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

Abstract

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

Introduction

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

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

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

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

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

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

Materials and methods

Study cohort and pneumococcal isolates

Pneumococcal isolates used in the study were from patients at three central hospitals in Malawi: Queen Elizabeth Central Hospital in Blantyre, Karonga Central Hospital in Karonga, and Kamuzu Central Hospital in Lilongwe. Samples were collected from people visiting Queen Elizabeth Central Hospital during 1997-2015 (53%) and from households in the Karonga district during 2009-2014 as part of two household studies (44%) (21). The remainder of the samples were collected from Kamuzu Central Hospital. Isolates were obtained either from the nasopharynx of carriers or the blood and cerebrospinal fluid (CSF) of patients and archived at the Malawi-Liverpool Wellcome Trust Clinical Research Programme Centre. In this study, the term "sterile sites" refers to blood and CSF, and "invasive samples" are those obtained from the sterile sites.

Whole-genome sequencing and quality control

Archived samples were sequenced under the Global Pneumococcal Sequencing project and Pneumococcal African Genomic Consortium at Wellcome Trust Sanger Institute in the United Kingdom. The bacterial DNA was extracted using QIAamp DNA mini kit and QIAgen Biorobot manufactured by QIAGEN. The whole genome sequencing was accomplished using the Illumina Genome Analyzer II and HiSeq platforms, generating 125 nucleotides paired-end reads. The quality of the reads was evaluated by Fastqc (22).

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

SeroBA version 1.23.4 was used to infer the serotype of samples (23). SeroBA applies a k-mer method to determine serotypes directly from the paired-end reads in FASTQ format. Any serotype with a relative frequency greater than 5% was considered as an abundant serotype. Fisher's exact test was applied to identify serotypes with a significant presence in the nasopharynx and sterile sites. P-values were

adjusted by the Benjamini-Hochberg method, and serotypes with an adjusted p-value less than 0.01 were considered significant. The odds ratio (OR) was calculated as follows OR = (*ad*)/(*bc*), where for serotype *k*, a is the number of invasive serotype *k*, *b* is the number of carriage serotype *k*, *c* is the number of invasive non-serotype *k,* and *d* is the number of carriage non-serotype *k*. For OR calculation, zero values were replaced by 0.5. Any abundant serotype with a significant presence in the nasopharynx was considered to have low invasiveness, and high invasiveness was assigned to abundant serotypes with a significant presence in the sterile sites. Fisher's exact test was also applied to identify serotypes whose frequencies changed significantly after introducing PCV13 in 2011 (adjusted p-value < 0.01).

Genome assembly and annotation

Genomes were assembled using Velvet Optimiser version 2.2.5 (24). The assembler was set to produce contigs longer than 500 base pairs. The starting hash value was 61, and the end hash value was 119. The quality of the assembled genomes was evaluated by Quast version 5.2(25). The assembled genomes were annotated using Prokka version 13.1 (26).

Pan-genome construction

The pan-genome for samples was generated using Roary version 3.12.0 (27). Roary was run to perform the core gene alignments with MAFFT version 7.313 (28). Genes in the pan-genome were divided into three groups based on their abundance among samples. Genes that existed in 100% of samples were considered as core genes, those that were not core but present in more than 95% of samples were regarded as a soft-core genes, and the rest were considered as accessory genes.

Analysis of the population structure

Since core genes are conserved between all samples, diversity in the core genome exists in the form of small-scale variations such as single nucleotide polymorphisms (SNPs) and short indels. SNPs and indels distribution in the core-genome was analyzed to gain insight into the population structure. The genetic

separation between samples was illustrated by the phylogenetic tree built from SNPs and indels in the alignment of core genes as the phylogeny markers. The core genome alignment was used as input into the IQ-TREE version 2 (29) to draw a phylogenetic tree visualized by iTol version 3 (30).

Diversity in the accessory genome exists in the form of large-scale gene presence-absence variations. The R package Nonnegative Matrix Factorization (31) visualized the pan-genome matrix as a gene presence-absence heatmap. Furthermore, to determine which factors, including isolation sites (nasopharynx, blood, and CSF), serotypes, geographical locations, and vaccination eras, have the greatest impact on the gene presence-absence profile of samples, a PCA of gene distribution was accomplished by the R package MixOmics version 6.20.0 (32).

Gene presence-absence statistical analysis

The phenotypic trait of samples was assigned based on the population structure and isolates' invasiveness. The gene presence-absence statistical test was performed using Scoary version 1.6.1643 (33). This tool scores the components of the pan-genome for associations with observed phenotypic traits while accounting for population stratification. The test was conducted between samples from different sources and serotypes to find putative virulence factors. Every gene with a Bonferroni corrected p-value less than 0.05 was reported as a significant gene.

Functional and gene ontology (GO) enrichment analysis

The list of significant genes was submitted to STRING webtool version 11.5 (34) for the functional enrichment analysis. The STRING database is a network that integrates information from all publicly available protein-protein interaction databases. The STRING's computational pipeline can predict direct (physical) and indirect (functional) interactions between proteins. Interactions in STRING are derived from 5 sources, including (i) genomic context predictions, (ii) high-throughput lab experiments, (iii) co-expression, (iv) automated text mining, and (v) previous knowledge in databases. Functional enrichment analysis in STRING uses information from classification systems such as the Kyoto Encyclopedia of Genes

and Genomes (KEGG) (35) and the Protein families database (Pfam) (36). The tool entitled "Multiple Sequences" was selected, and the *S. pneumoniae TIGR4* was chosen as the reference organism. STRING reports the associated enriched pathways with a false discovery rate of less than 0.05.

Results

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

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

Characteristics	**Categories**	**Nasopharynx**	**Blood**	**CSF**
Age (in years)	< 5	538	165	141
	5-19	109	42	60
	20-40	60	67	50
	> 40	7	24	7
	Unknown	111	70	26
Sex	Female	401	131	111
	Male	313	122	122
	Unknown	111	115	51
City	Blantyre	169	357	259
	Karonga	656	0	0
	Lilongwe	0	0	23
	Unknown	0	11	2
Sampling period		2009-2014	1997-2015	2000-2015

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

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

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

As illustrated in Fig 1 and S1 Table, of the abundant serotypes in the blood or CSF, serotypes 1 ($p = 1.96E-34$), 5 ($p = 3.97E-19$), and 12F ($p=5.29E-06$) had a significant presence among patients with a low frequency in the nasopharynx including serotype 1 (1%), serotype 5 (2%) and serotype 12F (0.6%). Given that the colonization phase is a pre-requisite for infection, serotypes 1, 5, and 12F may have a short period of nasopharyngeal colonization before infecting the sterile sites. Since serotypes 1 and 5 were also the most frequent in the entire cohort, they could be considered as the most prevalent serotypes with the highest invasiveness. In this study, the terms “significant invasive serotypes” or "hyper-invasive serotypes" were assigned to serotypes 1, 5, and 12F. Comparatively, abundant serotypes 16F and 19F in the nasopharynx had a significant presence among carriers suggesting that they have a lower potential to cause invasive disease. Other frequently detected serotypes, such as 6A, 6B, and 23F, were abundant and equally distributed among carriers and patients. They may need to colonize the nasopharynx longer than the hyper-invasive serotypes before infecting the sterile sites. Serotypes 6A, 6B, and 23F were previously reported as abundant and common serotypes between non-invasive carriers and invasive infection (37)(38).

