Early signals of vaccine driven perturbation seen in pneumococcal carriage population genomic data

Chrispin CHAGUZA[1,2,3,*], Ellen HEINSBROEK[2,4], Rebecca A. GLADSTONE[1], Terence TAFATATHA[5], Maaike ALAERTS[3,6], Chikondi PENO[3,7], Jennifer E. Cornick[2,3], Patrick MUSICHA[2,3,8,9], Naor BAR-ZEEV[2,3,10], Arox KAMNG'ONA[2,3,11], Aras KADIOGLU[2], Lesley MCGEE[12], William P. HANAGE[13], Robert F. BREIMAN[14], Robert S. HEYDERMAN[3,15], Neil FRENCH[2,3,¶], Dean B. EVERETT[3,7,¶], and Stephen D. BENTLEY[1,16,¶,*]

[1]Parasites and Microbes, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK
[2]Department of Clinical Infection, Microbiology and Immunology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK
[3]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi
[4]HIV & STI Department, National Infection Service, Public Health England, London, UK
[5]Malawi Epidemiology Intervention Research Unit (formerly KPS), Chilumba, Malawi
[6]Center of Medical Genetics, University of Antwerp, Antwerp, Belgium
[7]MRC Centre for Inflammation Research, Queens Medical Research Institute, University of Edinburgh, Edinburgh, UK
[8]Mahidol Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok, Thailand
[9]Nuffield Department of Medicine, University of Oxford, Oxford, UK
[10]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA
[11]Department of Biomedical Sciences, University of Malawi, College of Medicine, Blantyre, Malawi
[12]Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, USA
[13]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
[14]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, USA
[15]Division of Infection and Immunity, University College London, London, UK
[16]Department of Pathology, University of Cambridge, Cambridge, UK

[*]Corresponding author: C.C. (cc19@sanger.ac.uk) and S.D.B. (sdb@sanger.ac.uk).
[¶]N.F., D.B.E. and S.D.B. contributed equally to this work.

Abstract

Pneumococcal conjugate vaccines (PCV) have reduced pneumococcal diseases globally. Despite this, much remains to be learned about their effect on pathogen population structure. Here we undertook whole genome sequencing of 660 pneumococcal strains from asymptomatic carriers to investigate population restructuring in pneumococcal strains sampled before and after PCV13 introduction in a previously vaccine-naïve setting. We show substantial decreasing frequency of vaccine-type (VT) strains and their strain diversity post-vaccination in the vaccinated but not unvaccinated age groups indicative of direct but limited or delayed indirect effect of vaccination. Clearance of identical VT serotypes associated with multiple lineages occurred regardless of their genetic background. Interestingly, despite the increasing frequency of non-vaccine type (NVT) strains through serotype replacement, the serotype diversity was not fully restored to the levels observed prior to vaccination implying limited serotype replacement. The frequency of antibiotic resistant strains was low and remained largely unchanged post-vaccination but intermediate-penicillin-resistant lineages were reduced in the post vaccine population. Significant perturbations marked by changing frequency of accessory genes associated with diverse functions especially mobile genetic elements and bacteriocin activity were detected. This phylogenomic analysis demonstrates early vaccine-induced pneumococcal population restructuring not only at serotype but also accessory genome level.

Author summary

Different formulations of PCVs have been effective in reducing the invasive pneumococcal disease burden globally. Clinical trials have started to indicate high impact and effectiveness of PCV13 in Sub Saharan Africa (SSA) but there is limited understanding of how the introduction of PCVs alters the population structure of pneumococcal strains at serotype and genomic level. Here we investigated this using pneumococcal strains sampled pre- and post-PCV13 introduction from a previously vaccine naïve setting in Northern Malawi. Our findings reveal decrease in frequency of VT serotypes and their associated lineages in the largely vaccinated under-five population but not older individuals indicating a direct but limited or delayed indirect protection. The diversity of serotypes also decreased post-vaccination in VT strains in the under-fives but there was no change in NVT strains suggesting incomplete serotype replacement. At the genomic level, logistic regression revealed changing frequency of accessory genes largely associated with mobile genetic elements but

such changes did not include any antibiotic resistance genes. These findings show significant perturbations at serotype and accessory genome level in carried pneumococcal population after two years from PCV13 introduction but the pneumococcal population was still perturbed and had not returned to a new equilibrium state.

Introduction

Pneumococcal polysaccharide antigens covalently attached to carrier proteins elicit sufficient serotype-specific antibody responses against *Streptococcus pneumoniae* (the pneumococcus) and form the basis of pneumococcal conjugate vaccines (PCV). Different formulations of PCVs have been licensed and introduced globally, and clinical case-control, cohort and surveillance studies have documented high effectiveness of these vaccines on non-invasive [1] and invasive pneumococcal disease (IPD) [2]. For example, after introduction of PCV7, 69% and 57% reductions in IPD were observed in USA and UK respectively whereby vaccine type (VT) strains decreased by >90% [2, 3]. Few African countries introduced PCV7 because its projected low coverage of VTs as highly common invasive serotypes in this setting (e.g. serotypes 1 and 5) were not covered [4]. Regardless, PCV7 caused >85% reduction in IPD in South Africa, crucially in HIV-infected individuals [5]. Consistent with findings in high-income countries [6], higher-valent PCVs appears to be highly effective in reducing VT serotypes in carriage (>65%) and IPD (>80%) in Africa[7–9].

In addition to reducing IPD [2, 8, 9], PCV has an added benefit of reducing VT carriage [10]. However, the impact on overall carriage rate and density is not substantial [11, 12]. This is because PCV introduction induces serotype replacement whereby reduction of VTs substantial alters serotype competition dynamics thereby prompting an increase of non-vaccine type (NVT) strains uncommon prior to vaccination [13]. A well-known example of replacement is the upsurge of serotype 19A post-PCV7 introduction [14]. While invasive potential of replacement NVT serotypes is typically lower than for VTs, this is not universally true, and some strains retain propensity for invasive disease [15]. Therefore, understanding post-vaccination dynamics in pneumococcal population is crucial to inform future clinical interventions and it may be more informative to study the carriage population where evolution is ongoing, rather than studying isolates from disease which is considered to be an evolutionary dead-end.

In this study, we undertook whole genome sequencing (WGS) of 660 pneumococcal strains sampled before and after nation-wide introduction of PCV13 vaccine in a previously vaccine-naïve setting of Northern Malawi to investigate early changes in population structure, serotype composition and diversity, accessory genome variation. PCV13 was introduced in November 2011 in Malawi was via an accelerated '3+0' schedule (6, 10 and 14 weeks) with a limited catch-up for infants in the first year from introduction [16]. WGS of the isolates was undertaken by the Global Pneumococcal Sequencing (GPS) project (www.pneumogen.net), which to date has sequenced ≈23,000 pneumococcal strains sampled globally to map out post-vaccination evolution patterns of strains to inform future vaccine design. Here we focus on carriage samples from northern Malawi and present evidence of early vaccine-induced population restructuring at serotype, lineage and accessory genome level.

Results

Defining the population structure

Genomes were sequenced for 660 isolates collected from healthy carriers before and after introduction of PCV13 in a previously vaccine-naïve population in the Karonga district of northern Malawi (Fig 1a, S1 Data). A total of 45 serotypes and 169 sequence types (STs) were detected with evidence of serotype-switching within lineage due to recombination [17]. From a reference-free 1,050,021bp multiple sequence alignment of the 660 study genomes we identified 88,961 single nucleotide polymorphisms (SNP) which were used to infer their genetic population structure using an unsupervised hierarchical clustering algorithm [18] (S1-5 Fig). This approach identified twenty-three genomic clusters (GC) or lineages with one GC (GC23) appearing to be polyphyletic (Fig 1b). The GCs exhibited different within-lineage sequence diversity due to variations in recombination and age of the lineage (S6 Fig). We then used this defined population to investigate changes in frequency of serotypes, lineages and accessory genes post-PCV13 introduction while controlling for serotype category and sampling period.

Decrease of VT serotypes and their GCs signify PCV impact

Changes in frequency of GCs and their constituent serotypes before and after vaccination signaled substantial restructuring of the pneumococcal carriage population two years after PCV introduction. Using the Fisher's Exact test, three GCs namely

GC3 (P=0.002), GC16 (P=0.004) and GC17 (P=0.004) showed a post-vaccination increase in frequency but there was a decrease of GC10 (P=0.016) and GC19 (P=0.026) (Fig 2a,b, S1 Table). The reduced frequency of GCs, which were highly associated with VTs, largely reflected reduction of VTs in vaccinated children under five years old; a reduction that was not seen in the over five year old unvaccinated population (Fig 2c,d, S2 Table). We modelled the relationship between frequency of VT and NVT serotypes pre- and post-vaccination using linear regression and there was a smaller regression coefficient for VTs than NVT serotypes (Fig 2c,d). This showed that PCV13 reduced frequency of VT at a higher rate than NVT serotypes in the GCs. The majority of the GCs showed no change in odds ratio of VT serotypes in under-fives relative to over-fives while only four GCs (GC4,16,19,22) showed significant changes post-vaccination (S7 Fig, S3 Table). The decrease in the overall frequency of VT serotypes was evident (P=4.80 $\times 10^{-8}$, Fisher's Exact test) in strains sampled from under-fives but not over-fives (P=0.3739, Fisher's Exact test) (Fig 2e-g). The frequency of VT strains was also higher in strains from under-fives than over-fives (P=8.44$\times 10^{-4}$, Fisher's Exact test) but following vaccination their frequency became equal (33%) in both age groups (P=1, Fisher's Exact test) (Fig 2g, S4,5 Table). Similar tests revealed significant increase of four NVTs serotype 7C (P=0.001), 15B/C (P=0.004), 23A (P=0.017) and 28F (P=0.0001) after vaccination in under-fives while only 28F (P=0.029) increased in over-fives. By using a unique collection of isolates from both children and adults pre- and post-vaccination, these findings demonstrate substantial direct effects of PCV in children but either limited or incomplete indirect protection against carriage of VT strains in older individuals.

