PLOS | PATHOGENS

RESEARCH ARTICLE

The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome

Taj Azarian[1]*, Lindsay R. Grant[2], Brian J. Arnold[1], Laura L. Hammitt[2], Raymond Reid[2], Mathuram Santosham[2], Robert Weatherholtz[2], Novalene Goklish[2], Claudette M. Thompson[1], Stephen D. Bentley[3], Katherine L. O'Brien[2], William P. Hanage[1], Marc Lipsitch[1]

1 Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University; Cambridge, Massachusetts, United States of America, **2** Center for American Indian Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland; United States of America, **3** Wellcome Trust Sanger Institute, Cambridge, United Kingdom

* tazarian@hsph.harvard.edu

Check for updates

OPEN ACCESS

Citation: Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M, et al. (2018) The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. PLoS Pathog 14(4): e1006966. https://doi.org/10.1371/journal.ppat.1006966

Editor: Christoph Tang, University of Oxford, UNITED KINGDOM

Received: October 26, 2017

Accepted: March 9, 2018

Published: April 4, 2018

Copyright: © 2018 Azarian et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Whole-genome sequencing data are available from NCBI under BioProject PRJEB8327: https://www.ncbi.nlm.nih.gov/bioproject/PRJEB8327. In addition, a list of accession numbers and accompanying metadata is provided in the supplementary material. The authors may be contacted for any additional data requests.

Funding: This study was supported by R01 R01AI048935, the Grand Challenges in Global

Abstract

In the United States, the introduction of the heptavalent pneumococcal conjugate vaccine (PCV) largely eliminated vaccine serotypes (VT); non-vaccine serotypes (NVT) subsequently increased in carriage and disease. Vaccination also disrupts the composition of the pneumococcal pangenome, which includes mobile genetic elements and polymorphic non-capsular antigens important for virulence, transmission, and pneumococcal ecology. Antigenic proteins are of interest for future vaccines; yet, little is known about how the they are affected by PCV use. To investigate the evolutionary impact of vaccination, we assessed recombination, evolution, and pathogen demographic history of 937 pneumococci collected from 1998–2012 among Navajo and White Mountain Apache Native American communities. We analyzed changes in the pneumococcal pangenome, focusing on metabolic loci and 19 polymorphic protein antigens. We found the impact of PCV on the pneumococcal population could be observed in reduced diversity, a smaller pangenome, and changing frequencies of accessory clusters of orthologous groups (COGs). Post-PCV7, diversity rebounded through clonal expansion of NVT lineages and inferred in-migration of two previously unobserved lineages. Accessory COGs frequencies trended toward pre-PCV7 values with increasing time since vaccine introduction. Contemporary frequencies of protein antigen variants are better predicted by pre-PCV7 values (1998–2000) than the preceding period (2006–2008), suggesting balancing selection may have acted in maintaining variant frequencies in this population. Overall, we present the largest genomic analysis of pneumococcal carriage in the United States to date, which includes a snapshot of a true vaccine-naïve community prior to the introduction of PCV7. These data improve our understanding of pneumococcal evolution and emphasize the need to consider pangenome composition when inferring the impact of vaccination and developing future protein-based pneumococcal vaccines.

Health initiative through the Bill & Melinda Gates Foundation, the Native American Research Centers for Health (U26IHS300013/03), the Centers for Disease Control and Prevention National Vaccine Program Office, and the Thrasher Research Fund (02820-9). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: M.L. has consulted for Pfizer, Affinivax and Merck and has received grant support not related to this paper from Pfizer and PATH Vaccine Solutions. W.P.H. and M.L. have consulted for Antigen Discovery Inc. The authors have declared that no competing interests exist.

Author summary

Pneumococcal disease caused by the bacteria *Streptococcus pneumoniae* remains a significant cause of morbidity and mortality despite the existence of an effective vaccine. This is because the vaccines only target a small proportion of the total pneumococcal population. Introduction of vaccine in the United States removed vaccine serotypes leaving an open niche that was rapidly filled by non-vaccine serotypes. Forecasting which serotypes, and more generally which pneumococcal lineages, will increase in frequency in carriage and disease is an active area of research with significant public health importance. Here, we investigate the evolutionary impact of vaccination on the pneumococcal population using genomic data from a collection of 937 pneumococcal isolates collected from 1998–2012 among Native American communities. We find the impact of vaccine on the pneumococcal population could be observed in reduced diversity and changing frequencies of genes. Diversity subsequently rebounded through expansion and in-migration of non-vaccine lineages. Further, frequencies of genes coding for protein antigens important to host-pathogen interaction were initially disrupted but later returned to pre-vaccine values, suggesting selection may have acted in maintaining frequencies. These data improve our understanding of pneumococcal evolution and emphasize the need to consider genome composition when inferring the impact of vaccination.

Introduction

Pneumococcal conjugate vaccines (PCV) target capsular serotype-specific polysaccharides of the respiratory pathogen *Streptococcus pneumoniae*, which causes substantial morbidity and mortality [1,2]. Since the heptavalent PCV and 13-valent PCV were introduced in the United States (US) in 2000 and 2010, respectively, their effectiveness in reducing pneumococcal carriage and invasive disease has been well documented [3–6]. In communities where PCV has been introduced, the prevalence of vaccine serotypes (VT) in carriage and invasive disease consistently decreases, resulting in an overall reduction in pneumococcal disease. However, in a process called "serotype replacement," non-vaccine serotypes (NVT) subsequently increase in carriage after vaccine introduction, leading to slight increases in NVT-associated disease in almost all populations where the vaccine is introduced [7,8]. Because polysaccharide serotypes change rarely during pneumococcal evolution, common pneumococcal lineages typically contain only one or a few serotypes. As a result, PCV implementation removes lineages containing only VT from the population, while lineages including both VT and NVT experience genetic bottlenecks [9–11].

Forecasting which serotypes, and more generally which pneumococcal lineages, will increase in frequency in carriage and disease is an active area of research with significant public health importance. For *S. pneumoniae*, the most commonly used vaccines globally target a fraction of the more than 93 recognized capsular serotypes [12]. The bacteria's capsule (CPS) is the most important determinant of virulence and the strongest predictor of prevalence [13], as well as the target of PCVs; thus, changes in CPS serotype frequency have been the focus of many analyses of vaccine effect. However, selection acts on genes outside the operon determining CPS serotype. Whole-genome sequencing data has enabled investigation of variation in multiple genomic loci and genome content among pneumococci, focused on loci involved in host immunity and niche adaption. We focus here on two categories of proteins. The first is antigens (hereafter, when we use the generic term antigen we refer to proteins that elicit an

immune response, not to the polysaccharide capsule). Antigens such as pneumococcal surface proteins A and C (*pspA* and *pspC*) and pilus are of specific interest as possible targets for non-capsular polysaccharide based vaccines [14]. Together with other components of the pneumococcal genome, the capsule and non-capsular antigens comprise the overall antigenic profile of a pneumococcus [15–17]. Moreover, evolution among metabolic genes gives rise to distinct metabolic-profiles among pneumococcal lineages, which may be adapted for specific metabolic niches [18,19]. Thus, multiple loci may interface with the host, affecting the overall evolutionary success of a lineage at a population level.