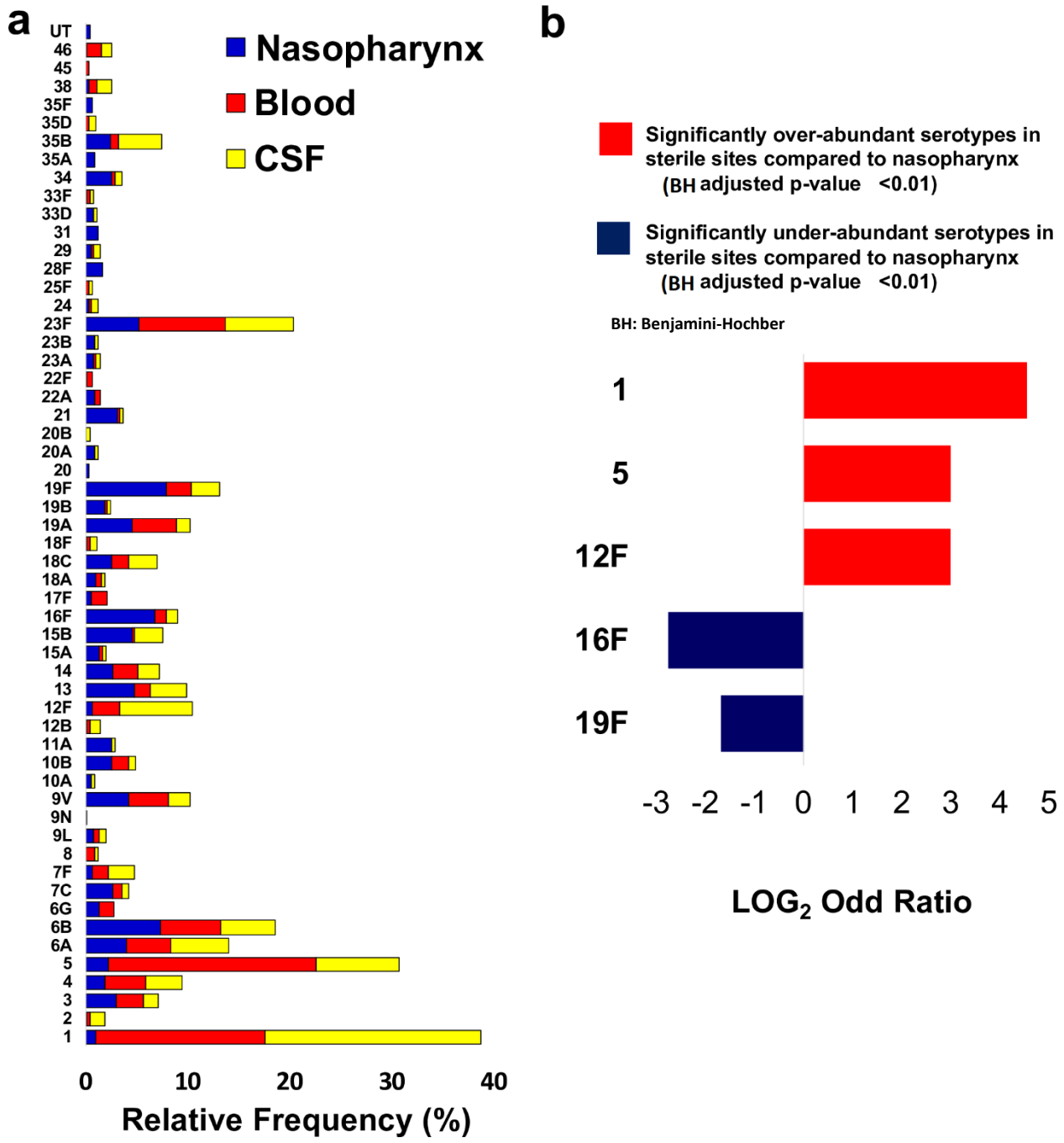

Fig 1. The distribution of the 56 pneumococcal serotypes assigned to 1477 samples from Malawi. (a) The relative frequency of each serotype in the nasopharynx of carriers, the blood of bacteremia patients, and the CSF of meningitis patients is shown in blue, red, and yellow, respectively (UT: Un-Typeable). (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).

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

High diversity in the pneumococcal pan-genome

For the genome assembly, the average optimized assembly hash value and the average N50 were 96 and 113,986, respectively. The mean assembled genome size was estimated at 2,116,779 nucleotides with a standard deviation of 106,481. This length is in the range of previously reported *S. pneumoniae* genome size (39). The length of the pan-genome was 5,178,167 base pairs, composed of 6,803 genes, including 729 core-(10.7%), 820 soft-core- (12.1%), and 5,254 accessory genes (77.2%). The pan-genome was open as it did not reach sampling saturation as the number of genes increased as the sample size grew (S2 Fig). In Fig 2, the pan-genome is demonstrated as a gene presence-absence heatmap that shows serotypes as the strongest determinant of the gene distribution in the pan-genome. The distinction was evident for the hyper-invasive serotypes 1, 5, and 12F since they clustered in unique clades. The sets of genes specifically present in different serotypes appear as small blue blocks in the heatmap.