Emergence and clonal expansion of NVT strains

The overall frequency of serotypes and their dynamics within GCs revealed clonal expansion and potential emergence through capsule switching (Fig 4). With the exception of serotype 28F, all serotypes were detected prior to vaccination suggesting that emergence of previously undetected serotypes following vaccination was uncommon. Therefore, clonal expansion of extant serotypes uncommon prior to vaccination rather than capsule-switching drove the increased frequency of NVT serotypes post-vaccination. The majority of the capsule-switches occurred pre-vaccination but the specific times when those events occurred could not be established due to short sampling frame, which could bias temporal phylogenetic signal [2 years] (Fig 3, S8 Fig, S3 Table). Six capsule-switch events were detected based on phylogenetic clustering and ST profiles namely serotype 11A to 20 in GC7, 13 to 19A in GC9, 16F

to 19F in GC13, 9V to 28F in GC21, and 7C to NT in GC23 (S8 Fig). The capsule-switched strains circulated at low frequency (<1%) both pre- and post-vaccination ruling them out as a driver for the NVT serotype replacement. Of these capsule-switched serotypes, only serotype 28F, which was not detected prior to vaccination in carriage and previously studied invasive datasets [19], underwent a significant clonal expansion post-vaccination (Fig 2e,f and Fig 3). Of the two serotype 28F lineages, the increase of serotype 28F strains was due to clonal expansion of strains in NVT GC2 rather than the serotype 9V to 28F vaccine-escape capsule-switched variants in GC21 (Fig 2e,f). No serotype 28F strains were detected pre-vaccination including in previous IPD datasets, which suggested either circulation at undetectable levels pre-vaccination or importation from other countries.

Additional strains with similar ST profiles were searched in the global collection of ≈20,000 strains with similar ST profiles to investigate whether there was importation of serotype 28F strains from other countries. Only two strains were identified from South Africa, a serotype 28F and 9V clustering with strains in GC2 and GC21 respectively both identified only post-13 introduction in South Africa (2013 and 2014). The closest matching strains from our setting were distinguished from the South African strains by 1,966 and 2,284 SNPs in GC2 and GC21, which implied no recent importation post-vaccination (Fig 4a,b). Furthermore, the maximum sequence divergence between serotype 28F strains in GC2 was 6,757 SNPs, which is a further contradiction of the hypothesis that serotype 28F was imported post-vaccination because recently imported clones typically exhibit less diversity caused by transmission bottleneck. With such a short time frame from PCV introduction (2 years), this would have been insufficient for the imported strains to accrue such high genetic diversity considering that the pneumococcal mutation rate is ≈4 SNPs per year [20]. Furthermore, although the capsule-switched 28F strains in GC21 clustered together with the serotype 9V strains, these serotypes had different STs, which suggests that the capsule-switch event occurred earlier prior to vaccination. Therefore, these findings demonstrate that newly emerging serotypes post-vaccination were not due to importation of novel serotypes from other countries but rather expansion of extant serotypes circulating at undetectable levels prior to vaccination.

Reduction of Simpson diversity index indicates PCV impact

Reduction in Simpson diversity among VT serotypes post-vaccination occurred in under-fives (P=0.022, resampling) further supporting direct PCV effect (Fig 5a, S6

Table). Reduction in Simpson diversity was also detected in NVT strains from under-fives but to a lesser extent while the opposite trend occurred in NVT strains from over-fives (S10a,b Fig, S7 Table), implying incomplete serotype replacement by NVT serotypes post-vaccination. Pre-vaccination diversity was similar for VT and NVT strains but following vaccination Simpson diversity was higher in NVT than VT strains following vaccination (P=0.004, resampling) (Fig 5b). The Simpson diversity index was higher for STs than serotypes both before (P=0.011, resampling) and after vaccination (P=0, resampling) (S9 Fig). No changes in Simpson diversity were detected for the composition of STs post-vaccination (S10 Fig, S6,7 Table). The high stability of Simpson diversity in NVT strains post-vaccination could imply limited serotype replacement by NVTs, which has resulted in incomplete restoration of the serotype diversity to the levels observed pre-vaccination following vaccine-induced clearance of VT strains.

Antibiotic resistance

An important set of pneumococcal accessory genes include those encoding proteins that confer resistance to antibiotics. Previous work [21] showed almost universal resistance and complete sensitivity of pneumococcal strains to co-trimoxazole and ceftriaxone respectively therefore we did not investigate these antibiotics. Resistance rates were detected genotypically by quantifying the frequency of the antibiotic resistance conferring genes namely chloramphenicol acetyltransferase encoding gene (cat_{pC194}) for chloramphenicol, macrolide efflux pump encoding genes (*mefA* and *mefE*) and ribosomal RNA methyltransferase (*ermB*) for erythromycin, ribosomal protection protein-encoding gene (*tetM*) for tetracycline. For penicillin, the minimum inhibitory concentrations (MIC) were genotypically predicted using allelic variation in the transpeptidase domain of the penicillin binding proteins (PBP) from which the binary resistant-susceptible phenotype was inferred using BSAC criteria [22] (Fig 6a-d). There were no significant changes in resistance rates against the four antibiotics (Fig 6e-h). Interestingly, despite no change in penicillin resistance rate post-vaccination, the MICs decreased significantly (P=0.0098, Student's t test) post-vaccination due to vaccine-induced clearance of intermediate resistant lineages particularly serotype 3 in GC12 (Fig 6i, S8 Table). This exemplifies how vaccine usage can be strategically employed to clear not only highly prevalent and antibiotic resistant pneumococcal lineages globally but also intermediate resistant lineages with high likelihood to express full resistance before they do [23]. Further genomic analysis revealed the existence of a diverse catalogue of mobile genetic elements (MGE)

which disseminated genes responsible for resistance against macrolides, tetracycline and chloramphenicol antibiotics (Fig 6j). The macrolide (erythromycin) resistance conferring genes were associated with Tn*916*, Tn*6003*, Tn*2009* and Tn*2010* elements but not Tn*1545*, which is common elsewhere [24] (Fig 6c,d). The *mefA/E* gene were disseminated by Tn*2009* and Tn*2010*-like elements while ermB was carried by Tn*6003* and Tn*916*-like conjugative elements (Fig 6c,d). Both cat_{pC194} and *tetM* genes were located on Tn*5253*-like conjugative elements but tetM was more common due to its association with additional independent elements mainly Tn*5251* and Tn*2009*-like elements (Fig 6c,d). These findings suggests no significant post-vaccination perturbation of the accessory gene pool associated with antibiotic resistance.

Vaccine-induced accessory genome dynamics

Distribution of intermediate frequency accessory genes detected in 5% to 95% of the strains were similar between different subsets of isolates including those from over-fives and under-fives regardless of serotype category both pre- and post-vaccination (Fig 7a, S12 Fig). No significant differences in frequency of accessory genes were detected among VT strains pre- and post-vaccination while three accessory genes showed significant change among NVT strains following vaccination (Fig 7a,b). The accessory genes which changed in frequency post-vaccination were detected by fitting a logistic regression model for the binary presence-absence genotype for each gene and assessing the coefficients for the sampling period. Other variables namely age group and serotype category were included in the model formulation to account for their group effects. Similar analysis was done to determine accessory genes associated with different age groups and serotype categories. Bonferroni adjustment was done to control for multiple testing and no genes were associated with specific age group (Fig 7d) but as expected many genes (≈400) were associated with different serotype categories (Fig 7e). Forty-two accessory genes showed significant change in frequency post-vaccination, of which approximately half (52.38%) increased post-vaccination (Table 1, Fig 7f, S9 Table). The accessory genes that increased post-vaccination encoded for a glycosyl transferase (odds ratio[OR]=3.34, P=9.71×10^{-6}), bacteriolysin family protein (OR=3.34, P=3.46×10^{-5}), type I restriction modification system (RMS) R subunit (OR=3.34, P=3.46×10^{-5}) and other diverse functions including sugar transport, MGE and phage activity. Conversely, genes whose frequency decreased were associated with bacteriocin gene *blpQ* (OR=0.19, P=5.20×10^{-6}) and other genes associated with bacteriocins, conjugative elements, two-component systems and phages (all OR≈0.41, P<1.52×10^{-2}). Interestingly, a capsule biosyn-

thesis gene (*wzx*) encoding a repeating unit of a flippase protein also showed a decreasing frequency (OR=0.43, P=0.042). Other genes with significant changes in frequency encoded proteins with a diverse functional repertoire but notably included MGEs [phages/transposase/insertion elements] (11/42), bacteriocins (3/42), RMS systems (2/42) and competence (2/42), which have been recently implicated to play a role in negative frequency dependent selection (NFDS) [25]. These findings indicate that PCV has perturbed pneumococcal population not only at serotype level but also accessory genome level.