Gene content varies tremendously among pneumococcal lineages [20,21]. The pneumococcal pangenome consists of "core genes" shared by ≥99% of strains and "accessory genes" present at frequencies ≤99%. Accessory genes may include polymorphic antigens, phage and plasmid-related chromosomal islands, and integrative and conjugative elements (ICE) harboring antimicrobial resistance genes. The latter are mobile genetic elements (MGE), which are often acquired through horizontal gene transfer (HGT) and may remain stable in pneumococcal lineages [21]. Variations in gene content among lineages of a bacterial species are associated with ecological niche specialization and are important for adaptation to changing environments, including selection by vaccine-induced and natural host immunity [21–23]. For the pneumococcus, MGE affect the bacteria's ability to recombine (i.e., competence) [24], antimicrobial susceptibility [25], and carriage duration [26]. Accessory loci may also be acted upon by negative frequency dependent selection (NFDS), hinting at their underlying role in non-serotype-specific immunity and *S. pneumoniae* ecology [27]. Taken together, gene variation beyond the capsular polysaccharide loci may significantly impact virulence, fitness, transmission, and, in turn, the overall epidemiology and ecology of pneumococcal strains.

Before PCV introduction, Navajo and White Mountain Apache (N/WMA) Native American communities in the Southwestern US experienced rates of invasive pneumococcal disease (IPD) 2–5 times higher than the general US population [28,29]. Pneumococcal carriage prevalence among N/WMA pre-PCV7 was 50% among all ages and 75% among children <2 years of age, significantly higher than the general population [30]. Thirty-eight percent of all pneumococcal carriage isolates were PCV7 serotypes [30]. After introduction of PCV7, carriage prevalence of PCV7 VT declined, and the rate of IPD among N/WMA caused by VT decreased by 89% [31]. However, carriage prevalence of NVT strains increased, resulting in no overall change in pneumococcal carriage prevalence among children or adults [5]. Also, despite increased NVT carriage there was no corresponding increase in the rate of IPD caused by NVT. After introduction of PCV13 in 2010, carriage of PCV13-specific serotypes declined by 60% among children <5 years of age within the first two years [6]. Yet, overall IPD rates among N/WMA still remain higher than those in the general US population [32].

Here, we analyze a sample of 937 pneumococci collected over 14 years and spanning before, during, and after the introduction of PCV7 and PCV13 vaccines among N/WMA. To understand the evolutionary impact of vaccination and characterize the shift from VT to NVT, we assessed the recombination, evolution, and pneumococcal population history, classified by serotype and by whole-genome sequencing data, across vaccine introduction periods. Furthermore, we investigated metabolic loci variation and pangenome composition over time, with a focus on pneumococcal antigens.

Methods

Study population and pneumococcal isolation

This study included pneumococci isolated from a subset of participants of three prospective, observational cohort studies of pneumococcal carriage among N/WMA families described

elsewhere (hereafter, "parent" studies) [1,6,30]. Briefly, participants living on reservations in the southwest USA were enrolled during three periods: 1998–2001, 2006–2008 and 2010–2012. Nasopharyngeal (NP) swab specimens were obtained during visits to Indian Health Services (IHS) facilities or the participants' home to determine pneumococcal carriage status (S1 Fig) [30]. A random subsample of isolates was selected from each time period, with an oversampling of isolates post-PCV7 (S2 Fig). A single isolate was chosen from each participant; however, previous pneumococcal carriage history was not assessed. With the exception of a subset of isolates collected from 2006–2008, all isolates were obtained from children $\leq$5 years of age.

DNA sequencing, *de novo* assembly, pangenome, and population structure

Genomic DNA from *S. pneumoniae* isolates were sequenced on the Illumina HiSeq, yielding $\geq$30-fold coverage per isolate. Paired-end 100 bp reads were filtered by quality and length. Serotypes were determined by mapping reads to concatenated CPS locus sequences of 93 pneumococcal serotypes using SRST2 [12,33]. Serotypes for isolates identified as serogroup 6 were further resolved using PneumoCaT [34]. Multilocus sequence type (MLST) was determined through a similar approach using SRST2. *De novo* genome assemblies were generated with Velvet [35] and annotated using Prokka v1.11 [36]. After annotation, the pangenome was analyzed with Roary, and a concatenated alignment of clusters of orthologous genes (COGs) shared among $\geq$99% of all strains (i.e., core genome) was abstracted [37]. Pneumococcal population structure was assessed using core genome SNPs with hierBAPS, which was run three times using maximum clustering sizes of 20, 40, and 60 [38]. A maximum likelihood (ML) phylogeny was estimated using RAxML v8.1.5 with GTR+Γ nucleotide substitution model and 100 bootstrap replicates [39]. Sequence clusters (SCs) (i.e., lineages) identified using hierBAPS were annotated on the core genome phylogeny. For the study period during which pediatric and adult isolates were collected (2006–2008), the proportion of isolates by SC was compared between age groups to 10,000 random deviates of a Dirichlet distribution [40].

Reference-based genome assembly and recombination analysis

A subsample of isolates from each SC and 25 publicly available reference genomes were aligned using Parsnp and visualized using Gingr to identify the most appropriate genome for reference-based mapping [41]. The phylogenetically closest genome was selected for reference-based mapping of isolates belonging to that SC. For four out of 27 SCs, a monophyletic match was not available; therefore, we generated references by refining, ordering, and concatenating the best draft assembly in the SC. A second *de novo* assembly was generated with SPAdes and assemblies were then merged using Zorro [42]. After this, SSPACE and GAPFILLER were used to scaffold the assembly and remove Ns [43,44]. Final contigs were ordered using Mauve, manually curated using ACT, and concatenated [45]. Filtered Illumina reads from isolates comprising each SC were mapped to the selected reference using SMALT v0.7.6 and SNPs were identified using SAMtools v1.3.1 [46]. SNPs were filtered requiring a depth of coverage of five and a minimum alternate allele frequency of 0.75. The output was analyzed as previously described to generate whole-genome multiple sequence alignments for each SC [9,47].

Next, we identified recombination among SCs using Gubbins [48]. Gubbins identifies SNPs introduced through recombination and allows censoring for downstream phylogenetic analysis. Results of Gubbins analyses were visualized using Phandango [49]. For SCs in which over 50% of the genome was censored due ancestral recombination events, we either sub-clustered SCs clearly delineated monophyletic clades (e.g., SC19 which was comprised of serotypes 15A and 17F) or removed divergent isolates that were significantly affected by recombination. Sub-clustered SCs were annotated on the ML phylogeny and then reanalyzed with Gubbins.

