in
584 Blantyre, Malawi, 2006-2018: Prospective Observational Time-Series and Case-Control Studies',
585 *SSRN Electronic Journal*, Dec. 2020, doi: 10.2139/SSRN.3745169.

586 [16] E. Heinsbroek *et al.*, 'Pneumococcal carriage in households in Karonga District, Malawi, before
587 and after introduction of 13-valent pneumococcal conjugate vaccination', *Vaccine*, 2018, doi:
588 10.1016/j.vaccine.2018.10.021.

589 [17] C. Cillóniz, C. Garcia-Vidal, A. Ceccato, and A. Torres, 'Antimicrobial Resistance Among
590 Streptococcus pneumoniae', in *Antimicrobial Resistance in the 21st Century*, Springer, 2018, pp.
591 13–38.

592 [18] A. Iranzadeh and N. J. Mulder, 'Bacterial Pan-Genomics', in *Microbial Genomics in Sustainable*
593 *Agroecosystems*, Springer, 2019, pp. 21–38.

594 [19] A. Embry, E. Hinojosa, and C. J. Orihuela, 'Regions of Diversity 8, 9 and 13 contribute to
595 Streptococcus pneumoniae virulence', *BMC Microbiol*, 2007, doi: 10.1186/1471-2180-7-80.

596 [20] R. Brückner, M. Nuhn, P. Reichmann, B. Weber, and R. Hakenbeck, 'Mosaic genes and mosaic
597 chromosomes-genomic variation in Streptococcus pneumoniae', *International Journal of Medical*
598 *Microbiology*. 2004. doi: 10.1016/j.ijmm.2004.06.019.

599 [21] A. W. Kamng'ona *et al.*, 'High multiple carriage and emergence of Streptococcus pneumoniae
600 vaccine serotype variants in Malawian children', *BMC Infect Dis*, vol. 15, no. 1, pp. 1–11, Jun.
601 2015, doi: 10.1186/S12879-015-0980-2/FIGURES/7.

602 [22] C. Chaguza *et al.*, 'Population genetic structure, antibiotic resistance, capsule switching and
603 evolution of invasive pneumococci before conjugate vaccination in Malawi', *Vaccine*, vol. 35, no.
604 35 Pt B, pp. 4594–4602, Aug. 2017, doi: 10.1016/J.VACCINE.2017.07.009.

605 [23] S. Andrews, 'FastQC: A quality control tool for high throughput sequence data.', *Babraham*
606 *Bioinformatics*, p. http://www.bioinformatics.babraham.ac.uk/projects/, 2010, doi: citeulike-
607 article-id:11583827.

608 [24] L. Epping, A. J. van Tonder, R. A. Gladstone, S. D. Bentley, A. J. Page, and J. A. Keane, 'SeroBA:
609 rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence
610 data', *Microb Genom*, vol. 4, no. 7, Jul. 2018, doi: 10.1099/MGEN.0.000186.

611 [25] S. Gladman and T. Seemann, 'VelvetOptimiser', *Free Software Foundation*. 2008. doi:
612 10.1016/S0925-8574(99)00040-3.

613 [26] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, 'QUAST: Quality assessment tool for genome
614 assemblies', *Bioinformatics*, 2013, doi: 10.1093/bioinformatics/btt086.

615 [27] T. Seemann, 'Prokka: Rapid prokaryotic genome annotation', *Bioinformatics*, vol. 30, no. 14, pp.
616 2068–2069, 2014, doi: 10.1093/bioinformatics/btu153.

617 [28] A. J. Page *et al.*, 'Roary: Rapid large-scale prokaryote pan genome analysis', *Bioinformatics*, vol.
618 31, no. 22, pp. 3691–3693, May 2015, doi: 10.1093/bioinformatics/btv421.

619 [29] K. Katoh and D. M. Standley, 'MAFFT multiple sequence alignment software version 7:
620 Improvements in performance and usability', *Mol Biol Evol*, 2013, doi: 10.1093/molbev/mst010.

621 [30] B. Q. Minh *et al.*, 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
622 Genomic Era', *Mol Biol Evol*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi:
623 10.1093/MOLBEV/MSAA015.

624 [31] I. Letunic and P. Bork, 'Interactive tree of life (iTOL) v3: an online tool for the display and
625 annotation of phylogenetic and other trees', *Nucleic Acids Res*, 2016, doi: 10.1093/nar/gkw290.

626 [32] R. Gaujoux, 'An introduction to NMF package', *BMC Bioinformatics*, 2010, doi: 10.1186/1471-
627 2105-11-367.