Discussion

The evolution of the pneumococcus in carriage is ever-changing therefore sampling this niche provides unrivalled opportunities to monitor and keep track of ongoing genomic adaptations over-time [5, 7]. Our genomic study demonstrates this by demonstrating changes in pneumococcal population structure following the introduction of PCV vaccine in previously vaccine-naïve low-income setting where, unlike in higher income settings such information is usually unavailable despite higher disease burden [26]. In contrast from other studies, our dataset is unique because it includes strains from both children and adults, which provides a more an opportunity to investigate PCV-induced population structuring in both vaccine-eligible and ineligible individuals. Our findings reveal changes at serotype, lineage and accessory genome level caused by vaccine-induced population restructuring specifically due to substantial clearance of VT serotypes in vaccinated under-five population but not unvaccinated over-five population. This reflects a direct consequence of PCV but limited and possibly delayed indirect protection in contrast to other settings e.g. UK where population herd immunity is higher albeit epidemiological differences with our population [3, 10]. Further ongoing and future studies will investigate factors mitigating herd effects, which may include higher HIV burden and vaccine scheduling. Lineage-specific changes revealed subtle serotype dynamics associated with clearance, clonal expansion and emergence of serotypes. Furthermore, while there was no change in the accessory gene pool associated with antibiotic resistance, there significant changes in other genes typically those encoding MGEs and bacteriocins. Additional statistical modelling shows that other accessory genes have undergone significant changes in frequency due to the population perturbations, which could be temporary and may return to equilibrium gene frequency due to NFDS [25]. Altogether, these findings reveal an early impact of PCV on serotype and lineage distribution, and accessory

gene dynamics in pneumococcal carriage post-vaccination.

The impact of PCV is shown by the reduction in frequency of VT serotypes, which occurred only in the largely vaccinated under-five population but not older and unvaccinated individuals. This implied high direct impact but limited indirect protection via herd immunity. While the frequency of VTs was higher in under-fives than over-fives, following vaccination these converged to the same amount (33%) but whether this is coincidental or represents some unknown phenomena threshold for VT frequency in the population remains unknown. Further assessment of Simpson diversity in serotype composition in VT and NVT strains provided further insights on the strain dynamics. Firstly, the Simpson diversity for ST composition remained largely unchanged post-vaccination [27, 28]. This suggests PCV did not significantly disrupt the ST composition in a similar fashion to the way it did for serotype composition possibly because of higher (≈3 times) saturation of STs than serotypes in pneumococcal population. Secondly, there was an increase of the diversity in NVT strains after vaccination albeit not statistically significant but serotype diversity was not restored to the levels observed pre-vaccination before clearance of VT serotypes, which implied that serotype replacement was incomplete. How serotype replacement in our setting will compare with other settings where settings remains to be fully understood once the population has reached a new equilibrium and replacement is complete.

Consistent with data from other settings [13], the observed modest serotype replacement has been driven by clonal expansion of extant serotypes prior to vaccination, which were masked due to competition by their VT counterparts. Unlike post-PCV7 introduction where serotype 19A was the dominant replacement serotype, in our setting replacement is driven by multiple serotypes including serotype 7C, 23A, 15B/C and 28F rather than a single dominant NVT clone. Furthermore, while certain capsule-switched strains were detected post-vaccination but whether these occurred post-vaccination due to vaccine pressure is unknown but there was evidence that the majority of these were likely extant prior to vaccination but at undetectable frequencies such as the serotype 9V to 28F vaccine-escape capsule-switched strains. Emergence of novel serotypes post-vaccination was uncommon with only serotype 28F detected post-vaccination only but this clone showed high sequence diversity and was not genetically similar to strains from other countries suggesting not only no recent importation post-vaccination but also circulation at undetectable levels prior to vaccina-

tion. Further studies are needed to assess the degree with which such replacement serotypes will cause disease in our setting and globally [29]. Taken together, these findings suggest there is incomplete serotype replacement in our setting after two years from PCV introduction mostly in under-fives compared to older age groups.

Clinically, the stability of the antibiotic resistance rates is not concerning considering that resistance rates were already lower than in IPD [19, 21, 30]. However, consistent with findings elsewhere [31, 32], the significant decrease of intermediate-penicillin resistance rate exemplifies the advantages of PCVs when strategically harnessed to thwart further emergence and expansion of clones with intermediate resistance to antimicrobials before they higher resistance is achieved [31]. While the accessory gene pool associated with antibiotic resistance did not change significantly, the changes detected in other accessory genes using logistic regression signaled that the pneumococcal population has been perturbed and these could likely indicate genes with the potential to drive NFDS. The perturbed-frequency genes revealed diverse functions with the majority associated with MGEs possibly reflecting their rapid mobility between pneumococcal strains. Other genes were associated with capsule biosynthesis (*wzx*) while those encoding for bacteriocin activity associated proteins important in mediating strain competition dynamics were also possibly under selection [33, 34]. These finding provide a clear evidence that PCV has induced population changes not only at serotype but also accessory genome level but these changes did not seemingly favour higher antibiotic resistance. The resultant phenotypic changes associated with the perturbed accessory genes and allelic variation in the core gene previously shown to drive metabolic shifts in strains post-vaccination warrants further investigation [35].

We have provided clear evidence of vaccine-induced perturbed pneumococcal population at serotype, lineage and accessory genome level after only two years following the introduction of PCV in a pneumococcal vaccine-naïve setting in Northern Malawi. This reflects a PCV impact, which is largely restricted to vaccinated under-fives but not unvaccinated older individuals highlighting limited manifestation of population herd immunity. Our data is timely and improves our understanding of the impact of PCV on pneumococcal population structure in high-disease burden but low-income settings. Our data revealed a perturbed pneumococcal carriage population after two years from PCV introduction but continued assiduous surveillance and WGS remain crucial to adequately monitor long-term effects of PCV particu-

larly after the equilibrium population dynamics have been re-established. Together with findings gained from surveillance and clinical trials of PCVs across SSA and globally, our findings will inform future conjugate vaccine design strategies and how their beneficial effects can be maximised especially in the most vulnerable tropical populations.

Materials and methods

Study population and isolate selection

A subset of 660 pneumococcal isolates collected through multiple household surveys were selected for WGS (S1 Data). These isolates were obtained from nasopharyngeal carriage in healthy children and adults in Karonga district of northern Malawi before vaccination in 2009 (n=370) and 2010 (n=112), and two years after vaccination in 2014 (n=178) were selected for WGS. The strains were representative of nasopharyngeal swabs and were samples from individuals from different age groups. By age group, 376 strains were sampled from the under-fives (<5 years old) and 130 strains from over-fives (≥5 years old) before vaccination while 106 and 48 strains were obtained from under-fives and over-fives post-vaccination. The nasal swabs were stored and processed as previously described [36]. The ethical approvals for the study were granted by National Health Sciences Research Committee in Malawi (NHSRC 490) and University of Malawi College of Medicine Research Ethics Committee (P.O8/14/1614).

Genomic DNA preparation and sequencing

Genomic DNA extraction was done using QIAamp DNA mini kit (Qiagen, Hilden, Germany), QIAgen Biorobot (Qiagen, Hilden, Germany) and Wizard® DNA Genomic DNA Purification Kit (Promega, Wisconsin, USA) as previously described [37]. Preparation of genomic DNA libraries and sequencing was done at Wellcome Sanger Institute using Illumina Genome Analyzer II and HiSeq platforms (Illumina, CA, USA). The length of reads ranged between 100 and 125 bases (median: 100) with a mean quality of 35.32 (S1 Fig) while mean number of mapped reads was 4,068,781 reads per isolate (S2 Fig). De novo sequence assemblies were generated using a pipeline [38], which uses Velvet v1.2.09 [39] and VelvetOptimiser v2.2.5 [40] (k-mers between 66.0% to 90.0% of the read length) to assemble reads, SSPACE Basic v2.0 [41] for assembly scaffolding (≈16 iterations), GapFiller v1.10 [42] for assembly gap closing and SMALT v0.7.4 (www.sourceforge.net/projects/smalt/) to re-map reads to the

assembly. The mean genome size, contig length and number of contigs were 2,125,143bp, 54,071bp and 48 contigs respectively while the mean N50 values was 129,431bp (S3,4 Fig). The sequence reads were deposited in the European Nucleotide Archive (S1 Data).

Capsule and sequence typing

The capsule types (serotypes) were defined using an *in silico* typing approach [43], which maps short sequence reads against reference sequences of the capsule-encoding polysaccharide synthesis (CPS) loci, which determines serotypes based antibody binding to its antigens [44]. The sequence types (ST) were inferred from assemblies based on loci of seven housekeeping genes for pneumococcal multilocus sequence typing (MLST) scheme [45] using MLSTcheck [46].

Presence of antibiotic resistance genes

The presence of antibiotic resistance genes for different antibiotics (tetracycline, chloramphenicol and erythromycin) and mobile genetic elements that disseminate them were investigated using nucleotide-BLAST v2.2.30 [47] with E-value of <0.001, sequence coverage >80% and nucleotide identity >80%. For penicillin whose resistance is caused by chromosomal mutations unlike genes disseminated by mobile genetic elements (MGE), the minimum inhibitory concentrations (MIC) were genotypically predicted using a robust analysis pipeline developed by the Centers for Disease Control and Prevention (CDC), which uses allelic variation in the transpeptidase domain of the penicillin binding proteins (PBP) - PBP1a, PBP2b and PBP2x to infer MICs [22]. The inferred MICs were translated into binary resistant-susceptible phenotype using the British Society for Antimicrobial Chemotherapy (BSAC) breakpoint [48].