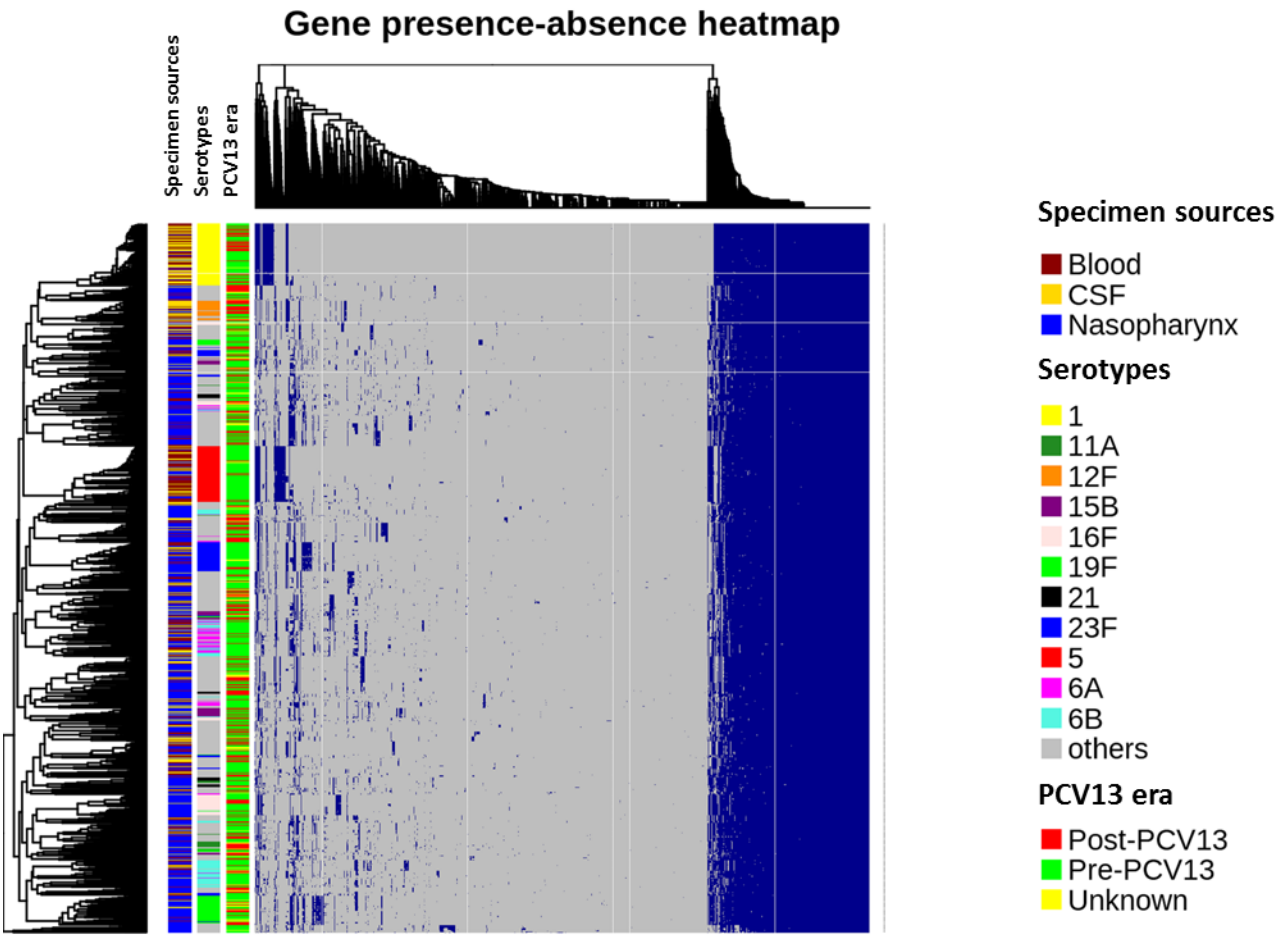

Fig 2. The pan-genome matrix of 1477 pneumococcal isolates from Malawi. The pan-genome is visualized as a gene presence-absence heatmap representing the hierarchical unsupervised clustering of samples based on the distribution of genes in the pan-genome. Each row is a sample, and each column is a gene. 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.

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

The maximum-likelihood tree characterized by the distribution of the small-scale variants (SNPs and indels) in the core-genome identified the hyper-invasive serotypes 1, 5, and 12F as monophyletic clusters (Fig 3). The distinction of hyper-invasive serotypes that were separated as unique clusters was

higher than other abundant serotypes, such as 6B, 19F, and 23F, which appeared as multiple clusters on the phylogenetic tree (Fig 3).