628 [33] F. Rohart, B. Gautier, A. Singh, and K. A. Lê Cao, 'mixOmics: An R package for 'omics feature
629 selection and multiple data integration', *PLoS Comput Biol*, 2017, doi:
630 10.1371/journal.pcbi.1005752.

631 [34] O. Brynildsrud, J. Bohlin, L. Scheffer, and V. Eldholm, 'Rapid scoring of genes in microbial pan-
632 genome-wide association studies with Scoary', *Genome Biol*, vol. 17, no. 1, p. 238, 2016, doi:
633 10.1186/s13059-016-1108-8.

634 [35] D. Szklarczyk *et al.*, 'STRING v11: Protein-protein association networks with increased coverage,
635 supporting functional discovery in genome-wide experimental datasets', *Nucleic Acids Res*, 2019,
636 doi: 10.1093/nar/gky1131.

637 [36] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, 'KEGG: Kyoto encyclopedia of
638 genes and genomes', *Nucleic Acids Research*, vol. 27, no. 1. pp. 29–34, 1999. doi:
639 10.1093/nar/27.1.29.

640 [37] A. Bateman *et al.*, 'The Pfam protein families database', *Nucleic Acids Res*, vol. 32, no. suppl_1,
641 pp. D138–D141, Jan. 2004, doi: 10.1093/NAR/GKH121.

642 [38] Q. Wang *et al.*, 'Serotype distribution of Streptococcus pneumoniae isolated from children
643 hospitalized in Beijing children's hospital (2013–2019)', *Vaccine*, vol. 38, no. 49, pp. 7858–7864,
644 2020, doi: 10.1016/j.vaccine.2020.10.005.

645 [39] M. Alizadeh Chamkhaleh *et al.*, 'Serotype distribution of Streptococcus pneumoniae among
646 healthy carriers and clinical patients: a systematic review from Iran', *European Journal of Clinical*
647 *Microbiology and Infectious Diseases*. 2020. doi: 10.1007/s10096-020-03963-z.

648 [40] N. L. Hiller *et al.*, 'Comparative genomic analyses of seventeen Streptococcus pneumoniae
649 strains: Insights into the pneumococcal supragenome', *J Bacteriol*, vol. 189, no. 22, pp. 8186–
650 8195, Nov. 2007, doi: 10.1128/JB.00690-07/SUPPL_FILE/SUPPLEMENTARY_TABLE_3.ZIP.

651 [41] S. Zhang and R. Meyer, 'The relaxosome protein MobC promotes conjugal plasmid mobilization
652 by extending DNA strand separation to the nick site at the origin of transfer', *Mol Microbiol*,
653 1997, doi: 10.1046/j.1365-2958.1997.4861849.x.

654 [42] C. Obert *et al.*, 'Identification of a candidate Streptococcus pneumoniae core genome and
655 regions of diversity correlated with invasive pneumococcal disease', *Infect Immun*, vol. 74, no. 8,
656 pp. 4766–4777, 2006, doi: 10.1128/IAI.00316-06.

657 [43] K. Yokoyama and H. Imamura, 'Rotation, structure, and classification of prokaryotic V-ATPase',
658 *Journal of Bioenergetics and Biomembranes*. 2005. doi: 10.1007/s10863-005-9480-1.

659 [44] J. Reizer, A. Reizer, and M. H. Saier, 'A functional superfamily of sodium/solute symporters',
660 *Biochim Biophys Acta*, vol. 1197, no. 2, pp. 133–166, Jun. 1994, doi: 10.1016/0304-
661 4157(94)90003-5.

662 [45] P. Gaudu, 'Aerobic respiration metabolism in lactic acid bacteria and uses in biotechnology', doi:
663 10.1146/annurev-food-022811-101255.

664 [46] C. Manzano *et al.*, 'Sortase-Mediated Pilus Fiber Biogenesis in Streptococcus pneumoniae',
665 *Structure*, 2008, doi: 10.1016/j.str.2008.10.007.

666 [47] J. Lemieux, S. Woody, and A. Camilli, 'Roles of the sortases of Streptococcus pneumoniae in
667 assembly of the RlrA pilus', *J Bacteriol*, 2008, doi: 10.1128/JB.00379-08.

668 [48] L. Williams, F. Stapleton, and N. Carnt, 'Microbiology, lens care and maintenance', *Contact*
669 *Lenses*, pp. 65–96, Jan. 2019, doi: 10.1016/B978-0-7020-7168-3.00004-0.

670 [49] J. D. Langereis and M. I. de Jonge, ' Non-encapsulated Streptococcus pneumoniae , vaccination as
671 a measure to interfere with horizontal gene transfer ', *Virulence*, 2017, doi:
672 10.1080/21505594.2017.1309492.