Gene annotation, alignment and population structure

The sequenced and assembled genomes were annotated using Prokka v1.11 [49]. We identified core and accessory genes, present in ≥99% and <99% of the strains respectively, by clustering coding sequences using Roary v3.6.1 pan-genome pipeline [50]. The core- and pan-genome were comprised of 660 and 9,472 genes respectively and a 1,050,021bp core-genome alignment with 88,961bp single nucleotide polymorphism (SNP) positions identified using Snp-Sites v2.3.2 [51] was generated from Roary analysis (S5 Fig). The core SNPs were clustered into genomic clusters (GCs) using the

hierarchical clustering module (hierBAPS) in BAPS v6.0 [18]. DNA sequence translation and format conversion was done using BioPython [52] while nucleotide-BLAST v2.2.30 [47] and ACT v13.0.0 [53] was used for sequence comparison.

Phylogenetic tree construction

Maximum likelihood phylogenetic trees of the strains was constructed from the core gene alignment using FastTree-SSE3 v2.1.3 [54]. The GC or lineage-specific trees were constructed using RAxML v7.0.4 [55] from whole genome alignments after removing regions with putative recombination events using Gubbins v1.4.10 [56]. The phylogenetic trees were generated using a general time reversible (GTR) model [57], Gamma heterogeneity between nucleotide sites [58] and 100 bootstrap replicates parameters [59]. The phylogenetic tree was rooted at the midpoint of the branch separating most divergent strains. Visualisation of the tree Together with the strain's metadata was done using iToL v2.1 [60].

Statistical analysis

The changes in frequency of serotypes, antibiotic resistance and accessory genome content were assessed using Fisher's Exact test. Odds ratios for detecting GCs and serotypes pre- and post-vaccination were determined (pseudo-counts of 1 to avoid division by zeros). Changes in the composition of serotypes and STs were detected by using Simpson diversity index and the P-values were detected by resampling using Jackknife approach (www.comparingpartitions.info). Logistic regression was used to assess changes in gene presence-absence patterns of the intermediate frequency accessory genes (present at frequency from 5% to 95%) before and after vaccination while controlling for the effects of age group and serotype category. The reference levels for each variable in the regression were as follows: 'pre-vaccination' for sampling period, 'over-five' for age group and 'NVT' for serotype category. The estimated coefficients for each variable were extracted and summarised and P-values were for each gene were adjusted using Bonferroni correction to account for multiple comparisons. The statistical analysis was done using R v3.1.2 (R Core Team, 2013), GraphPad Prism v7.0 (GraphPad Software, California, USA) and Python v2.7.9 (Python Software Foundation).

Ethics statement

Written informed consent was obtained from adults while parents, guardians and caregivers of child participants. The ethical approvals for the study were granted by National Health Sciences Research Committee in Malawi (approval #: NHSRC 490 and 1232), the London School of Hygiene and Tropical Medicine (approval #5345) and the University of Liverpool (approval #: 670) and University of Malawi College of Medicine Research Ethics Committee (approval #: P.O8/14/1614). Nasopharyngeal swabs were collected from healthy children and adults as previously described and the samples were de-identified in the analysis.

Author contributions

C.C., N.F., D.B.E. and S.D.B. conceived and designed the study. N.F., D.B.E. and S.D.B. supervised the study. N.F., D.B.E., L.M., R.F.B. and S.D.B. secured funding. E.H., T.T. and N.F. collected samples. M.A., A.W.K., J.E.C. and C.P. performed molecular and microbiology experiments. S.D.B. supervised whole genome sequencing and genomic analysis. C.C. and R.A.G. checked quality of the sequence assemblies. C.C. performed genomic and statistical analyses. P.M. assisted with data analysis and interpretation. C.C. and S.D.B. wrote initial draft of the paper. L.M., R.F.B., A.K., W.P.H. and R.S.H. contributed to data interpretation. All authors contributed to writing and reviewing of the paper.

Competing interests

The authors declare no competing financial interests.

Acknowledgements

We acknowledge work by clinical and laboratory staff at Malawi Epidemiology and Intervention Research Unit (MEIRU) and Malawi-Liverpool-Wellcome Trust Clinical Research Programme (MLW) who collected and prepared the samples respectively, and sequencing and informatics teams at Wellcome Sanger Institute. This work was funded by Bill and Melinda Gates Foundation (grant #OPP1034556) and Wellcome UK (grant #084679/Z/08/Z). C.C. acknowledge PhD studentship funding from Commonwealth Scholarship Commission. The contents of this paper are solely

504 responsibility of the authors and does not necessarily represent official views of their
505 affiliated institutions and funding agencies.

506 References

507 1. Ben-Shimol, S. *et al.* Impact of Widespread Introduction of Pneumococcal Con-
508 jugate Vaccines on Pneumococcal and Nonpneumococcal Otitis Media. *Clin In-*
509 *fect Dis* **63,** 611–8. ISSN: 1058-4838 (2016).

510 2. Whitney, C. G. *et al.* Decline in Invasive Pneumococcal Disease after the Intro-
511 duction of Protein–Polysaccharide Conjugate Vaccine. *New England Journal of*
512 *Medicine* **348,** 1737–1746. ISSN: 0028-4793 (2003).

513 3. Miller, E., Andrews, N. J., Waight, P. A., Slack, M. P. & George, R. C. Herd
514 immunity and serotype replacement 4 years after seven-valent pneumococcal
515 conjugate vaccination in England and Wales: an observational cohort study.
516 *Lancet Infect Dis* **11,** 760–8. ISSN: 1473-3099 (2011).

517 4. Gordon, S. B. *et al.* Poor potential coverage for 7-valent pneumococcal conju-
518 gate vaccine, Malawi. *Emerg Infect Dis* **9,** 747–9. ISSN: 1080-6040 (Print) 1080-
519 6040 (2003).

520 5. Gottberg, A. *et al.* Effects of vaccination on invasive pneumococcal disease in
521 South Africa. *N Engl J Med* **371.** doi:10.1056/NEJMoa1401914. http://dx.doi.
522 org/10.1056/NEJMoa1401914 (2014).

523 6. Moore, M. R. *et al.* Effect of use of 13-valent pneumococcal conjugate vaccine
524 in children on invasive pneumococcal disease in children and adults in the USA:
525 analysis of multisite, population-based surveillance. *Lancet Infect Dis.* ISSN:
526 1473-3099. doi:10.1016/s1473-3099(14)71081-3. http://ac.els-cdn.com/
527 S1473309914710813/1-s2.0-S1473309914710813-main.pdf (2015).

528 7. Hammitt, L. L. *et al.* Population effect of 10-valent pneumococcal conjugate
529 vaccine on nasopharyngeal carriage of Streptococcus pneumoniae and non-typeable
530 Haemophilus influenzae in Kilifi, Kenya: findings from cross-sectional carriage
531 studies. *Lancet Glob Health* **2.** doi:10.1016/S2214-109X(14)70224-4. http:
532 //dx.doi.org/10.1016/S2214-109X(14)70224-4 (2014).

533 8. Mackenzie, G. A. *et al.* Effect of the introduction of pneumococcal conjugate
534 vaccination on invasive pneumococcal disease in The Gambia: a population-
535 based surveillance study. *Lancet Infect Dis* **16,** 703–11. ISSN: 1473-3099 (2016).

9. Cohen, C. *et al.* Effectiveness of the 13-valent pneumococcal conjugate vaccine against invasive pneumococcal disease in South African children: a case-control study. *The Lancet Global Health.* ISSN: 2214-109X. doi:10.1016/S2214-109X(17)30043-8. http://dx.doi.org/10.1016/S2214-109X(17)30043-8http://www.thelancet.com/pdfs/journals/langlo/PIIS2214-109X(17)30043-8.pdf (2017).

10. Van Hoek, A. J. *et al.* Pneumococcal carriage in children and adults two years after introduction of the thirteen valent pneumococcal conjugate vaccine in England. *Vaccine* **32,** 4349–55. ISSN: 0264-410x (2014).

11. Brugger, S. D., Frey, P., Aebi, S., Hinds, J. & Muhlemann, K. Multiple colonization with S. pneumoniae before and after introduction of the seven-valent conjugated pneumococcal polysaccharide vaccine. *PLoS One* **5,** e11638. ISSN: 1932-6203 (2010).

12. Dunne, E. M. *et al.* Effect of pneumococcal vaccination on nasopharyngeal carriage of Streptococcus pneumoniae, Haemophilus influenzae, Moraxella catarrhalis, and Staphylococcus aureus in Fijian children. *J Clin Microbiol* **50,** 1034–8. ISSN: 0095-1137 (2012).

13. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* **45,** 656–63. ISSN: 1546-1718 (2013).

14. Gladstone, R. A. *et al.* Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine* **33,** 2015–21. ISSN: 0264-410x (2015).

15. Weinberger, D. M., Malley, R. & Lipsitch, M. Serotype replacement in disease following pneumococcal vaccination: A discussion of the evidence. *Lancet* **378,** 1962–1973. ISSN: 0140-6736 1474-547X (2011).

16. Bar-Zeev, N. *et al.* Methods and challenges in measuring the impact of national pneumococcal and rotavirus vaccine introduction on morbidity and mortality in Malawi. *Vaccine* **33,** 2637–45. ISSN: 0264-410x (2015).

17. Wyres, K. L. *et al.* Pneumococcal Capsular Switching: A Historical Perspective. *Journal of Infectious Diseases* **207,** 439–449 (2013).

18. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Molecular Biology and Evolution.* doi:10.1093/molbev/mst028. http://mbe.oxfordjournals.org/content/early/2013/02/13/molbev.mst028.abstracthttp://mbe.oxfordjournals.org/content/30/5/1224.full.pdf (2013).