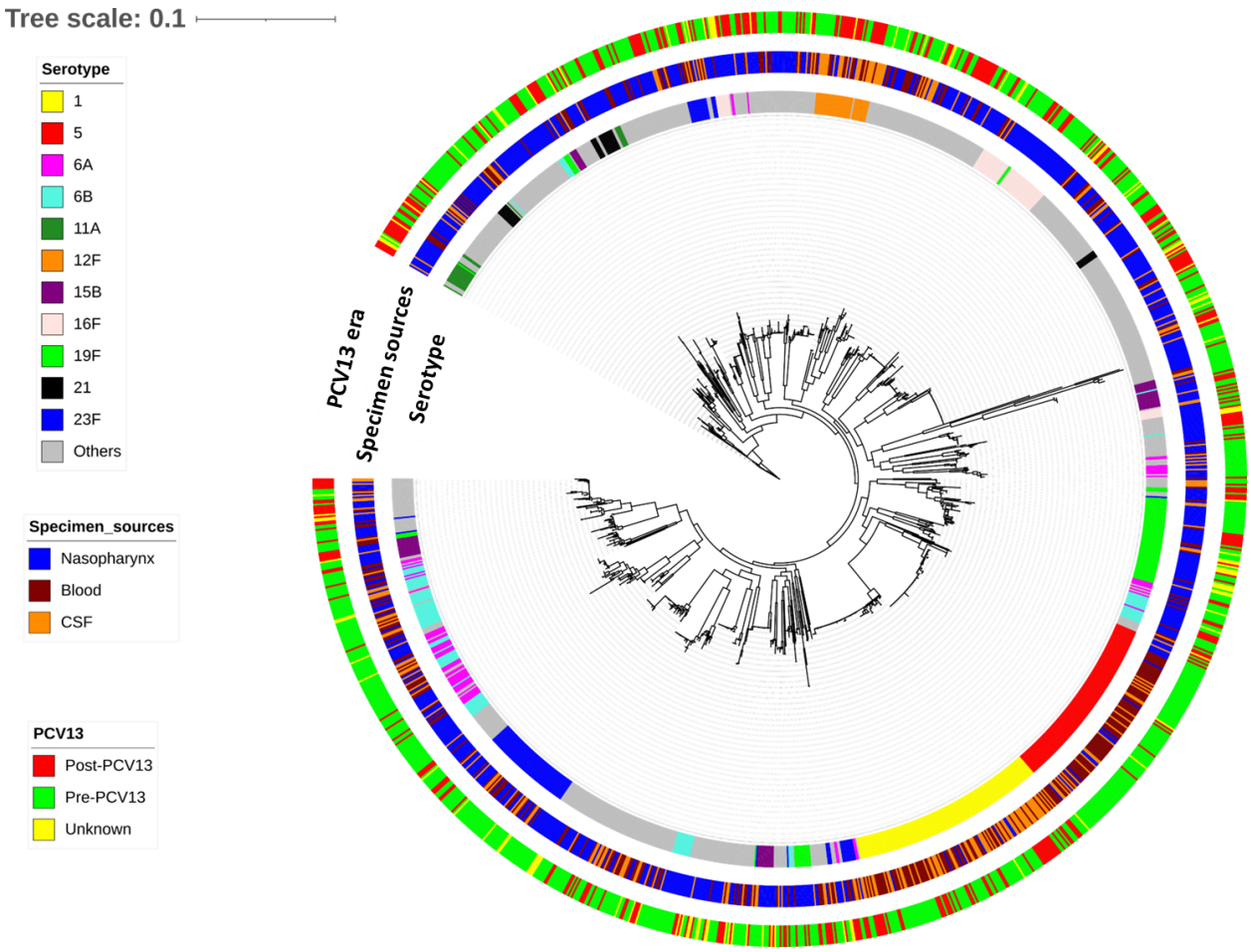

Fig 3. The phylogenetic population structure of 1477 pneumococcal samples from Malawi. The phylogenetic tree was built based on the multiple sequence alignment of the core genome using the maximum likelihood method. Colors on the loops show the serotypes, specimen sources (isolation sites), and PCV13 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.

The PCA of large-scale variants distribution in the accessory-genome (gene presence/absence) showed serotypes 1 and 5 clustered distantly from other strains (Fig 4). Meanwhile, a moderate level of separation was observed for serotypes 12F, 19F, and 23F. This highlighted the distinction of serotypes 1 and 5 in their gene content as observed in their SNPs profiles demonstrated by the phylogenetic tree.

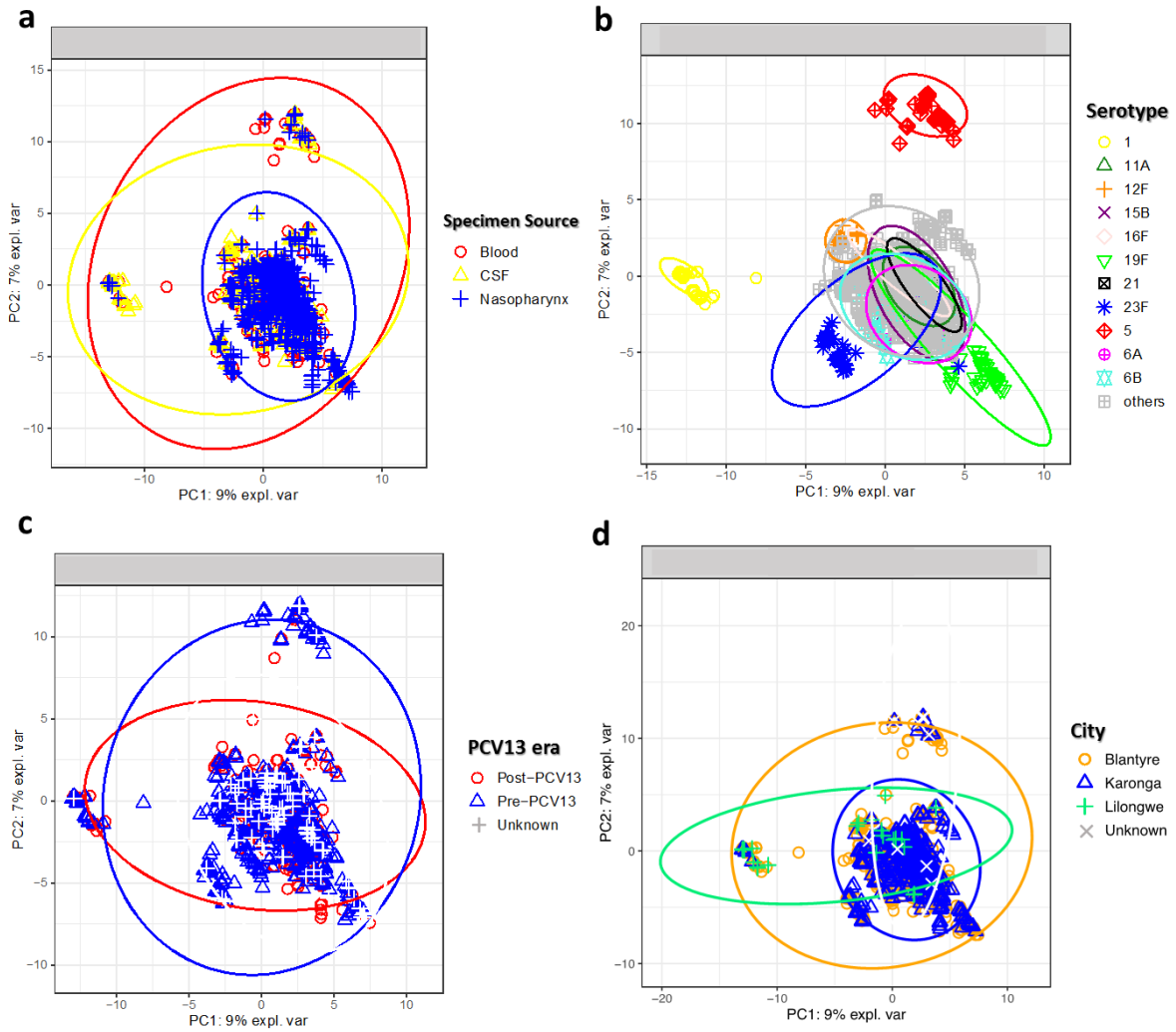

Fig 4. The PCA of the gene distribution in the pan-genome of pneumococcal isolates from 1477 Malawians. The PCA of variants (gene presence-absence) in the accessory-genome indicates the influence of (a) specimen sources (isolation sites), (b) serotypes, (c) PCV13 (vaccination) era, and (d) geographical 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

population likely contained invasive serotypes that could bias the test between carriage and disease samples to identify potential virulence genes.

c) Population structure: The significant abundance of the hyper-invasive serotypes in the patient group and their highest genetic distinction would skew the test between the carriage and patient groups. The difference between carriage and disease groups would actually be the difference between the carriage and hyper-invasive serotypes (and not all strains in the blood and CSF).

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

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

based analysis identified 27 significant genes including 11 genes significantly present in the blood and CSF and 16 genes in the nasopharynx (Table 2 and 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

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 (40).

Significant genes in the nasopharynx (absent from samples in the blood and CSF) were from RD10, which is known as SecY2A2 island and is responsible for the secretion of pneumococcal serine-rich repeat protein or PsrP (S4 Fig)(41). RD10 contains several glycosyltransferases and secretory components that were significantly absent from the genome of samples collected from blood and CSF.