673 [50] D. van der Windt, H. J. Bootsma, P. Burghout, C. E. van der Gaast-de Jongh, P. W. M. Hermans,
674 and M. van der Flier, 'Nonencapsulated Streptococcus pneumoniae resists extracellular human
675 neutrophil elastase- and cathepsin G-mediated killing', *FEMS Immunol Med Microbiol*, 2012, doi:
676 10.1111/j.1574-695X.2012.01028.x.

677 [51] U. M. Talbot, A. W. Paton, and J. C. Paton, 'Uptake of Streptococcus pneumoniae by respiratory
678 epithelial cells', *Infect Immun*, 1996.

679 [52] C. J. Lee, S. D. Banks, and J. P. Li, 'Virulence, immunity, and vaccine related to streptococcus
680 pneumoniae', *Crit Rev Microbiol*, 1991, doi: 10.3109/10408419109113510.

681 [53] L. E. Keller, D. A. Robinson, and L. S. McDaniel, ' Nonencapsulated Streptococcus pneumoniae: :
682 Emergence and Pathogenesis ', *mBio*, 2016, doi: 10.1128/mbio.01792-15.

683 [54] W. P. Hausdorff, 'The roles of pneumococcal serotypes 1 and 5 in paediatric invasive disease',
684 *Vaccine*, 2007, doi: 10.1016/j.vaccine.2006.09.009.

685 [55] J. E. Cornick *et al.*, 'Region-specific diversification of the highly virulent serotype 1 Streptococcus
686 pneumoniae', *Microb Genom*, vol. 1, no. 2, 2015.

687 [56] J. Leimkugel *et al.*, 'An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern
688 Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics', *J*
689 *Infect Dis*, vol. 192, no. 2, pp. 192–199, 2005.

690 [57] B. D. Gessner, J. E. Mueller, and S. Yaro, 'African meningitis belt pneumococcal disease
691 epidemiology indicates a need for an effective serotype 1 containing vaccine, including for older
692 children and adults', *BMC Infect Dis*, 2010, doi: 10.1186/1471-2334-10-22.

693 [58] J. E. Cornick *et al.*, 'Invasive streptococcus pneumoniae in children, Malawi, 2004-2006', *Emerg*
694 *Infect Dis*, 2011, doi: 10.3201/eid1706.101404.

695 [59] C. Obert *et al.*, 'Identification of a Candidate Streptococcus pneumoniae core genome and
696 regions of diversity correlated with invasive pneumococcal disease', *Infect Immun*, vol. 74, no. 8,
697 pp. 4766–4777, Aug. 2006, doi: 10.1128/IAI.00316-06.

698 [60] D. M. Walters, V. L. Stirewalt, and S. B. Melville, 'Cloning, sequence, and transcriptional
699 regulation of the operon encoding a putative N-acetylmannosamine-6-phosphate epimerase
700 (nanE) and sialic acid lyase (nanA) in Clostridium perfringens', *J Bacteriol*, 1999.

701 [61] L. E. Bakeeva, K. M. Chumakov, A. L. Drachev, A. L. Metlina, and V. P. Skulachev, 'The sodium
702 cycle. III. Vibrio alginolyticus resembles Vibrio cholerae and some other vibriones by flagellar
703 motor and ribosomal 5S-RNA structures', *BBA - Bioenergetics*, 1986, doi: 10.1016/0005-
704 2728(86)90115-5.

705 [62] M. H. Saier, 'Families of transmembrane sugar transport proteins', *Molecular Microbiology*. 2000.
706 doi: 10.1046/j.1365-2958.2000.01759.x.

707 [63] P. D. Boyer, 'THE ATP SYNTHASE—A SPLENDID MOLECULAR MACHINE', *Annu Rev Biochem*, 2002,
708 doi: 10.1146/annurev.biochem.66.1.717.

709 [64] A. Bidossi *et al.*, 'A functional genomics approach to establish the complement of carbohydrate
710 transporters in Streptococcus pneumoniae', *PLoS One*, 2012, doi: 10.1371/journal.pone.0033320.

711 [65] C. M. Buckwalter and S. J. King, 'Pneumococcal carbohydrate transport: Food for thought', *Trends*
712 *in Microbiology*. 2012. doi: 10.1016/j.tim.2012.08.008.

713 [66] M. H. Saier, 'The Bacterial Phosphotransferase System: New Frontiers 50 Years after Its
714 Discovery', *Journal of Molecular Microbiology and Biotechnology*. 2015. doi: 10.1159/000381215.