Analysis of VT and NVT lineage population dynamics

For comparison between vaccine periods, isolates were subdivided into three epochs and six sub-epochs by year of collection: pre-PCV7 sub-epochs 1A (1998) and 1B (1999–2001); post-PCV7 sub-epochs 2A (2006) and 2B (2007–2008); PCV13 sub-epochs 3A (2010) and 3B (2011–2012) (S2 Fig). Collection years were grouped to balance sample sizes among sub-epochs. To determine the representativeness of the genomic sample to the parent studies from which the sample was drawn, we compared the serotype distribution and serotype diversity (Simpson's D) of unique carriage isolates from the three parent studies of pneumococcal carriage [1,6,30], by epoch, to that of the sample. Core genome alignments were generated for isolates in each sub-epoch using Roary, and population genomic statistics including Tajima's *D* [50], Watterson's estimator (Θ_w) [51], and nucleotide diversity were calculated for each period using 0-fold and 4-fold degenerate sites. The ratio of diversity at non-synonymous sites to synonymous sites ($\pi N/\pi S$) was also calculated as a measure of selection. The same statistics were calculated for each SC. Code for calculating population genetic statistics using Roary output is available at http://github.com/c2-d2/Projects/NWMA_Pneumo/.

ML phylogenies of SCs, inferred from recombination-censored alignments, were used to test temporal signal by assessing correlation between strain isolation date and root-to-tip distance. SCs with poor root-to-tip correlation were assessed for residual recombination and phylogenetic signal. SCs determined to have sufficient temporal signal were analyzed with BEAST v1.8.2 [52]. For each SC or sub-SC a combination of strict and relaxed molecular clock models and constant and Gaussian Markov random field (SkyGrid) demographic models [53] were tested using recombination-free SNP alignments, ascertainment bias correction [54,55], and HKY nucleotide substitution model. For SCs in which the coefficient of variation for relaxed molecular clock models was high (i.e., significant rate heterogeneity across the tree), a random local clock (RLC) model was also tested [56]. Markov chain Monte Carlo lengths for each model run ranged from 150 million to 1 billion depending on the size of the SC and length of the SNP alignment. MCMC chains were sampled to obtain 10,000 trees and 10,000 parameter estimates in the posterior distribution. Effective sampling size (ESS) values were assessed to determine sufficient mixing using Tracer v1.6.0, and runs with ESS values of 200 for all parameters were accepted. Marginal likelihood estimates (MLE) were obtained for each model using path-sampling and stepping-stone analysis, and models were compared using Bayes Factors [57,58]. Parameter estimates for the evolutionary rate, root height (i.e., TMRCA), and N_e were obtained from the best-fit model. For SCs in which SkyGrid demographic models were fit, the slope of the N_e change over time was calculated to determine directionality, and the 95% highest posterior density (HPD) was used to determine significance.

Variation in metabolic loci, accessory genome content, and non-capsular antigens among epochs

To assess the impact of PCV7 on the pneumococcal pangenome we compared frequencies of polymorphisms in core genes and accessory genome COGs among sub-epochs, focusing on antigens and metabolic loci for the core genome and on antigens for the accessory genome analysis. We identified metabolic genes using coding sequences found in *S. pneumoniae* reference strain D39 (RefSeq: NC_008533.1) that were annotated as "Metabolism" according to KEGG Orthology (KO) groupings of the KEGG database (http://www.genome.jp/kegg/) and were assigned to a known metabolic pathway (KEGG pathway spd01100). Pangenome analysis using Roary was repeated including D39, and COGs found in the core genome (i.e., present among all 937 taxa) with $\geq$90 BLAST identity to metabolic genes were abstracted. A concatenated alignment of core metabolic COGs was then constructed, and biallelic SNP sites were identified. To

assess changes to the accessory genome, we obtained the binary presence-absence matrix of accessory COGs present in frequencies ranging from 5–95% among all taxa. This frequency range was conservatively selected to mitigate the effect of genome assembly and annotation errors in COG identification. Last, we used a previously described method to identify the variants of 19 polymorphic antigens [15]. These antigens have measurable interactions with the host immune system, and therefore are thought to be under the greatest population level host immune pressure. Ten additional antigens were evaluated (*lysM*, *lytB/C*, *pcpA*, *pcsB*, *phtE*, *piaA*, *piuA*, *psaA*, *SP2027*, *pce*) but were excluded because they were deemed nearly monomorphic due to their low nucleotide diversity.

Using the concatenated nucleotide alignment of metabolic loci and a binary presence absence alignment accessory COGs and antigen variants, ML phylogenies were inferred using RAxML with GTRGAMMA (nucleotide) or GTRCAT (binary) substitution model and 100 bootstrap replicates. The cophenetic (patristic) distances of each phylogeny were read into R, and the meandist function in the package *vegan* was used to calculate within-group distances for three population groupings: serogroup, serotype, and SC. Within-group distances for population stratifications were then compared. For each set of genomic loci (metabolic, accessory COGs, and antigens), frequencies were computed for each of the six sub-epochs. Mean squared errors (MSEs) were then calculated to assess changes in frequencies from Epoch 1A. This was done by subsampling 75 individuals with replacement from each sub-epoch and performing 1000 bootstrap replicates of each comparison (e.g., Epoch 1A vs. 1B, 1A vs. 2A, 1A vs. 2B, and so on). The significance of changes in antigen distributions among epochs was additionally tested by comparing the proportion of antigen variants between Epochs 1–3 to 10,000 random deviates of a Dirichlet distribution.

Ethics statement

The Navajo Nation, White Mountain Apache tribe and the IRBs of the Johns Hopkins Bloomberg School of Public Health, the Navajo Nation and the Phoenix Area IHS approved this study. During the original pneumococcal carriage studies from which these isolates were obtained, written informed consent was obtained from adult participants and from caregivers of child participants. Assent was obtained from children 7–17 years. Isolates were obtained from NP swabs, as previously described, and de-identified for analysis.

Results

Population structure

We analyzed genomic data from a total of 937 pneumococcal carriage isolates collected from N/WMA Native Americans in Southwestern US between 1998 and 2012. All isolates were obtained from children $\leq$5 years of age with the exception of 125 isolates (13.3% of total) collected from individuals 6–76 years of age during 2006–2008. Isolates collected from 1998–2001 (n = 274) were obtained from communities that served as the control for cluster-randomized PCV7 trials and therefore represent a vaccine naïve population. Isolates collected during 2006–2008 (n = 398) represent the post-PCV7 pneumococcal population, and isolates from 2010–2012 (n = 265) were sampled during the implementation of PCV13 (S2 Fig). Whole-genome sequencing data has been deposited in NCBI sequence read archive (SRA) under accession number ERP009399, BioProject PRJEB8327. Individual accession numbers are provided in supplementary file 1.