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

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

There were 57 significantly absent genes from serotypes 1, 5, and 12F (Fig 5.a). The most significant ones were from RD8a, a region composed of two operons, RD8a1 (SP_1315-1324) and RD8a2 (SP_1325-SP_1331) (S5 Fig). Genes in RD8a1 include eight *ntp* genes that code for the multiunit *V-type proton/sodium ATP synthase* complex that produces ATP through oxidative phosphorylation in the presence of a Na^+ gradient across the membrane (42). RD8a2 includes *neuraminidase*, *N-acetylneuraminate lyase (nanA),* and *N-acetylmannosamine-6-phosphate epimerase (nanE).* These genes cleave carbohydrates from the glycoproteins on the surface of epithelial cells. Other genes in RD8a2 encode the Sodium/solute symporter subunits that use Na^+ gradient to import the carbohydrates (43). Symporter means the channel transports the solute (carbohydrates) and co-solute (Na^+) in the same direction using the energy stored in an inwardly directed sodium gradient. In summary, genes in RD8a operons work together to produce ATP, cleave carbohydrates from the host epithelial cells, and import them into the bacterial cell. RD8a was not present in any samples assigned to serotypes 1, 5, and 12F, whereas it was present in other abundant serotypes such as 6A, 6B, 16F, and 19F. The under-represented pathway (sp00190) in the hyper-invasive serotypes was linked to genes in RD8a. The lack of RD8a in the hyper-invasive serotypes may relate to their speed at entering the blood and CSF, where they do not have access to free oxygen molecules that are required for oxidative phosphorylation (44).

Other significant genes absent from the hyper-invasive serotypes were from RD4 and RD7. RD4 is composed of a group of sortase enzymes that catalyze the assembly of pilins into pili and the anchoring of pili and other surface proteins to the cell wall (45)(46). The pilus is a hair-like structure associated with bacterial adhesion and colonization (47). Due to the possible short colonization period of the hyper-invasive serotypes, they may not benefit from genes in RD4 that assemble the pilus. Genes in RD7 have remained uncharacterized to date.

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

conserved in serotype 5. RD10 was also conserved in 100% of serotypes 16F and 19F. The summary of the gene present-absent analysis is illustrated in Fig 5.b.

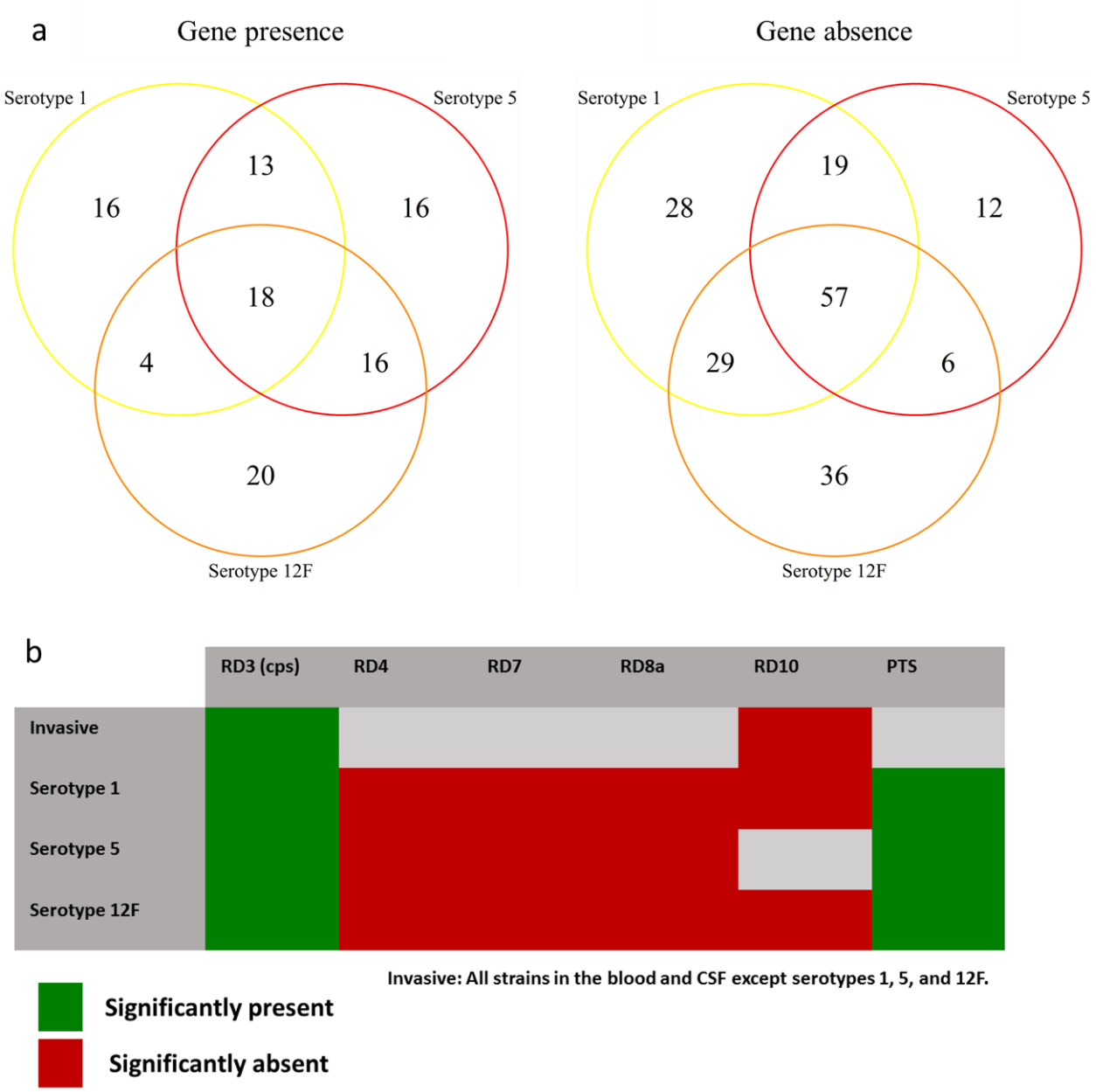

Fig 5. The summary of the gene presence-absence analysis. (a) The number of significant genes present-absent in the hyper-invasive serotypes. The gene presence-absence analysis was applied using Scoary to compare the gene pools of the hyper-invasive serotypes and serotypes 16F and 19F. P-values were 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.