715 [67] B. A. Bensing, B. W. Gibson, and P. M. Sullam, 'The Streptococcus gordonii Platelet Binding
716 Protein GspB Undergoes Glycosylation Independently of Export', *J Bacteriol*, 2004, doi:
717 10.1128/JB.186.3.638-645.2004.

718 [68] B. A. Bensing and P. M. Sullam, 'Transport of preproteins by the accessory Sec system requires a
719 specific domain adjacent to the signal peptide', *J Bacteriol*, 2010, doi: 10.1128/JB.00373-10.

720 [69] M. Yamaguchi *et al.*, 'Streptococcus pneumoniae Invades Erythrocytes and Utilizes Them to
721 Evade Human Innate Immunity', *PLoS One*, 2013, doi: 10.1371/journal.pone.0077282.

722 [70] M. Bandara *et al.*, 'The accessory Sec system (SecY2A2) in Streptococcus pneumoniae is involved
723 in export of pneumolysin toxin, adhesion and biofilm formation', *Microbes Infect*, 2017, doi:
724 10.1016/j.micinf.2017.04.003.

725 [71] R. Wu and H. Wu, 'A molecular chaperone mediates a two-protein enzyme complex and
726 glycosylation of serine-rich streptococcal adhesins', *Journal of Biological Chemistry*, 2011, doi:
727 10.1074/jbc.M111.239350.

728

729

730 ## Supporting information

731 **S1 Fig. Characteristics of the 1477 pneumococcal isolates used in the study.** (a) The relative frequency
732 of serotypes in the entire cohort, samples were assigned to 56 serotypes. For each sample, the in-silico
733 serotyping was accomplished by SeroBA. (b) Frequency of isolates in the pre- and post-PCV13 eras in
734 Malawi. (c) Frequency of isolates obtained from each specimen source.

735 **S2 Fig. The pan-genome of 1477 pneumococcal samples isolated in Malawi was obtained from 1997 to**
736 **2015.** The pan-genome is an open pan-genome, which means the number of total genes increases
737 unlimitedly when the sample size grows. The dashed line represents the number of total genes, and the
738 solid line represents the number of conserved genes in the pan-genome.

739 **S3 Fig. The three-dimensional PCA of the gene distribution in the vaccine types.** For each serotype and
740 for downsampling, 10 samples were randomly selected from the nasopharynx, blood, and CSF. The PCA
741 was conducted using the R package MixOmics. Hyper-invasive serotypes 1, 5, and 12F clustered
742 separately from other strains.

743 **S4 Fig. Genes in RD10 are absent from serotypes 1 and 12F but conserved in serotype 5, 16F, and 19F.**
744 Genes from RD10 encode the components of the secretory system SecA2/Y2 that transports
745 glycoproteins to the bacterial cell surface, which are required for binding to the human proteins on the
746 surface of epithelial cells and erythrocytes.

747 **S5 Fig. RD8a consists of two operons RD8a1 (SP_1315-1324) and RD8a2 (SP_1325-1331).** This region is
748 not detected in the significant invasive serotypes 1, 5, and 12F, but it is present in more than 80% of
749 serotype 16F and 19F that significantly dominates the nasopharynx. The important biological processes
750 carried out by these genes are the transport of ions across the membrane and the synthesis of ATP
751 molecules.

752 **S1 Table. Statistical analysis of serotypes' prevalence across specimen sources.**

753 **S2 Table. Gene presence-absence analysis (Invasive vs Nasopharyngeal, serotypes 1, 5, and 12F were**
754 **excluded).**

755 **S3 Table. Gene presence-absence analysis (Serotype 1 vs 16F & 19F).**

756 **S4 Table. Gene presence-absence analysis (Serotype 5 vs 16F & 19F).**

757 **S5 Table. Gene presence-absence analysis (Serotype 12F vs 16F & 19F).**

758 Figure captions

759 **Fig 1. The distribution of the 56 pneumococcal serotypes assigned to 1477 samples from Malawi.** (a)
760 The relative frequency of each serotype in the nasopharynx of carriers, the blood of bacteremia
761 patients, and the CSF of meningitis patients is shown in blue, red, and yellow, respectively (UT: Un-
762 Typeable). (b) The log-transformed odds ratio of the significantly over- and under-abundant serotypes in
763 the sterile sites (blood and CSF). Fisher's exact test was applied to identify serotypes with a significant
764 differential abundance among carriers and patients (nasopharynx and sterile sites) at the significance
765 level of the Benjamini-Hochberg adjusted p-value < 0.01 (BH: Benjamini-Hochberg).