Pangenome analysis of *de novo* genome assemblies identified 8,674 COGs, of which 1,111 were present in $\geq$ 99% of strains (i.e., the core genome). Analysis of population structure using hierBAPS identified 27 SCs, two of which (SC27 and SC4) were polyphyletic in the ML

phylogeny (Fig 1). SC27 was comprised of low frequency genotypes whereas SC4 contained three distinct monophyletic clades that were bifurcated by branches with low bootstrap support. Based on recombination analysis using Gubbins and assessment of temporal signal (i.e.,

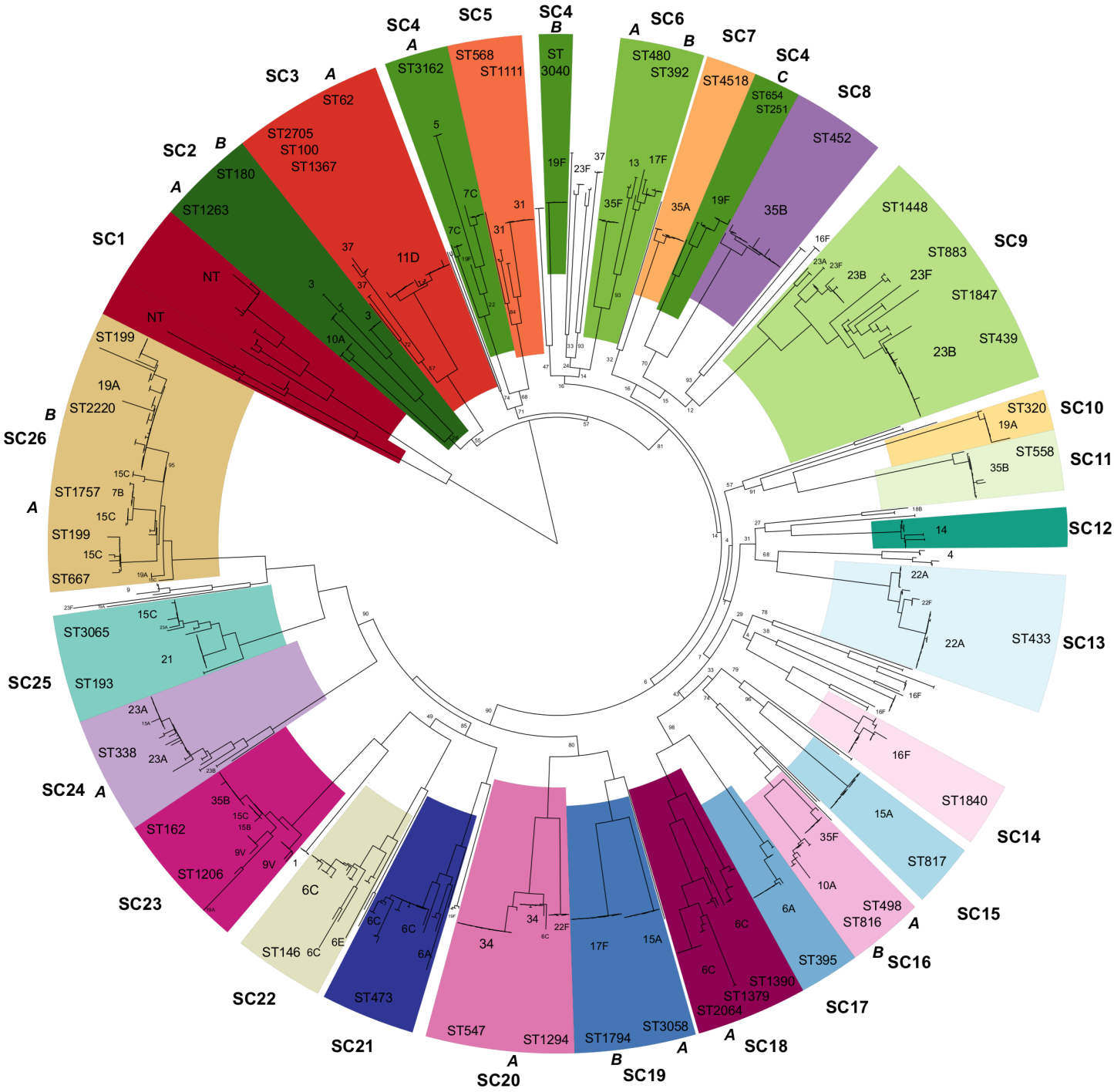

Fig 1. Phylogeny of pneumococcal population structure. Maximum likelihood phylogeny of 937 pneumococcal carriage isolates inferred from an alignment of 1,111 core COGs including 78,525 polymorphic sites using RAxML with GTR+Γ nucleotide substitution model and 100 bootstrap replicates. Clades are colored by sequence cluster (SC), which are labeled on the outside ring. Some SCs are further divided into monophyletic sub-clusters (A, B, C) based on ancestral recombination history. Pneumococcal serotypes and MLST are labeled on each clade and bootstrap values are labeled on internal branches.

https://doi.org/10.1371/journal.ppat.1006966.g001

molecular clock), SC4 as well as 10 other SCs were further subdivided, as it was evident that substantial ancestral recombination events occurred on branches separating dominant monophyletic clades. This subdivision is consistent with the biological definitions of lineages or subpopulations [59,60]. Subsequent analysis focused on 33 SCs or sub-SCs that varied in size from 10 to 71 isolates (Table 1). The proportion of isolates belonging to each SC differed between age groups for only four of 27 SCs, among isolates collected from 2006–2008. SC07 (serotype 35A) and SC15 (serotype 15A) were more common among children ≤5 years of age, 0.8% and 1.6% adults compared to 3.3% and 4.4% children, respectively ($p = 0.03$ and 0.05). SC08

Table 1. Demography and vaccine type composition of pneumococcal sequence clusters. For each sequence cluster (SC) or sub-cluster, the number of isolates, proportion of PCV7 and PCV13 vaccine types, and recombination rate (*r/m*) are listed. The best fit demographic and molecular clock model as determined through comparison of marginal likelihood estimates using BEAST are specified. The *Ne* directionality (constant, exponentially increasing or decreasing) were determined by assessing the slope of the *Ne* during the study period (1998–2012) and 95% Highest Posterior Density (HPD).