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

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

Discussion

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

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

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

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

The second possible scenario relates to the differential gene expression pattern of shared genes in the ubiquitous serotypes. Although the type-specific *cps* genes were identified in the isolates of both nasopharynx and sterile sites, the expression pattern of these genes could vary within each strain, which would contribute to the invasiveness of some strains. Several studies described a cycle of encapsulation and un-encapsulation amongst pneumococcal strains. Isolates benefit from mutations in the *cps* locus to either cease or re-start the capsule expression (48). The lack of a capsule at the epithelial surface enables the bacterium to expose its surface proteins on the cell wall underneath the capsule and promote adherence to the host epithelial cells. It has been estimated that 15% of isolates in the upper respiratory tract are unencapsulated and adhere to the respiratory epithelial cells more efficiently than encapsulated isolates (49)(50). Lack of capsule also facilitates acquiring virulence and resistance genes from other isolates. The thick capsule prevents immunoglobulins from interacting with the pathogen surface proteins during disease. Meanwhile, the negatively charged CPS interferes with the function of the host phagocytes (51)(52). Taken together, the presence of the *cps* locus in the genome of isolates assigned to the same serotype does not necessarily reflect the encapsulation of all cells.

Serotypes 1 and 5 can infect all age groups and cause severe IPDs (53). Serotype 1 is genetically distinct between different geographical regions (54) and is known as the leading cause of pneumococcal meningitis in Africa (55)(56). Our findings support previous research showing that serotypes 1, 5, and 12F are the major cause of IPDs in Malawi (57). Serotype 1 was persistently dominant in pre- and post-PCV13 eras, serotype 5 was only predominant in pre-PCV3, and serotype 12F emerged after vaccination. The study also characterized the high genetic distinction of the hyper-invasive serotypes in Malawi by identifying significant genes that are present or absent in their genome structure compared to other serotypes. Many of the significant genes present in the nasopharynx and sterile sites were homologous and had the same function, and there were significant genes with an unknown function (hypothetical proteins). However, of greatest interest, we did find several significant genes with specific functions that could explain the difference between the biology of the hyper-invasive serotypes and nasopharyngeal samples.

Genes from RD8a were not present in the genome of serotypes 1, 5, and 12F. The presence of this region has been reported to be associated with the virulence of serotypes 6B and 14 in the United States (58). We found this region conserved in serotypes 13, 14, 16F, and 19F in Malawi. One may hypothesize that RD8a may be required by serotypes that colonize the nasopharynx longer than serotypes 1, 5, and 12F. The functions of genes in RD8a strengthen the assumption to some extent that RD8a may be essential for long nasopharyngeal colonization. RD8a harbors genes such as *neuraminidase*, *nanA,* and *nanE* that cleave terminal sialic acid residues from mucoglycans and epithelial glycoconjugates and assist the pathogen in penetrating the mucus layer and adhering to the epithelial cells in the nasopharynx. In the meantime, because free carbohydrates are scarce in the upper respiratory tract (59), cleaved sialic acid can be used as the carbon source for metabolism. During colonization, secretion of the pneumococcal toxins elevates the level of sodium ions (Na^+) in the nasopharynx (60), the sodium-solute symporter in RD8a imports a wide variety of substrates with the sodium ions into the cell (61).

Most importantly, the *ntp* gene cluster in RD8a encodes the *V-type sodium ATP synthase* that pumps the extra sodium ions out of the cell (42) and uses the sodium-motive force for oxidative phosphorylation and ATP synthesis (62). Oxidative phosphorylation is the final step of aerobic respiration that requires free oxygen molecules for ATP synthesis. Pneumococci are facultative anaerobes that can either perform aerobic or anaerobic respiration with or without oxygen. In the upper respiratory tract, they access atmospheric oxygen molecules that can be used by *ntp* genes to perform oxidative phosphorylation. However, genes in RD8a may not be beneficial for hyper-invasive serotypes that supposedly do not stay in the nasopharynx for long. Thus, the level of aerobic ATP synthesis is presumably higher in serotypes 13, 14, 16F, and 19F in contrast with serotypes 1, 5, and 12F, which lack RD8a.

Pneumococci can ferment up to 30 types of carbohydrates, imported mainly by two types of membrane transporters, including ATP-binding cassette (ABC) transporters and PTS transporters (63). The major differences between ABC and PTS transporters are: (i) ABC transporters use energy from ATP, but PTS transporters use energy from phosphoenolpyruvate, and (ii) ABC transporters do not modify the imported substrate, but PTS transporters phosphorylate the incoming sugar upon transport. Generally, ABC transporters require more energy than PTS transporters, albeit they can transport longer and more complicated carbohydrates (64). Unlike isolates in the nasopharynx, serotypes 1, 5, and 12F have access to more simple and free host dietary carbohydrates in the blood and the central nervous system. Due to a potential lower ATP synthesis level in serotypes 1, 5, and 12F (due to the lack of RD8a), they may prefer to use PTS transporters to uptake sugars such as fructose and lactose, and that is why genes that encode PTS transporters are significantly more present in the hyper-invasive serotypes. In addition to the sugar uptake, PTS transporters regulate several pathways in bacteria, such as gene expression and communication between cells. Thus, the phenotypic effects of the PTS transporters should not be limited just to their ability to import carbohydrates (65).

Genes from RD10 were absent from serotypes 1 and 12F. However, they were conserved in serotypes 5, 16F, and 19F. Moreover, the location-based gene presence-absence analysis showed that RD10 was significantly present in the nasopharyngeal samples compared to samples from sterile sites. These genes have homology to the general secretion pathway protein B sceA2/Y2 system components in *Streptococcus gordonii* that contribute to the secretion of the general secretion pathway protein B involved in infective endocarditis (66). In the *S. pneumoniae* genome, the homolog of general secretion pathway protein B is the PsrP that is transported to the bacterial cell surface by the SecA2/Y2 system encoded by genes in RD10. Studies conducted on *Streptococcus gordonii* showed that the presence of SecA2/Y2 promotes the adhesion to epithelial cells in the nasopharynx and erythrocytes in the blood (67)(68), which is why SecA/Y2 is significantly present in nasopharyngeal samples and serotype 5 (abundant in the blood). The presence of the secA2/Y2-like component should also facilitate the export of pneumolysin that enhances adhesion to the host cell and survival in the blood (69)(70).

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

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.