766 **Fig 2. The pan-genome matrix of 1477 pneumococcal isolates from Malawi.** The pan-genome is
767 visualized as a gene presence-absence heatmap that represents the hierarchical unsupervised clustering
768 of samples based on the distribution of genes in the pan-genome. Each row is a sample, and each
769 column is a gene. A blue dot denotes the presence of each gene. On the right side of the heatmap, the
770 large blue block shows core genes present in all samples. The left side of the heatmap represents the
771 accessory genome along with the clustering bands. In addition to the significant serotypes 1, 5, 12F, 16F,
772 and 19F, other abundant serotypes, including 6A, 6B, and 23F, as well as serotypes with source-based p-
773 value < 0.05, including 21, 11A, and 15B, are also highlighted on the heatmap.

774 **Fig 3. The phylogenetic population structure of 1477 pneumococcal samples from Malawi.** The
775 phylogenetic tree was built based on the multiple sequence alignment of the core genome using the
776 maximum likelihood method. Colors on the loops show the serotypes, specimen sources (isolation sites),
777 and PCV13 eras. In addition to the significant serotypes 1, 5, 12F, 16F, and 19F, other abundant

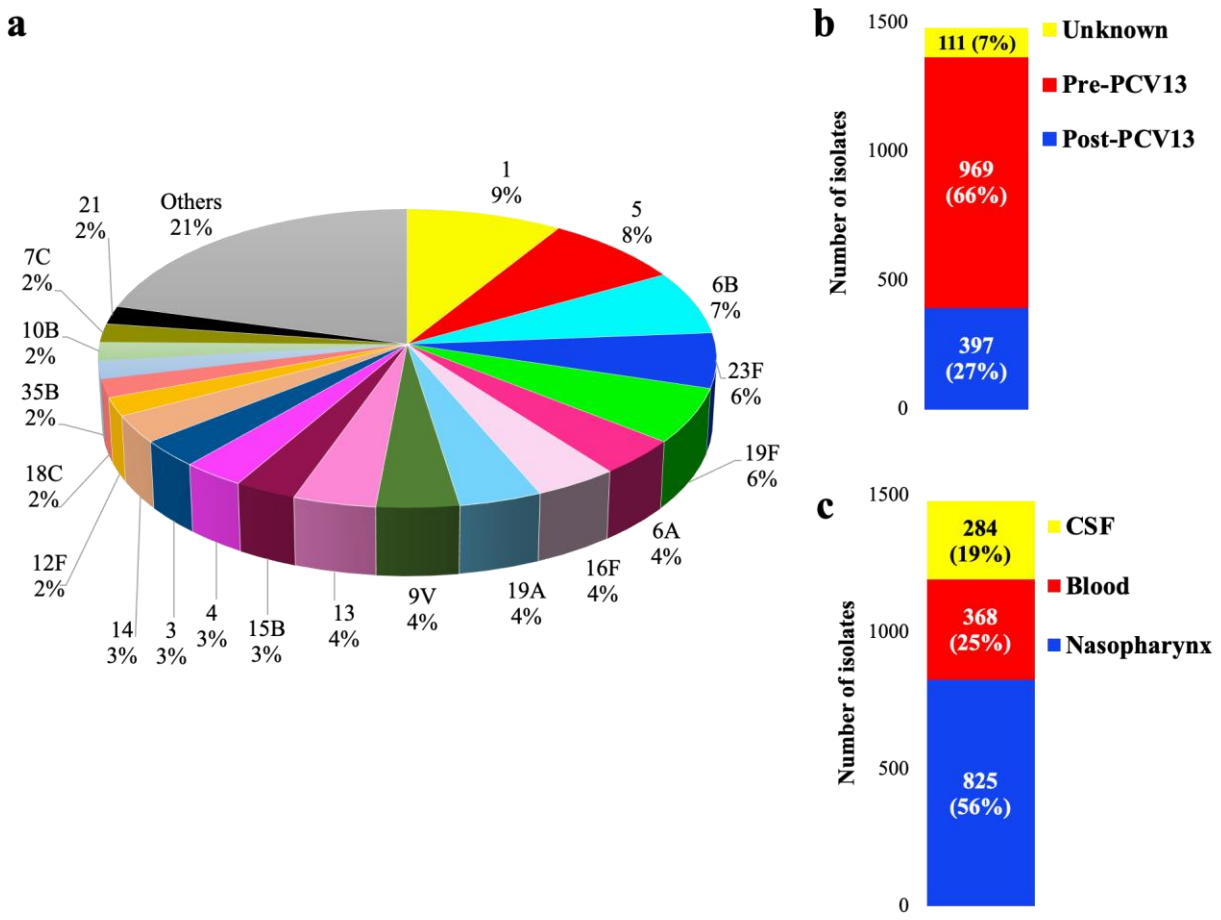
778 serotypes, including 6A, 6B, and 23F, as well as serotypes with source-based p-value < 0.05, including 21,
779 11A, and 15B, are also highlighted on the tree.