570 19. Chaguza, C. *et al.* Population genetic structure, antibiotic resistance, capsule
571 switching and evolution of invasive pneumococci before conjugate vaccination
572 in Malawi. *Vaccine.* ISSN: 0264-410X. doi:doi.org/10.1016/j.vaccine.2017.07.009.
573 http://www.sciencedirect.com/science/article/pii/S0264410X17309052http:
574 //ac.els-cdn.com/S0264410X17309052/1-s2.0-S0264410X17309052-main.pdf
575 (2017).

576 20. Croucher, N., Harris, S., Fraser, C. & Quail, M. Rapid pneumococcal evolution
577 in response to clinical interventions. *Science* **331.** doi:10.1126/science.1198545.
578 http://dx.doi.org/10.1126/science.1198545 (2011).

579 21. Everett, D. B. *et al.* Ten years of surveillance for invasive Streptococcus pneu-
580 moniae during the era of antiretroviral scale-up and cotrimoxazole prophylaxis
581 in Malawi. *PLoS One* **6,** e17765. ISSN: 1932-6203 (Electronic) 1932-6203 (Link-
582 ing) (2011).

583 22. Li, Y. *et al.* Penicillin-Binding Protein Transpeptidase Signatures for Tracking
584 and Predicting β-Lactam Resistance Levels in Streptococcus pneumoniae. *mBio*
585 **7.** doi:10.1128/mBio.00756-16. http://mbio.asm.org/content/7/3/e00756-
586 16.abstracthttp://mbio.asm.org/content/7/3/e00756-16.full.pdf (2016).

587 23. Henriques-Normark, B. & Normark, S. Bacterial vaccines and antibiotic resis-
588 tance. *Ups J Med Sci* **119,** 205–8. ISSN: 0300-9734 (2014).

589 24. Xu, X. *et al.* Distribution of Serotypes, Genotypes, and Resistance Determi-
590 nants among Macrolide-Resistant Streptococcus pneumoniae Isolates. *Antimi-*
591 *crobial Agents and Chemotherapy* **54,** 1152–1159. ISSN: 0066-4804 1098-6596
592 (2010).

593 25. Corander, J. *et al.* Frequency-dependent selection in vaccine-associated pneu-
594 mococcal population dynamics. *Nature Ecology & Evolution.* ISSN: 2397-334X.
595 doi:10.1038/s41559-017-0337-x. https://doi.org/10.1038/s41559-017-0337-x
596 (2017).

597 26. O'Brien, K. L. *et al.* Burden of disease caused by Streptococcus pneumoniae
598 in children younger than 5 years: global estimates. *Lancet* **374,** 893–902. ISSN:
599 1474-547X (Electronic) 0140-6736 (Linking) (2009).

600 27. Chang, Q. *et al.* Stability of the pneumococcal population structure in Mas-
601 sachusetts as PCV13 was introduced. *BMC Infect Dis* **15,** 68. ISSN: 1471-2334
602 (Electronic) 1471-2334 (Linking) (2015).

28. Azarian, T. *et al.* The impact of serotype-specific vaccination on phylodynamic parameters of Streptococcus pneumoniae and the pneumococcal pan-genome. *PLoS Pathog* **14,** e1006966. ISSN: 1553-7366 (2018).

29. Masala, G. L., Lipsitch, M., Bottomley, C. & Flasche, S. Exploring the role of competition induced by non-vaccine serotypes for herd protection following pneumococcal vaccination. *J R Soc Interface* **14.** ISSN: 1742-5662. doi:10.1098/rsif.2017.0620. http://rsif.royalsocietypublishing.org/content/royinterface/14/136/20170620.full.pdf (2017).

30. Musicha, P. *et al.* Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998-2016): a surveillance study. *Lancet Infect Dis* **17,** 1042–1052. ISSN: 1473-3099 (2017).

31. Su, L.-H. *et al.* Evolving pneumococcal serotypes and sequence types in relation to high antibiotic stress and conditional pneumococcal immunization. *Scientific Reports* **5,** 15843. ISSN: 2045-2322 (2015).

32. Kyaw, M. H. *et al.* Effect of introduction of the pneumococcal conjugate vaccine on drug-resistant Streptococcus pneumoniae. *N Engl J Med* **354,** 1455–63. ISSN: 1533-4406 (Electronic) 0028-4793 (Linking) (2006).

33. Son, M. R. *et al.* Conserved mutations in the pneumococcal bacteriocin transporter gene, blpA, result in a complex population consisting of producers and cheaters. *MBio* **2.** doi:10.1128/mBio.00179-11 (2011).

34. Dawid, S., Roche, A. M. & Weiser, J. N. The blp bacteriocins of Streptococcus pneumoniae mediate intraspecies competition both in vitro and in vivo. *Infect Immun* **75,** 443–51. ISSN: 0019-9567 (Print) 0019-9567 (2007).

35. Watkins, E. R. *et al.* Vaccination Drives Changes in Metabolic and Virulence Profiles of Streptococcus pneumoniae. *PLoS Pathog* **11,** e1005034. ISSN: 1553-7366 (2015).

36. Heinsbroek, E. *et al.* Pneumococcal Acquisition Among Infants Exposed to HIV in Rural Malawi: A Longitudinal Household Study. *American Journal of Epidemiology* **183,** 70–78. ISSN: 0002-9262 1476-6256 (2016).

37. Everett, D. B. *et al.* Genetic characterisation of Malawian pneumococci prior to the roll-out of the PCV13 vaccine using a high-throughput whole genome sequencing approach. *PLoS One* **7.** doi:10.1371/journal.pone.0044250. http://dx.doi.org/10.1371/journal.pone.0044250 (2012).

636 38. Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and
637 improvement pipeline for Illumina data. *Microbial Genomics* **2.** doi:doi:10.
638 1099/mgen.0.000083. http://mgen.microbiologyresearch.org/content/journal/
639 mgen/10.1099/mgen.0.000083 (2016).

640 39. Zerbino, D. & Birney, E. Velvet: Algorithms for de novo short read assembly
641 using de Bruijn graphs. *Genome Res* **18,** 821 –829 (2008).

642 40. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing
643 technologies. *Curr Protoc Bioinformatics* **Chapter 11,** Unit 11 5. ISSN: 1934-
644 340X (Electronic) 1934-3396 (Linking) (2010).

645 41. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffold-
646 ing pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–9. ISSN: 1367-
647 4811 (Electronic) 1367-4803 (Linking) (2011).

648 42. Nadalin, F., Vezzi, F. & Policriti, A. GapFiller: a de novo assembly approach to
649 fill the gap within paired reads. *BMC Bioinformatics* **13,** S8. ISSN: 1471-2105
650 (2012).

651 43. Epping, L. *et al.* SeroBA: rapid high-throughput serotyping of Streptococcus
652 pneumoniae from whole genome sequence data. *Microb Genom.* ISSN: 2057-
653 5858. doi:10.1099/mgen.0.000186. http://www.microbiologyresearch.org/
654 docserver/fulltext/mgen/mgen.000186.zip/mgen000186.pdf (2018).

655 44. Bentley, S. *et al.* Genetic analysis of the capsular biosynthetic locus from all 90
656 pneumococcal serotypes. *PLoS Genet* **2.** doi:10.1371/journal.pgen.0020031.
657 http://dx.doi.org/10.1371/journal.pgen.0020031 (2006).

658 45. Enright, M. C. & Spratt, B. G. A multilocus sequence typing scheme for Strep-
659 tococcus pneumoniae: identification of clones associated with serious invasive
660 disease. *Microbiology* **144 (Pt 11),** 3049–60. ISSN: 1350-0872 (Print) 1350-
661 0872 (1998).

662 46. Page, A., Taylor, B. & Keane, J. Multilocus sequence typing by blast from de
663 novo assemblies against PubMLST. *The Journal of Open Source Software* **1.**
664 doi:doi:10.21105/joss.00118. http://dx.doi.org/10.21105/joss.00118 (2016).

665 47. Altschul, S. *et al.* Gapped BLAST and PSI-BLAST: a new generation of pro-
666 tein database search programs. *Nucleic Acids Res* **25,** 3389 –3402 (1997).

667 48. Andrews, J. M. BSAC standardized disc susceptibility testing method. *J An-*
668 *timicrob Chemother* **48 Suppl 1,** 43–57. ISSN: 0305-7453 (Print) 0305-7453
669 (2001).

49. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–9. ISSN: 1367-4803 (2014).

50. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3. ISSN: 1367-4803 (2015).

51. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* **2.** doi:doi:10.1099/mgen.0.000056. http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000056 (2016).

52. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–3. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (2009).

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

54. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26,** 1641–1650 (2009).

55. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688 –2690 (2006).

56. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research.* doi:10.1093/nar/gku1196. http://nar.oxfordjournals.org/content/early/2014/11/20/nar.gku1196.abstracthttp://nar.oxfordjournals.org/content/early/2014/11/20/nar.gku1196.full.pdf (2014).

57. Tavaré, S. in *American Mathematical Society: Lectures on Mathematics in the Life Sciences* 57–86 (Amer Mathematical Society, 1986). ISBN: 0821811673. doi:citeulike-article-id:4801403. http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0821811673.

58. Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* **10,** 1396–401. ISSN: 0737-4038 (Print) 0737-4038 (1993).

59. Felsenstein, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39,** 783–791. ISSN: 0014-3820 (1985).

703 60. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and dis-
704 play of phylogenetic trees made easy. *Nucleic Acids Research* **39,** W475–W478
705 (2011).