References

1. Swarthout TD, Fronterre C, Lourenço J, Obolski U, Gori A, Bar-Zeev N, et al. High residual carriage of vaccine-serotype Streptococcus pneumoniae after introduction of pneumococcal conjugate

vaccine in Malawi. Nature Communications 2020 11:1 [Internet]. 2020 May 6 [cited 2022 Aug 8];11(1):1–12. Available from: https://www.nature.com/articles/s41467-020-15786-9

2. Kamthunzi P. Impact of PCV13 vaccination in Blantyre, Malawi. Lancet Glob Health [Internet]. 2021 Jul 1 [cited 2022 Aug 8];9(7):e893–4. Available from: http://www.thelancet.com/article/S2214109X21002588/fulltext

3. Paixão L, Oliveira J, Veríssimo A, Vinga S, Lourenço EC, Ventura MR, et al. Host glycan sugar-specific pathways in streptococcus pneumonia: Galactose as a key sugar in colonisation and infection. PLoS One. 2015;

4. Bogaert D, de Groot R, Hermans PWM. Streptococcus pneumoniae colonisation: The key to pneumococcal disease. Lancet Infectious Diseases. 2004.

5. Backhaus E, Berg S, Andersson R, Ockborn G, Malmström P, Dahl M, et al. Epidemiology of invasive pneumococcal infections: Manifestations, incidence and case fatality rate correlated to age, gender and risk factors. BMC Infect Dis. 2016;

6. Kadioglu A, Weiser JN, Paton JC, Andrew PW. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. Nature Reviews Microbiology. 2008.

7. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: Transmission, colonization and invasion. Nature Reviews Microbiology. 2018.

8. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, et al. Pneumococcal capsules and their types: Past, present, and future. Clin Microbiol Rev. 2015;

9. Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. Infect Immun. 2007;

10. Brown J, Hammerschmidt S, Orihuela C. Streptococcus Pneumoniae: Molecular Mechanisms of Host-Pathogen Interactions. Streptococcus Pneumoniae: Molecular Mechanisms of Host-Pathogen Interactions. 2015.

11. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, et al. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large cps fragment from an oral streptococcus. mBio. 2020;

12. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. The Lancet. 2011.

13. Feikin DR, Kagucia EW, Loo JD, Link-Gelles R, Puhan MA, Cherian T, et al. Serotype-Specific Changes in Invasive Pneumococcal Disease after Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites. PLoS Med. 2013;

14. McCollum ED, Nambiar B, Deula R, Zadutsa B, Bondo A, King C, et al. Impact of the 13-valent pneumococcal conjugate vaccine on clinical and hypoxemic childhood pneumonia over three years in central Malawi: An observational study. PLoS One. 2017;

15. Bar-Zeev N, Swarthout TD, Everett DB, Alaerts M, Msefula J, Brown C, et al. Impact and Effectiveness of 13-Valent Pneumococcal Conjugate Vaccine on Population Incidence of Vaccine and Non-Vaccine Serotype Invasive Pneumococcal Disease in Blantyre, Malawi, 2006-2018: Prospective Observational Time-Series and Case-Control Studies. SSRN Electronic Journal [Internet]. 2020 Dec 31 [cited 2022 Jul 9]; Available from: https://papers.ssrn.com/abstract=3745169

16. Heinsbroek E, Tafatatha T, Phiri A, Swarthout TD, Alaerts M, Crampin AC, et al. Pneumococcal carriage in households in Karonga District, Malawi, before and after introduction of 13-valent pneumococcal conjugate vaccination. Vaccine. 2018;

17. Cillóniz C, Garcia-Vidal C, Ceccato A, Torres A. Antimicrobial Resistance Among Streptococcus pneumoniae. In: Antimicrobial Resistance in the 21st Century. Springer; 2018. p. 13–38.

18. Iranzadeh A, Mulder NJ. Bacterial Pan-Genomics. In: Microbial Genomics in Sustainable Agroecosystems. Springer; 2019. p. 21–38.

19. Embry A, Hinojosa E, Orihuela CJ. Regions of Diversity 8, 9 and 13 contribute to Streptococcus pneumoniae virulence. BMC Microbiol. 2007;

20. Brückner R, Nuhn M, Reichmann P, Weber B, Hakenbeck R. Mosaic genes and mosaic chromosomes-genomic variation in Streptococcus pneumoniae. International Journal of Medical Microbiology. 2004.

21. Crampin AC, Dube A, Mboma S, Price A, Chihana M, Jahn A, et al. Profile: The Karonga health and demographic surveillance system. Int J Epidemiol. 2012;

22. Andrews S. FastQC: A quality control tool for high throughput sequence data. Babraham Bioinformatics [Internet]. 2010;http://www.bioinformatics.babraham.ac.uk/projects/. Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:FastQC+a+quality+control+tool+for+high+throughput+sequence+data.#0

23. Epping L, van Tonder AJ, Gladstone RA, Bentley SD, Page AJ, Keane JA. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. Microb Genom [Internet]. 2018 Jul 1 [cited 2022 Jul 11];4(7). Available from: https://pubmed.ncbi.nlm.nih.gov/29870330/

24. Gladman S, Seemann T. VelvetOptimiser. Free Software Foundation. 2008.

25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;

26. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

27. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 May 25;31(22):3691–3.

28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;

29. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol [Internet]. 2020 May 1 [cited 2021 Dec 6];37(5):1530–4. Available from: https://academic.oup.com/mbe/article/37/5/1530/5721363

30. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;

31. Gaujoux R. An introduction to NMF package. BMC Bioinformatics. 2010;

32. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ‘omics feature selection and multiple data integration. PLoS Comput Biol. 2017;

33. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol [Internet]. 2016;17(1):238. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1108-8

34. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;

35. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Vol. 27, Nucleic Acids Research. 1999. p. 29–34.

36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res [Internet]. 2004 Jan 1 [cited 2021 Dec 6];32(suppl_1):D138–41. Available from: https://academic.oup.com/nar/article/32/suppl_1/D138/2505358

37. Wang Q, Shi W, Li Y, Gao W, Yuan L, Dong F, et al. Serotype distribution of Streptococcus pneumoniae isolated from children hospitalized in Beijing children's hospital (2013–2019). Vaccine. 2020;38(49):7858–64.

38. Alizadeh Chamkhaleh M, Esteghamati A, Sayyahfar S, Gandomi-Mohammadabadi A, Balasi J, Abdiaei H, et al. Serotype distribution of Streptococcus pneumoniae among healthy carriers and clinical patients: a systematic review from Iran. European Journal of Clinical Microbiology and Infectious Diseases. 2020.

39. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, et al. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: Insights into the pneumococcal supragenome. J Bacteriol [Internet]. 2007 Nov [cited 2022 Dec 24];189(22):8186–95. Available from: https://journals.asm.org/doi/10.1128/JB.00690-07

40. Zhang S, Meyer R. The relaxosome protein MobC promotes conjugal plasmid mobilization by extending DNA strand separation to the nick site at the origin of transfer. Mol Microbiol. 1997;

41. Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, Tuomanen EI, et al. Identification of a candidate Streptococcus pneumoniae core genome and regions of diversity correlated with invasive pneumococcal disease. Infect Immun. 2006;74(8):4766–77.