780 **Fig 4. The PCA of the gene distribution in the pan-genome of pneumococcal isolates from 1477**
781 **Malawians.** The PCA of variants (gene presence-absence) in the accessory-genome indicates the
782 influence of (a) specimen sources (isolation sites), (b) serotypes, (c) PCV13 (vaccination) era, and (d)
783 geographical locations on the gene presence-absence profile of pneumococcal isolates in Malawi.
784 Serotypes 1 and 5 were clearly separated from other samples.

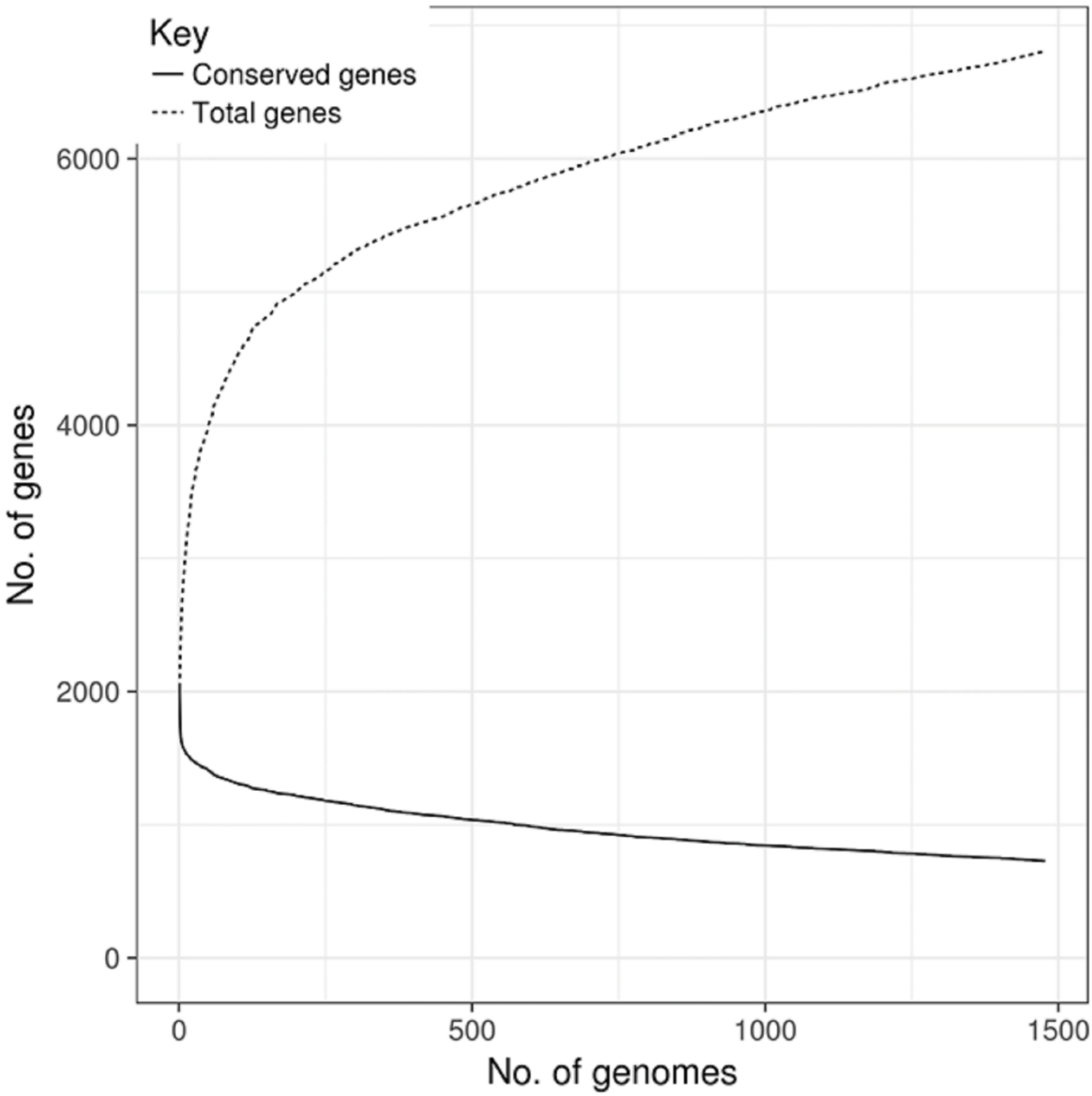
785 **Fig 5. The summary of the gene presence-absence analysis. (a) The number of significant genes**
786 **present-absent in the hyper-invasive serotypes.** The gene presence-absence analysis was applied using
787 Scoary to compare the gene pools of the hyper-invasive serotypes and serotypes 16F and 19F. P-values
788 were corrected by the Bonferroni method, and significant genes had an adjusted p-value of less than
789 0.05. (b) The significant presence and absence of RDs in samples from blood and CSF is shown as a
790 presence-absence heatmap.