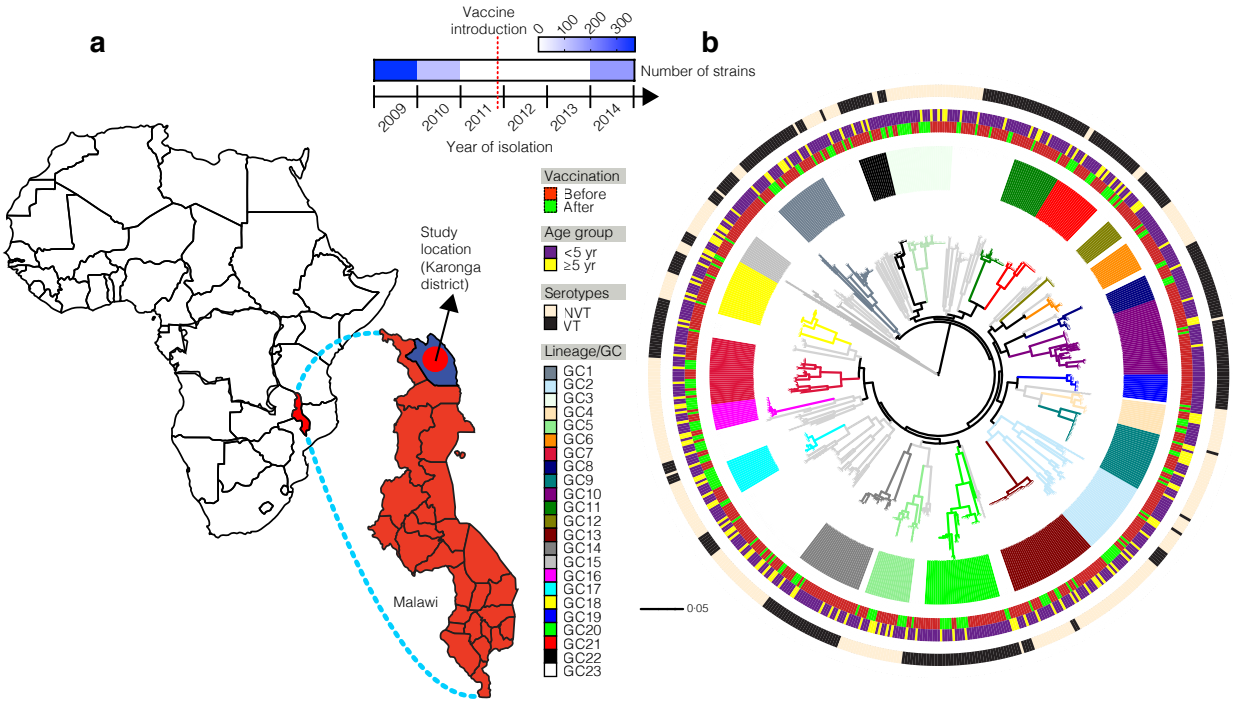

Fig 1: Core gene maximum likelihood phylogeny of carried pneumococci. (a) Sampling location of the study strains. (b) Maximum likelihood phylogeny of the carried pneumococcal strains reconstructed using core genome SNPs from 660 study strains demonstrating their genetic similarity and diversity. The strains' metadata namely GC (lineage), sampling period, age group and serotype category of the subjects shown by coloured strips around the phylogeny. The colours for phylogenetic branches correspond to the inferred GCs as shown in the metadata key next to the tree. The tree was rooted at the mid-point of the branch separating 'classical' non-typeable (NT) pneumococcal GC (GC15) from rest of the strains. Detailed characteristics of the study strains are provided in S1 Data.

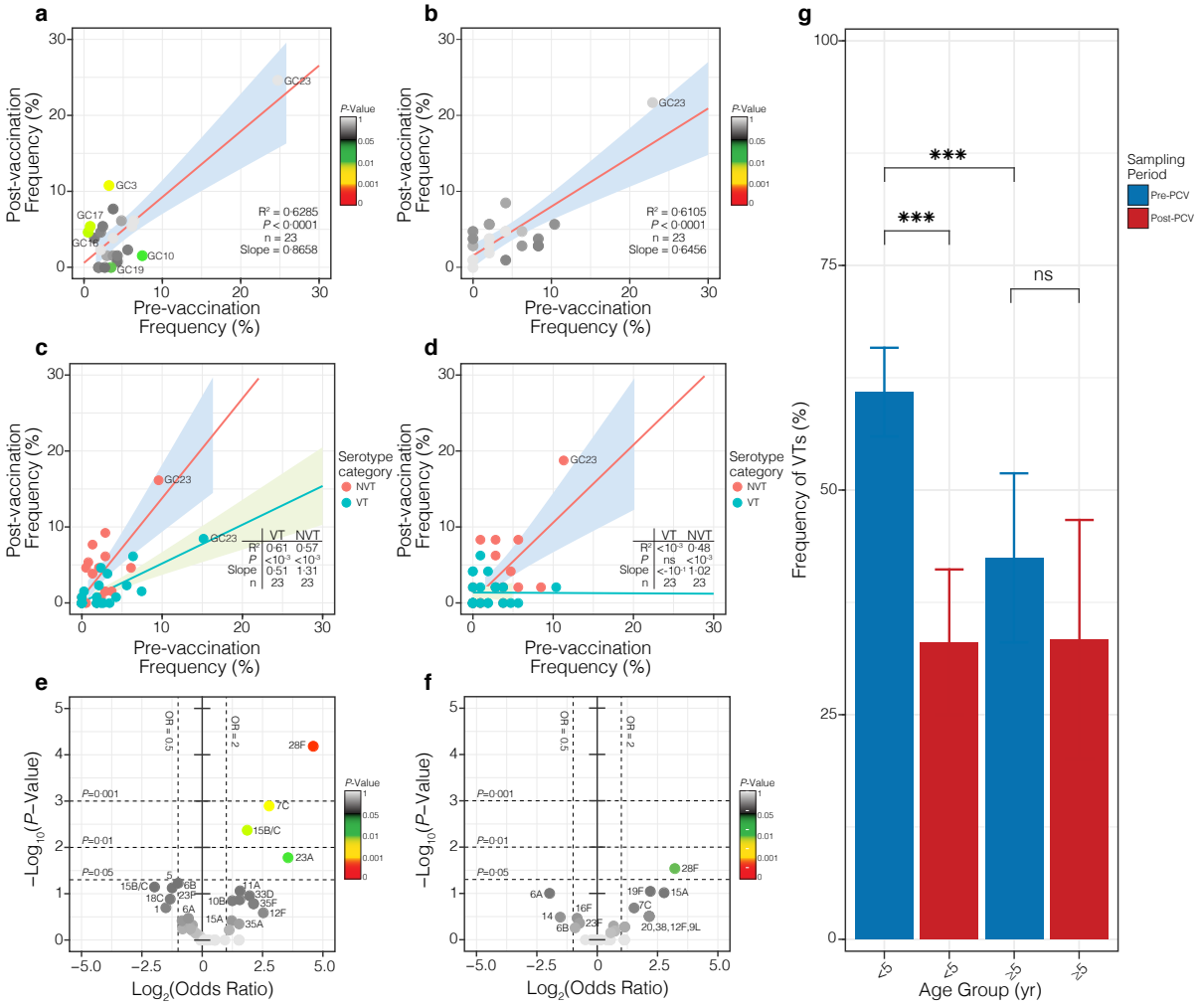

Fig 2: Frequency of GCs and serotypes in carriage. The scatter plot showing frequency of GCs before and after vaccination in (a) under-fives and (b) over-fives. Fitted linear regression for the overall trend in GC frequency of VT and NVT strains pre- and post-vaccination. (c) The scatter plots showing frequency of VT serotypes in GCs pre- and post-vaccination in (d) under-fives and (e) over-fives. The regression lines are bordered by 95% CI. The changes of individual serotype frequency are shown by volcano plots for strains in (f) under-fives and (g) over-fives where the x-axis shows magnitude ($\log_2$ odds ratio) and y-axis shows statistical significance (-$\log_{10}$ P-value) of change after compared to before vaccination. Statistically significant changes are marked with asterisks: 'ns': not significant, P<0.05 (*), P<0.01 (**) and P<0.001 (***).