42. Yokoyama K, Imamura H. Rotation, structure, and classification of prokaryotic V-ATPase. Journal of Bioenergetics and Biomembranes. 2005.

43. Reizer J, Reizer A, Saier MH. A functional superfamily of sodium/solute symporters. Biochim Biophys Acta [Internet]. 1994 Jun 29 [cited 2022 Jan 11];1197(2):133–66. Available from: https://pubmed.ncbi.nlm.nih.gov/8031825/

44. Gaudu P. Aerobic respiration metabolism in lactic acid bacteria and uses in biotechnology. [cited 2022 Aug 21]; Available from: www.annualreviews.org

45. Manzano C, Contreras-Martel C, el Mortaji L, Izoré T, Fenel D, Vernet T, et al. Sortase-Mediated Pilus Fiber Biogenesis in Streptococcus pneumoniae. Structure. 2008;

46. Lemieux J, Woody S, Camilli A. Roles of the sortases of Streptococcus pneumoniae in assembly of the RlrA pilus. J Bacteriol. 2008;

47. Williams L, Stapleton F, Carnt N. Microbiology, lens care and maintenance. Contact Lenses. 2019 Jan 1;65–96.

48. Langereis JD, de Jonge MI. Non-encapsulated Streptococcus pneumoniae , vaccination as a measure to interfere with horizontal gene transfer . Virulence. 2017;

49. van der Windt D, Bootsma HJ, Burghout P, van der Gaast-de Jongh CE, Hermans PWM, van der Flier M. Nonencapsulated Streptococcus pneumoniae resists extracellular human neutrophil elastase- and cathepsin G-mediated killing. FEMS Immunol Med Microbiol. 2012;

50. Talbot UM, Paton AW, Paton JC. Uptake of Streptococcus pneumoniae by respiratory epithelial cells. Infect Immun. 1996;

51. Lee CJ, Banks SD, Li JP. Virulence, immunity, and vaccine related to streptococcus pneumoniae. Crit Rev Microbiol. 1991;

52. Keller LE, Robinson DA, McDaniel LS. Nonencapsulated Streptococcus pneumoniae : Emergence and Pathogenesis . mBio. 2016;

53. Hausdorff WP. The roles of pneumococcal serotypes 1 and 5 in paediatric invasive disease. Vaccine. 2007;

54. Cornick JE, Chaguza C, Harris SR, Yalcin F, Senghore M, Kiran AM, et al. Region-specific diversification of the highly virulent serotype 1 Streptococcus pneumoniae. Microb Genom. 2015;1(2).

55. Leimkugel J, Adams Forgor A, Gagneux S, Pflüger V, Flierl C, Awine E, et al. An outbreak of serotype 1 Streptococcus pneumoniae meningitis in northern Ghana with features that are characteristic of Neisseria meningitidis meningitis epidemics. J Infect Dis. 2005;192(2):192–9.

56. Gessner BD, Mueller JE, Yaro S. African meningitis belt pneumococcal disease epidemiology indicates a need for an effective serotype 1 containing vaccine, including for older children and adults. BMC Infect Dis. 2010;

57. Cornick JE, Everett DB, Broughton C, Denis BB, Banda DL, Carrol ED, et al. Invasive streptococcus pneumoniae in children, Malawi, 2004-2006. Emerg Infect Dis. 2011;

58. Obert C, Sublett J, Kaushal D, Hinojosa E, Barton T, Tuomanen EI, et al. Identification of a Candidate Streptococcus pneumoniae core genome and regions of diversity correlated with

invasive pneumococcal disease. Infect Immun [Internet]. 2006 Aug [cited 2022 Jan 15];74(8):4766–77. Available from: https://pubmed.ncbi.nlm.nih.gov/16861665/

59. Walters DM, Stirewalt VL, Melville SB. Cloning, sequence, and transcriptional regulation of the operon encoding a putative N-acetylmannosamine-6-phosphate epimerase (nanE) and sialic acid lyase (nanA) in Clostridium perfringens. J Bacteriol. 1999;

60. Bakeeva LE, Chumakov KM, Drachev AL, Metlina AL, Skulachev VP. The sodium cycle. III. Vibrio alginolyticus resembles Vibrio cholerae and some other vibriones by flagellar motor and ribosomal 5S-RNA structures. BBA - Bioenergetics. 1986;

61. Saier MH. Families of transmembrane sugar transport proteins. Molecular Microbiology. 2000.

62. Boyer PD. THE ATP SYNTHASE—A SPLENDID MOLECULAR MACHINE. Annu Rev Biochem. 2002;

63. Bidossi A, Mulas L, Decorosi F, Colomba L, Ricci S, Pozzi G, et al. A functional genomics approach to establish the complement of carbohydrate transporters in Streptococcus pneumoniae. PLoS One. 2012;

64. Buckwalter CM, King SJ. Pneumococcal carbohydrate transport: Food for thought. Trends in Microbiology. 2012.

65. Saier MH. The Bacterial Phosphotransferase System: New Frontiers 50 Years after Its Discovery. Journal of Molecular Microbiology and Biotechnology. 2015.

66. Bensing BA, Gibson BW, Sullam PM. The Streptococcus gordonii Platelet Binding Protein GspB Undergoes Glycosylation Independently of Export. J Bacteriol. 2004;

67. Bensing BA, Sullam PM. Transport of preproteins by the accessory Sec system requires a specific domain adjacent to the signal peptide. J Bacteriol. 2010;

68. Yamaguchi M, Terao Y, Mori-Yamaguchi Y, Domon H, Sakaue Y, Yagi T, et al. Streptococcus pneumoniae Invades Erythrocytes and Utilizes Them to Evade Human Innate Immunity. PLoS One. 2013;

69. Bandara M, Skehel JM, Kadioglu A, Collinson I, Nobbs AH, Blocker AJ, et al. The accessory Sec system (SecY2A2) in Streptococcus pneumoniae is involved in export of pneumolysin toxin, adhesion and biofilm formation. Microbes Infect. 2017;

70. Wu R, Wu H. A molecular chaperone mediates a two-protein enzyme complex and glycosylation of serine-rich streptococcal adhesins. Journal of Biological Chemistry. 2011;

Supporting information

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 downsampling, 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 serotypes 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.

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

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

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

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

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

Figure captions

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).

Fig 2. The pan-genome matrix of 1477 pneumococcal isolates from Malawi. The pan-genome is visualized as a gene presence-absence heatmap that represents 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.

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.

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 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.