SC	Isolates	% PCV7	% PCV13	SNP Sites	Recombination Rate (r/m)	Demographic Model	Clock Model	*Ne* slope (1998–2012)	*Ne* Direction
02-A	19	0.0%	100.0%	4,137	1.49	Constant	Relaxed	-	→
02-B	10	0.0%	0.0%	375	1.17	SkyGrid	Relaxed	0.032	→
03-A	35	0.0%	0.0%	1,048	2.27	SkyGrid	Relaxed	-0.041	→
04-A	19	0.0%	0.0%	1,239	1.65	Constant	Relaxed	-	→
04-B	10	100.0%	0.0%	147	0.01	SkyGrid	Relaxed	-0.027	→
04-C	14	100.0%	0.0%	408	10.31	Constant	Relaxed	-	→
05	21	0.0%	0.0%	1,248	0.93	Constant	Relaxed	-	→
06-A	10	0.0%	0.0%	178	0.04	SkyGrid	Strict	-0.031	→
06-B	11	0.0%	0.0%	679	2.28	SkyGrid	Relaxed	-0.002	→
07	15	0.0%	0.0%	556	1.33	Constant	Relaxed	-	→
08	28	0.0%	0.0%	660	6.09	SkyGrid	Relaxed	-0.020	→
09-A	71	26.8%	0.0%	2,049	4.50	Constant	Relaxed	-	→
10	12	0.0%	100.0%	150	15.03	Constant	Relaxed	-	→
11	18	0.0%	0.0%	370	1.97	SkyGrid	Strict	-0.067	↘
12	13	100.0%	0.0%	293	1.23	Constant	Relaxed	-	→
13	41	0.0%	0.0%	979	2.12	SkyGrid	Relaxed	-0.021	→
14	19	0.0%	0.0%	530	7.99	SkyGrid	Relaxed	0.008	→
15	19	0.0%	0.0%	544	0.05	Constant	Relaxed	-	→
16-A	11	0.0%	0.0%	270	0.36	Constant	Relaxed	-	→
16-B	12	0.0%	0.0%	2,954	2.30	SkyGrid	Relaxed	0.084	→
17	13	0.0%	100.0%	176	0.61	SkyGrid	Random Local	-0.290	↘
18-A	21	0.0%	0.0%	733	5.04	Constant	Relaxed	-	→
19-A	15	0.0%	0.0%	165	0.00	Constant	Relaxed	-	→
19-B	21	0.0%	0.0%	283	0.10	Constant	Relaxed	-	→
20-A	35	0.0%	0.0%	827	3.39	Constant	Random Local	-	→
21	28	0.0%	21.4%	1,561	10.27	Constant	Relaxed	-	→
22	27	0.0%	0.0%	987	14.11	Constant	Relaxed	-	→
23	41	0.0%	2.4%	677	8.24	Constant	Relaxed	-	→
24-A	32	0.0%	0.0%	753	15.00	SkyGrid	Relaxed	0.013	→
25	32	0.0%	0.0%	889	6.33	Constant	Relaxed	-	→
26-AB	84	0.0%	56.0%	2,344	5.40	SkyGrid	Relaxed	-0.059	↘
26-A	40	0.0%	0.0%	923	4.81	SkyGrid	Strict	-0.110	↘
26-B	44	0.0%	100.0%	599	3.83	SkyGrid	Relaxed	-0.020	→
27*	95	18.9%	13.7%	-	-	-	-	-	-

*SC27 is polyphyletic, comprised of several at low frequencies

https://doi.org/10.1371/journal.ppat.1006966.t001

(serotype 35B) and SC26 (serotypes 19A/15C) were more common among adults, 5.6% and 14.4% adults compared to 2.2% and 8.4% children, respectively ($p = 0.05$ and 0.04).

Representativeness of sequenced isolates

For temporal comparison, we divided study periods into three epochs and six sub-epochs (1A/B, 2A/B, 3A/B) (S2 Fig). To verify representativeness of isolates used for genome sequencing in this study, we obtained prevalence data on 3,868 carriage events from children ≤5 years of age in the parent N/WMA carriage studies from which the genomic sample was drawn. This included 1227 events from Epoch 1, 1038 from Epoch 2, and 1603 from Epoch 3. For the major epochs, the proportions of NVT, PCV7, and PCV13 serotypes in our sample were comparable with the serotype dynamics characterized by the three N/WMA parent studies (Fig 2A). The exception was the proportion of NVT and PCV7 VT in Epoch 1, which was due to differences between serological and genomic assignment of serogroup 6 isolates. In Epoch 1, serotypes 6B and 6C were both assigned to serotype 6B by the Quellung reaction used in the parent carriage study. This was subsequently resolved in the current study using a genomic approach to determine serotype, and later carriage studies were able to distinguish 6B from 6C. In pre-PCV Epoch 1, 26.3% of the sample was comprised of PCV7 VT, mostly serotypes 23F, 9V, 14, and 19F. Post-PCV7, the proportion of PCV7 VT in Epoch 2 fell to 1.8%. The prevalence of PCV13 VTs declined steadily from 17.5% in Epoch 1 to 11.3% in Epoch 3. The reduction in PCV13-specific VT after the introduction of PCV7 was likely due to the cross-reactivity of the 6B component of PCV7 with serotype 6A [61], which can be inferred from the elimination of SC17 (serotype 6A) after Epoch 1 (Fig 3).

Population dynamics: Serotypes and lineages

Fluctuations in serotype distribution were reflected in measures of serotype diversity. Simpson's D, which summarizes diversity as the probability that two isolates chosen at random are different, increased from Epoch 1 to 2, reflecting an increase in previously low-frequency NVT serotypes as well as the introduction of previously unobserved serotypes (Fig 2B). Fig 3 illustrates how the composition of the 27 main SCs changed during each of the three epochs. Of two lineages containing PCV7 VT only in Epoch 1, one (SC12) disappeared after vaccination, and another remained, with only PCV7 NVT isolates in Epochs 2 and 3. In SCs containing both PCV7 VT and NVT, the VT lineages were largely eliminated. After Epoch 1, the composition of the pneumococcal population in our sample and parent carriage studies shifted to a predominance of NVT and PCV13 VT, with the largest increases in serotypes 23B and 15C. While in most cases the NVT increases arose from serotypes previously observed in Epoch 1, serotypes belonging to SC10, SC22, and SC24 were not detected until Epoch 2. PCV13 VTs in our sample were not significantly impacted between Epoch 2 and 3. Further comparison of PCV13 implementation data from N/WMA communities during Epoch 3 sampling demonstrated incomplete vaccine coverage and persistence of PCV13 vaccine serotypes (S3 Fig). This finding is consistent with the previous observation that the impact of PCV13 on carriage among underimmunized children was not detected until vaccine coverage in the community reached 58% [6]. This coverage level was not attained until February 2011, at which point 52% of the Epoch 3 sample had been collected. As a result, our assessment of the impact of PCV13 on the overall pneumococcal population was limited.