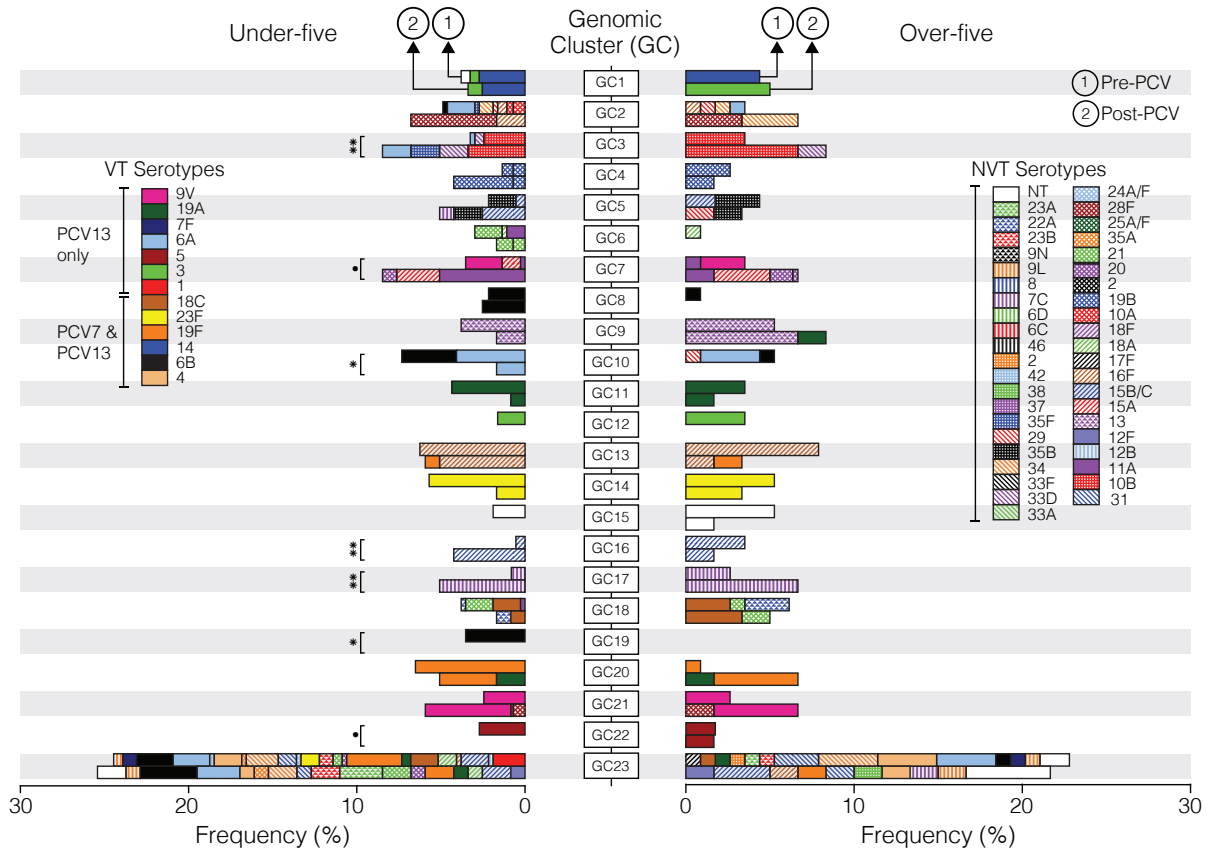

Fig 3: Dynamics of pneumococcal GCs and serotypes. The leftward facing stacked bar graph shows frequency of GCs in under-fives while (b) the rightward facing bar graph shows frequency of GCs and their constituent serotypes in over-fives before and after PCV introduction. The bar graphs are aligned by genomic clusters (GC) for easy comparisons of frequency of serotypes pre- and post-vaccination between the two age groups. The serotypes are distinguished by different colours in the bar graphs as described in the key. GC23 is the 'bin' cluster consisting of unclustered strains not placed in clusters GC1-22. The GCs whose frequency changed significantly post-vaccination are marked with asterisks: P<0.05 (*) and P<0.01 (**) and those with borderline significance P<0.095 (.). The Fisher's exact test was used to determine P-values.

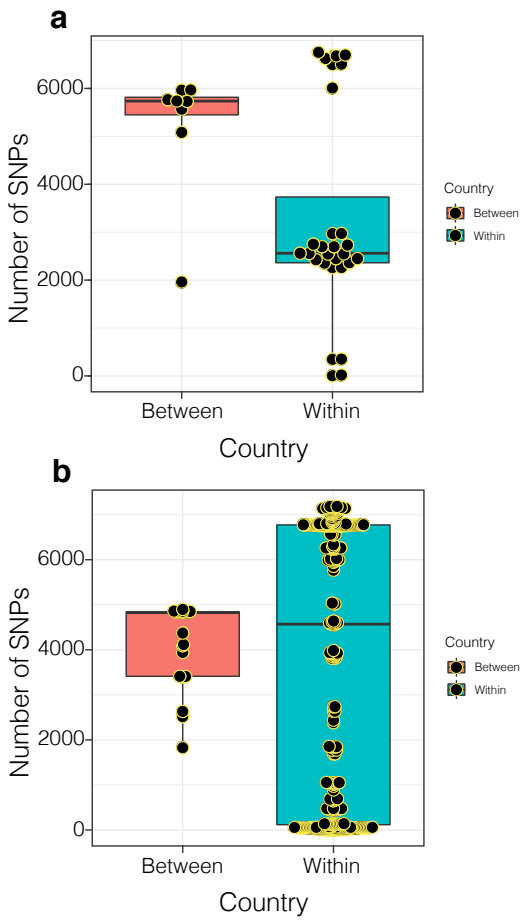

Fig 4: Genetic diversity of a recently emerged serotypes (28F). Boxplots showing within (Malawi) and between country (Malawi and South Africa) genetic diversity of serotype 28F strains showing in (a) GC2 (b) GC21. Lineage GC21 also include serotype 9V strains, which some of which underwent a capsule-switch to acquire a serotype 28F capsule. Additional details are provided in in S11 Fig

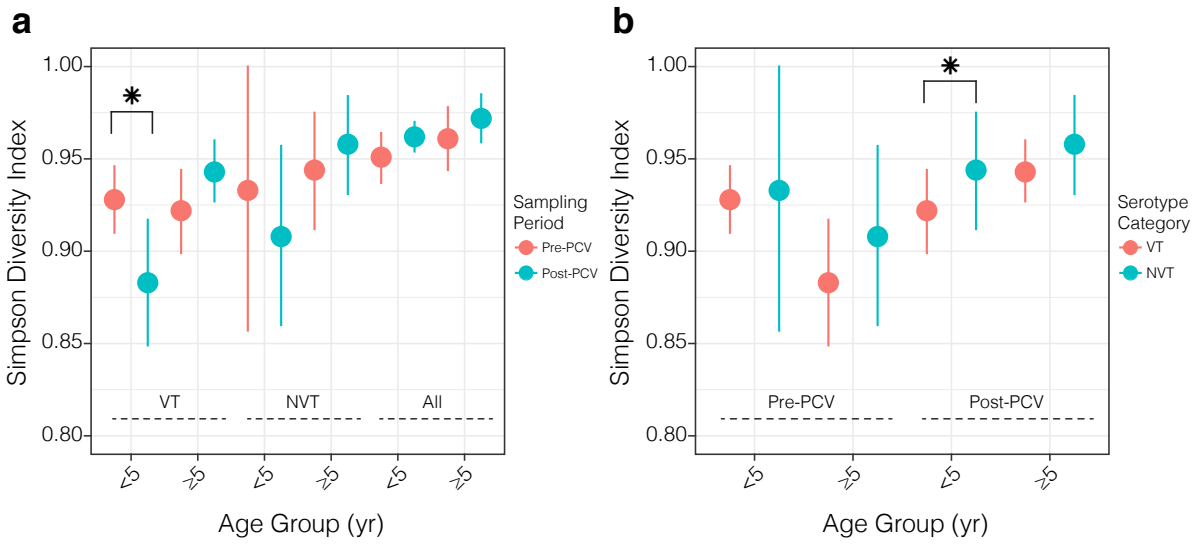

Fig 5: Serotype composition and diversity in context of PCV. (a) Simpson diversity index for composition of serotypes between pre- and post-vaccination datasets among VT, NVT and all strains among isolates. (b) Simpson diversity index for composition of serotypes between VT and NVT strains sampled pre- and post-vaccination. Statistically significant changes are marked with asterisks: 'ns': not significant, P<0.05 (*), P<0.01 (**) and P<0.001 (***). The estimates and P-values for frequency of VTs and Simpson diversity are summarised in S6,7 Table.