791 Data summary

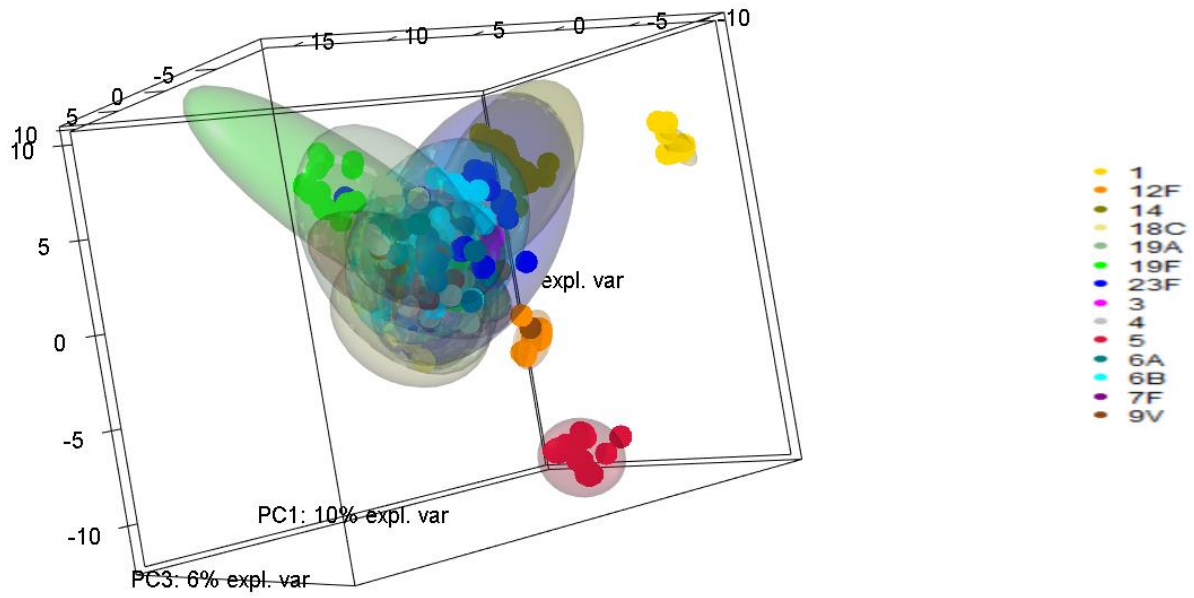
792 **S6 Table. Samples IDs on European Nucleotide Archive (ENA)**



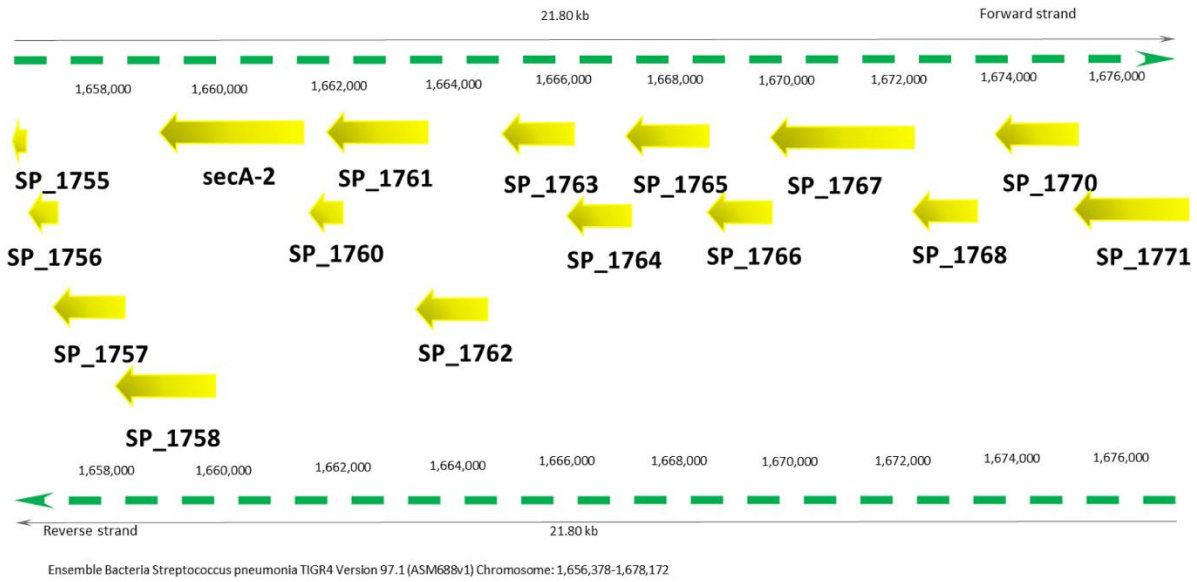
S1 Fig. Characteristics of the 1477 pneumococcal isolates used in the study. (a) The relative frequency of serotypes in the entire cohort, samples were assigned to 56 serotypes. For each sample, the in-silico serotyping was accomplished by SeroBA. (b) Frequency of isolates in the pre- and post-PCV13 eras in Malawi. (c) Frequency of isolates obtained from each specimen source.