Population genetic parameters

We used Watterson's theta (Θ_W)–proportional to the number of polymorphic sites—and Tajima's *D* to assess the impact of vaccine on population level genetic diversity and population

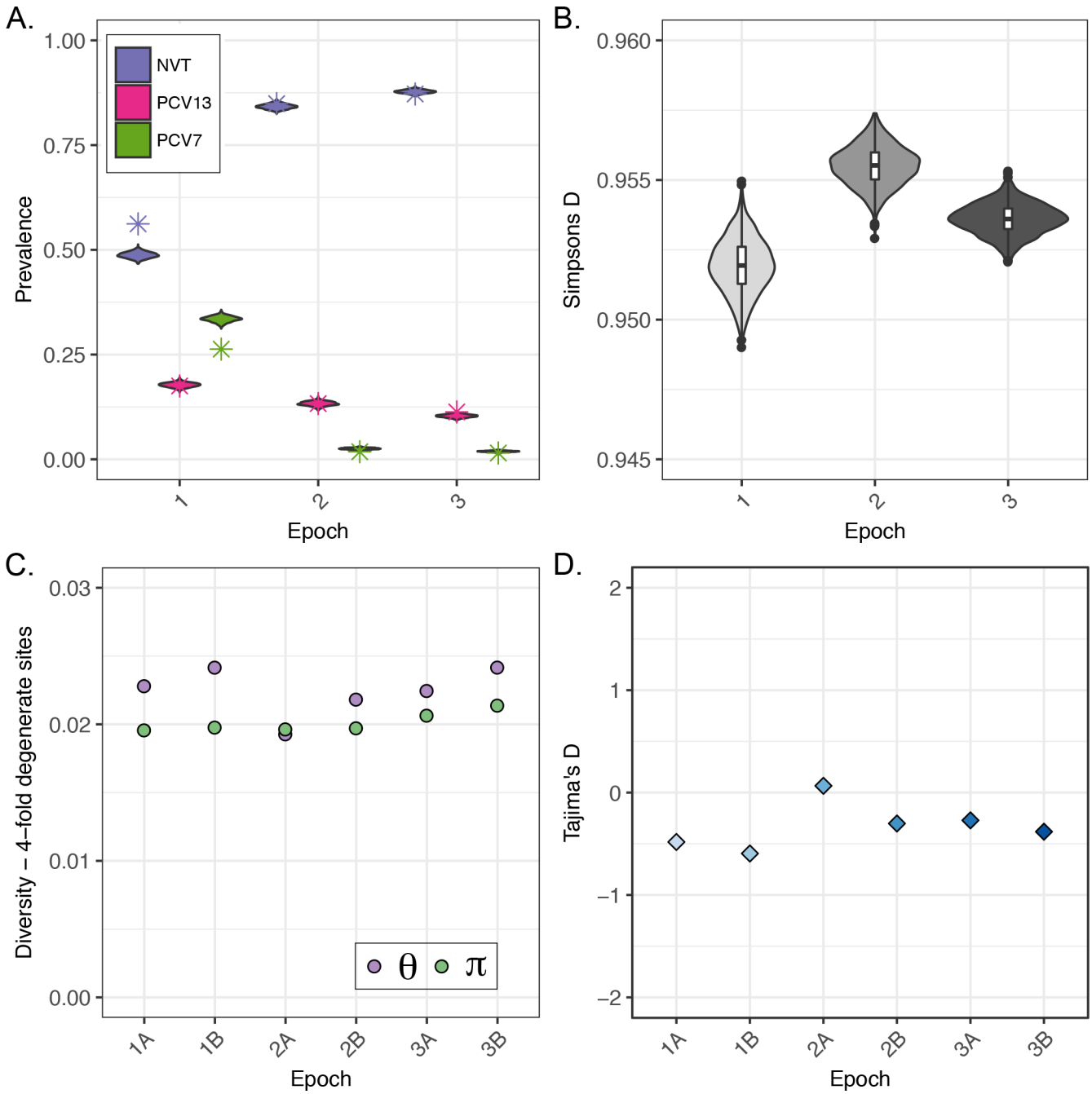

Fig 2. Pneumococcal population dynamics pre- and post-vaccine, 1998–2012. The study periods were subdivided into three epochs and six sub-epochs: Pre-PCV7 [Epoch 1: A (n = 105), B (n = 169)], Post-PCV7 [Epoch 2: A (n = 79), B (n = 319)], and PCV13-Intermediate [Epoch 3: A (n = 119), B (n = 146)]. A.) The proportions of vaccine types (VT) for each epoch from three parent studies (violin) and current study subsample (asterisk). Parent study VT proportions are estimated from serotypes of 3,868 carriage events from previous N/WMA studies [1,6,30]. PCV7 VT include serotypes 4, 6B, 9V, 14, 18C, 19F and 23F. PCV13 vaccine types (minus PCV7 types) include only serotypes 1, 3, 5, 6A, 7F, 19A. Violin plots represent the realization of 1000 bootstrap replicates subsampling with replacement from each epoch. Asterisks represent the point estimates for VT proportions in the current study. B.) Simpson's Diversity Index of serotype diversity across study periods estimated from three parent studies (n = 3,868). This measure summarizes the number and abundance of each serotype. C.) Population genetic measures of diversity, Watterson estimator (Θ_w) and π (average number of pairwise differences), estimated from 4-fold (synonymous) degenerate sites of taxa in sub-epochs of the current study. D.) Population genetic statistic Tajima's D estimated from the core genomes of taxa in sub-epochs of the current study. Negative values of Tajima's D indicate many sites with a rare minor alleles.