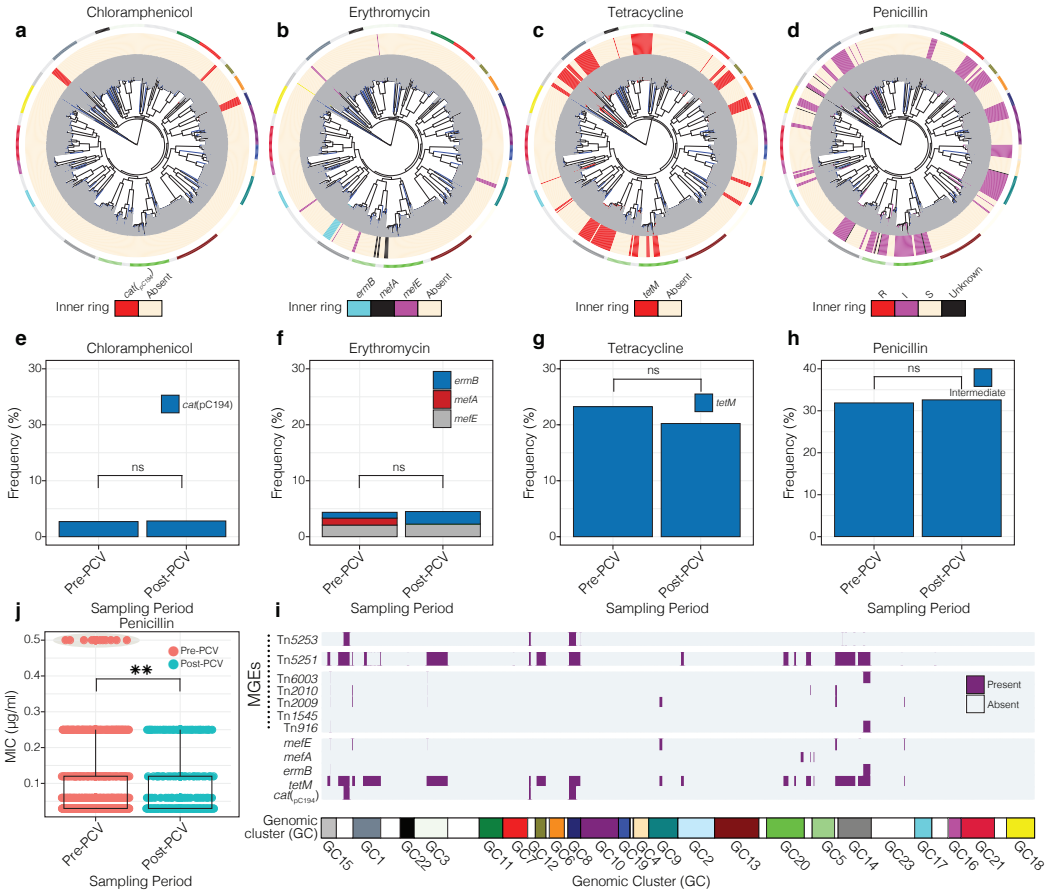

Fig 6: Distribution of antibiotic resistance genes and MGEs. (a) Distribution of cat_{pC194} chloramphenicol resistance gene. (b) Distribution of *mefA*, *mefE* and *ermB* erythromycin resistance genes. (c) Distribution of *tetM* tetracycline resistance gene. (d) Distribution of penicillin MICs. Branches of the maximum likelihood phylogenies and innermost ring surrounding the trees in a-c are coloured by presence and absence of the genes as shown in the key at the bottom of the phylogeny. The outermost ring around the phylogenies shows the GCs corresponding to those in Fig 1b and the colour strip at the bottom of this figure. (e-h) Frequency of genotypic antibiotic resistance rates for chloramphenicol, erythromycin, tetracycline and intermediate penicillin resistance pre- and post-vaccination. (i) Distribution of penicillin MICs pre- and post-vaccination. (j) The horizontal panel shows the distribution of the antibiotic resistance conferring genes and MGEs that disseminate them. The subsets with statistically significant changes are marked with asterisks: 'ns': not significant, P<0.05 (∗), P<0.01 (∗∗) and P<0.001 (∗∗∗).

706

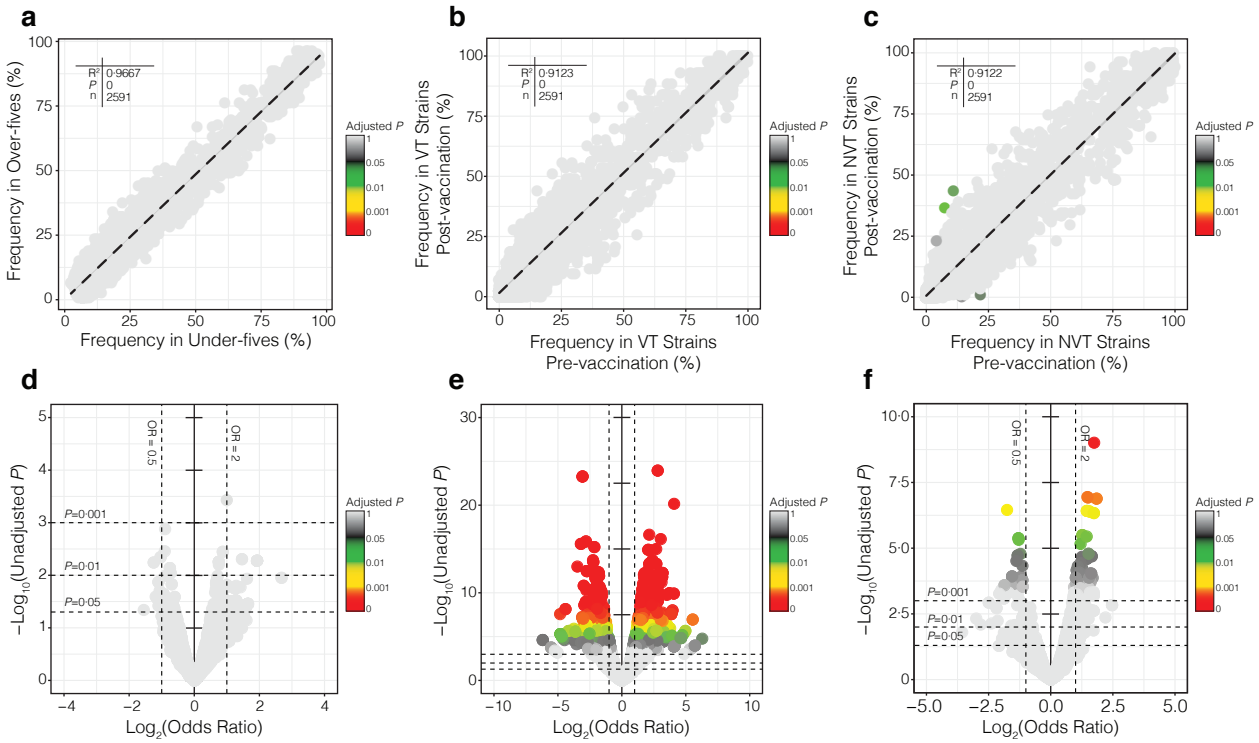

Fig 7: Pneumococcal accessory genome dynamics. The distribution of 2,591 intermediate frequency accessory genes in the entire pneumococcal population. (a) Scatter plot showing frequency of accessory genes between isolates sampled from under-fives and over-fives pre-vaccination. (b) Scatter plot showing frequency of genes among VT isolates pre- and post-vaccination. (c) Scatter plot showing frequency of genes among NVT isolates pre- and post-vaccination. Coefficients from linear regression and are labelled on the plots. Volcano plots show magnitude ($\log_2$ odds ratio) on the x-axis and statistical significance (-$\log_{10}$ *P*-value) for *P*-values and odds ratio for the association of accessory genes with different variables namely (d) age group, (e) serotype category and (f) sampling period. The points were coloured by adjusted *P*-values after correcting for multiple testing using Bonferroni method.

707

Table 1: Coefficients for the effect of sampling period on presence and absence of accessory genes using logistic regression accounting for age and serotype category after Bonferroni adjustment for multiple testing (only top 25 hits shown). Full list is shown in S9 Table.

Accessory gene	Odds ratio (95% CI)	Raw *P*-value	Adjusted *P*-value	Description/product
COG_6818	3.34 (2.24,4.99)	3.75×10^{-09}	9.71×10^{-06}	Glycosyl transferase
COG_7104	3.34 (2.20,5.06)	1.33×10^{-08}	3.46×10^{-05}	Double glycine cleavage site bacteriolysin superfamily
*hsdR*_1	3.34 (2.20,5.06)	1.33×10^{-08}	3.46×10^{-05}	Type I restriction-modification system, R subunit
COG_2038	2.78 (1.91,4.06)	1.12×10^{-07}	2.91×10^{-04}	ABC transporter permease
COG_900	2.78 (1.91,4.06)	1.12×10^{-07}	2.91×10^{-04}	IS*630*-Spn1, transposase *Orf1*
COG_5542	2.84 (1.93,4.18)	1.17×10^{-07}	3.03×10^{-04}	Replication protein
ptrB	0.30 (0.19,0.47)	3.48×10^{-07}	9.02×10^{-04}	Prolyl oligopeptidase family protein
COG_4000	2.95 (1.94,4.48)	3.96×10^{-07}	1.03×10^{-03}	Phage protein gp27
rlmN	2.95 (1.94,4.48)	3.96×10^{-07}	1.03×10^{-03}	Ribosomal RNA large subunit methyltransferase N
COG_4863	3.34 (2.09,5.33)	4.59×10^{-07}	1.19×10^{-03}	Hypothetical protein
COG_5488	2.54 (1.76,3.65)	5.24×10^{-07}	1.36×10^{-03}	Phage protein
COG_842	0.40 (0.28,0.58)	6.98×10^{-07}	1.81×10^{-03}	ABC transporter permease
IS861 truncation	0.40 (0.28,0.58)	6.98×10^{-07}	1.81×10^{-03}	IS*861* truncation
COG_1881	2.74 (1.82,4.14)	1.59×10^{-06}	4.11×10^{-03}	SNF2 family protein
COG_802	2.74 (1.82,4.14)	1.59×10^{-06}	4.11×10^{-03}	Transposase
COG_4974	3.00 (1.91,4.72)	1.93×10^{-06}	5.00×10^{-03}	Membrane associated protein
COG_1144	2.41 (1.67,3.50)	3.15×10^{-06}	8.17×10^{-03}	2-isopropylmalate synthase
COG_3978	0.41 (0.28,0.60)	4.17×10^{-06}	1.08×10^{-02}	Rep protein
COG_5509	0.41 (0.28,0.60)	4.17×10^{-06}	1.08×10^{-02}	Phage protein gp27
COG_6116	0.41 (0.28,0.60)	4.17×10^{-06}	1.08×10^{-02}	ABC transporter permease
saeS	0.41 (0.28,0.60)	4.21×10^{-06}	1.09×10^{-02}	Histidine kinase
COG_6305	0.41 (0.28,0.60)	4.77×10^{-06}	1.24×10^{-02}	Bacteriocin
COG_748	0.41 (0.28,0.60)	4.77×10^{-06}	1.24×10^{-02}	Hypothetical protein
blpPQ	0.19 (0.09,0.39)	5.20×10^{-06}	1.35×10^{-02}	Bacteriocin BlpPQ
COG_2726	0.41 (0.28,0.61)	5.84×10^{-06}	1.51×10^{-02}	Tn*5252*, *Orf* 9 protein