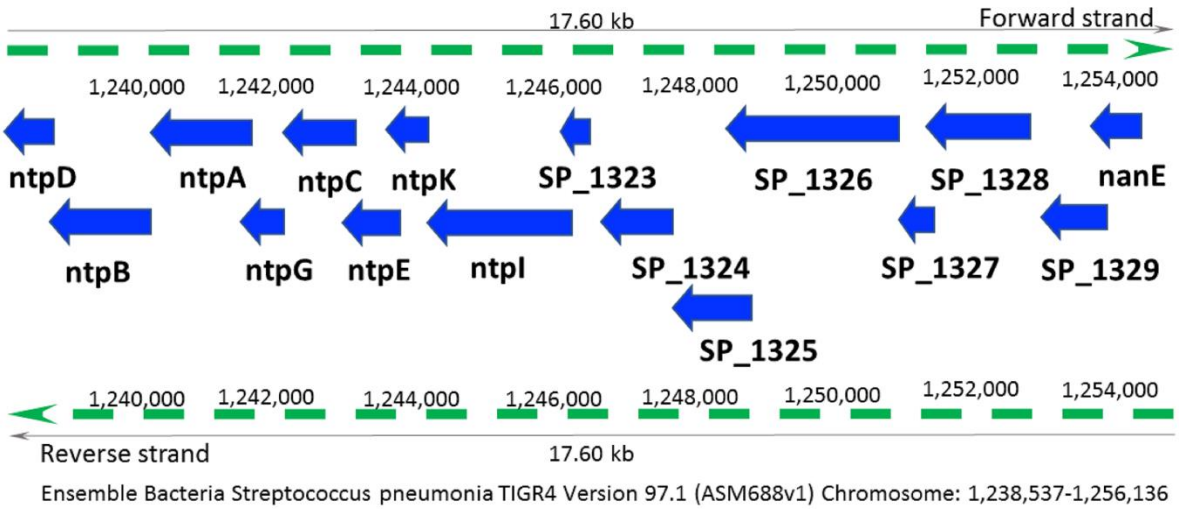
S2 Fig. The pan-genome of 1477 pneumococcal samples isolated in Malawi was obtained from 1997 to 2015. The pan-genome is an open pan-genome, which means the number of total genes increases unlimitedly when the sample size grows. The dashed line represents the number of total genes, and the solid line represents the number of conserved genes in the pan-genome.



S3 Fig. The three-dimensional PCA of the gene distribution in the vaccine types. For each serotype and for down sampling, 10 samples were randomly selected from the nasopharynx, blood, and CSF. The PCA was conducted using the R package MixOmics. Hyper-invasive serotypes 1, 5, and 12F clustered separately from other strains.



S4 Fig. Genes in RD10 are absent from serotype 1 and 12F but conserved in serotype 5, 16F, and 19F. Genes from RD10 encode the components of the secretory system SecA2/Y2 that transports glycoproteins to the bacterial cell surface, which are required for binding to the human proteins on the surface of epithelial cells and erythrocytes.



S5 Fig. RD8a consists of two operons RD8a1 (SP_1315-1324) and RD8a2 (SP_1325-1331). This region is not detected in the significant invasive serotypes 1, 5, and 12F, but it is present in more than 80% of serotype 16F and 19F that significantly dominates the nasopharynx. The important biological processes carried out by these genes are the transport of ions across the membrane and the synthesis of ATP molecules.