https://doi.org/10.1371/journal.ppat.1006966.g002

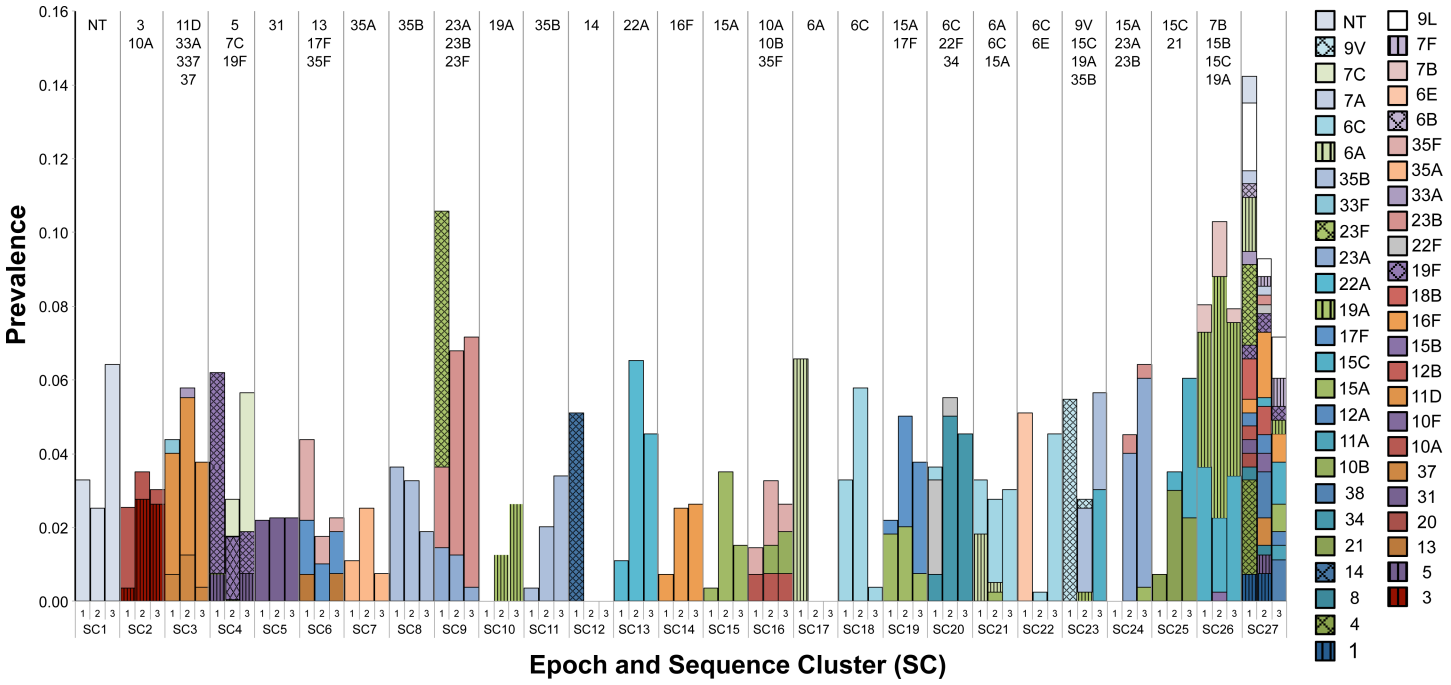

Fig 3. Population structure of pneumococcal carriage isolates among N/WMA. Populations are divided into sequence clusters (SC) based on genomic data and further subdivided into three epochs based on collection date. The bars represent the proportion of population comprising each SC during an epoch and are stratified by serotype composition. Solid bars represent non-vaccine serotypes; checkered hatched pattern PCV7 vaccine types; and vertical line pattern PCV13 vaccine types (those not included in PCV7). Serotypes comprising each SC are also labeled above each column. All SCs except SC4 and SC27 are monophyletic. SC4 is has three distinct sub-clades and SC27 include polyphyletic lineages present at minor frequencies in the population.

https://doi.org/10.1371/journal.ppat.1006966.g003

size. Under neutrality and constant population size, $\Theta_W = 2N_e\mu$, where N_e is the effective population size and μ is the mutation rate [51]. Selective removal of several clusters of related strains, such as lineages or sub-lineages associated with VT, should lead to a reduction in Θ_W. A related measure, Tajima's *D*, tests for evidence of population growth, with negative values suggesting population expansion (due to the presence of rare variants at high frequencies) and positive values suggesting balancing selection or population contraction [50]. Consistent with our expectations, Θ_W decreased from Epoch 1B to 2A, illustrating an overall decrease in pneumococcal genomic diversity, while the average number of pairwise differences (π) was unaffected (Fig 2C). Tajima's *D* values computed for the polymorphic nucleotide sites in the core genome increased from -0.59 in Epoch 1B to 0.07 in 2A, signifying a removal of rare variants consistent with a species-wide population bottleneck (Fig 2D). By Epoch 3B both Θ_W and Tajima's *D* returned to pre-PCV7 levels while π increased. No discernible changes in either measure were associated with PCV13 introduction.

Contributions of population processes

After the population genetic bottleneck induced by PCV7's removal of VT, genetic diversity (i.e., Θ_W) may have been augmented by 1) clonal expansion of NVT lineages due to selection or genetic drift (to increase Θ_W such lineages would have to have been so rare post-bottleneck that they were not sampled), 2) introduction of new lineages, or 3) recombination. We hence examined evidence for each of these among individual SCs. Recombination rates (*r/m*) varied among SCs, ranging from 0 to 15.0, averaging 4.25 (Table 1 and S4 Fig). While coalescent analysis found SCs varied in mutation rates (S5 Fig), there was no significant difference between the median evolutionary rates of NVT and VT SCs (95% CI: -1.06e-06–8.54e-06, $F(1,29)$ =

2.55, $p = 0.12$). Therefore, high evolutionary rates among NVT lineages were not solely responsible for recovering the diversity lost due to the removal of PCV7 VT.

To investigate the contribution of introduction of new lineages or expansion of previously unsampled ones, we estimated the TMRCAs (i.e., lineage age) of SCs. Overall, the median TMRCA was 1955 and ranged from 1839 (SC21: 6A/C ST473) to as recent as 2000 (SC10: 19A ST320) (S6 Fig). Two SCs that were not identified during Epoch 1 sampling emerged following vaccination: SC10 (S7 Fig), which is all type 19A and ST320, and SC24, largely comprised of serotype 23A (S8 Fig) related to PMEN clone Colombia^{23F}-26. Estimated TMRCA for SC10 was 2000 [95% HPD: 1996–2004]. The lineage age, taken together with its low level of genetic diversity ($\Theta_w = 0.0006$) and negative Tajima's *D* value (-2.15), suggests that this SC was introduced after the implementation of PCV7 among southwest Native Americans and is currently experiencing population expansion. SC24 was first identified in 2006 during Epoch 2, but its most recent common ancestor was estimated at 1958 [95% HPD: 1928–1980], near the median TMRCA among all SCs. Considering its prevalence in Epoch 2 and moderate level of diversity ($\Theta_w = 0.003$), it is likely that SC24 was not recently introduced and that its was present in the population before PCV7 but at a sufficiently low frequency not to be sampled until 2006, by which time its frequency may have increased. Furthermore, SC24's low Tajima's *D* value (-1.63) is consistent with population expansion.

No detectable signal of vaccine impact on effective population size

We hypothesized that post-PCV7 changes in pneumococcal populations would be visible as decreases in the effective population size (N_e) of predominantly VT lineages and increases in those of predominantly NVT lineages. The effective population size can be interpreted as the number of genomes contributing offspring to the next generation, and changes in N_e can be used to measure population growth or contraction. Inferring demography among SCs identified that over half (56%) fit constant population size models based on MLEs (Table 1). Furthermore, while the remainder of SCs best fit a fluctuating N_e model (i.e., Skygrid), assessment of N_e trajectories identified only three that were significantly different from a constant size based on HPDs. These three SCs (SC11, SC17, and SC26-A) were found to be decreasing throughout the study period; one was PCV13 VT (SC 17) and two were NVT (SC11 and SC 26-A). To assess bias potentially introduced by removing recombination, we tested the association between recombination rates and inferred demography, which we found to not be significant ($F(1,30) = 0.44$, $p = 0.51$) [62]. Overall, these findings show that the relatively subtle increases in sample frequencies of individual SCs containing NVT are not visible as departures from a constant N_e.

The impact of vaccine on the pneumococcal pangenome

To test the hypothesis that selective removal of PCV7 VT disrupted accessory genome content, we compared accessory size and frequencies of 2370 COGs and 53 variants of 19 antigens between pre-PCV7 Epoch 1 to post-PCV7 epochs. Further, we tested the concurrent effect on metabolic loci by assessing frequencies of 22,434 biallelic SNPs found among 256 metabolic genes present in the core genome. For metabolic loci, accessory COGs, and antigen variants, within-group diversity was minimized when SC population groupings were assigned, compared to serogroup and serotype (S9 Fig). The introduction of PCV7 resulted in an overall reduction in pangenome size, illustrated by the difference in logarithmic pangenome curves for Epochs 2A and 3B (S10 Fig). A comparison of pre-PCV7 Epochs 1A and 1B provided a baseline estimate of stochastic, temporal fluctuations in frequencies in the absence of an effect of vaccine. Plotting COG frequencies in subsequent

epochs demonstrated perturbation in pneumococcal accessory COGs frequencies following introduction of PCV7 (S11 Fig). This perturbation is characterized by the dispersion of frequency scatterplots comparing Epochs 1A vs. 2A [$R^2 = 0.96$, MSE = $8.26x10^{-3}$ (95% CI: $8.32x10^{-3}$–$8.40x10^{-3}$)] and 2B [$R^2 = 0.98$, MSE = 6.65 $x10^{-3}$ (95% CI: $6.60x10^{-3}$–$6.70x10^{-3}$)] (Figs 4 and S11). This effect was also observed when comparing the frequencies of polymorphic antigens and metabolic loci between epochs (Figs 5, S12 and S13). For all sets of genomic loci, MSE in comparison to Epoch 1A are smaller for 1B than for any of the subsequent epochs, illustrating the disruption caused by PCV7. While this observation alone could be explained by drift leading to increasing divergence in frequencies over time, a further observation cannot: in each example, MSEs decreased from Epoch 2 to 3, indicating metabolic loci, accessory COGs, and antigen frequencies were trending back toward pre-PCV7 values (Fig 5). This trend was observed when isolates collected from individuals >5 years of age were removed from Epoch 2 and the analysis repeated. This led us to compare Epoch 3A (post-PCV7/pre-PCV13) to previous sub-epochs to determine whether the pre-PCV7 Epochs 1A/B or the immediately preceding Epoch 2B were better predictors of COG/antigen frequencies. For accessory genome COG frequencies and metabolic loci, Epoch 2B was a better predictor of 3A frequencies; however, for antigens, pre-PCV7 Epoch 1B was the best predictor of Epoch 3A frequencies (S14 Fig). Taken together, we found that antigen variant frequencies largely returned to pre-PCV7 values; however, some perturbations were not resolved (Fig 6). This was due largely to *pspC* groups 1/5 ($p = 0.01$) and *srtH* Var-I ($p = 0.004$), which remained at higher frequencies at Epoch 3, and *rrgA* Var-I ($p<0.001$), which was completely removed from the population.

Discussion

The impact of PCV7 introduction on pneumococcal serotype distributions has been well-characterized in the N/WMA and other communities, but the pneumococcal genome-wide impact has been investigated in fewer populations [3,63]. We studied genomes from a sample spanning the introduction of PCV7 and PCV13, which, based on serotype distribution, were representative of the full set of data from which the sample was drawn. Beyond the expected impact on serotypes, we find the effect of vaccine on the pneumococcal population could be observed as changes in population level diversity, metabolic loci, size of the pneumococcal pangenome, and frequencies of accessory genes including polymorphic antigens. We further illustrate how pneumococcal genomic diversity and frequencies of accessory genome COGs rebounded after the population bottleneck induced by the selective removal of VT lineages by PCV7. These findings help explain how the frequency distribution of polymorphic antigens, for example, largely return to baseline frequencies after being disrupted by vaccine.

The post-PCV7 pneumococcal population in N/WMA saw the complete removal of two SCs and a significant reduction in prevalence of three. The population bottleneck was characterized by changes in levels and patterns of genomic diversity, decreasing Θ_W and increasing Tajima's *D* (Fig 2). Subsequently, the removal of VT pneumococci was counterbalanced by the expansion of SC9 and the emergence of two previously unobserved SCs, SC10 and SC24. In Epoch 2, we identified minor variations in the distribution of SCs by age group for four SCs. As none of the SCs contained PCV7 VT, differences likely resulted from variation in acquired serotype-specific immunity among children and adults [64]. Overall, population structure of SCs was comparable, consistent with pneumococcal transmission dynamics and the wide-ranging impact of the PCV7 vaccine on carriage in children and adults [5]. Despite the changes in the prevalence of SCs over time, no consistent pattern of change in the N_e of these SCs was detectable through coalescent analysis of individual SCs (Table 1). This lack of signal may be

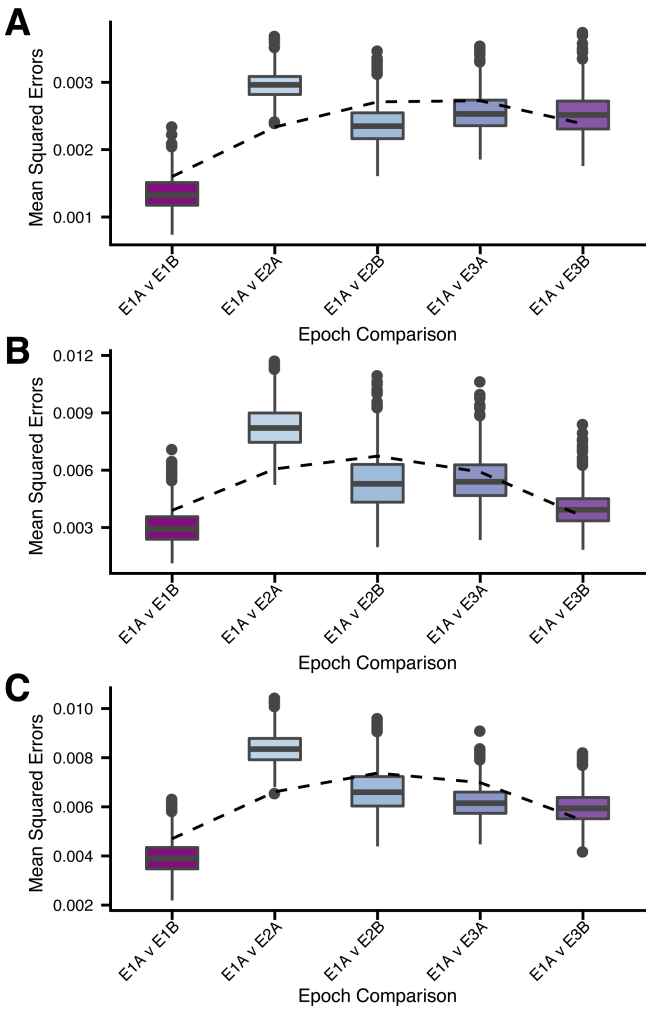

Fig 4. Mean squared errors (MSEs) comparing changes in genomic loci frequencies between Epoch 1A and all subsequent sub-epochs. For each comparison, 75 individuals were subsampled with replacement from each sub-epoch and 1,000 bootstrap replicates were performed. A.) MSEs for sub-epoch comparison of frequencies of 22,434 biallelic SNP sites found among 256 metabolic genes. B.) MSEs for sub-epoch comparison of frequencies of 53 variants of 19 polymorphic antigens. C.) MSEs for sub-epoch comparison of frequencies of 2370 COGs found from 5–95% among all 937 taxa.

https://doi.org/10.1371/journal.ppat.1006966.g004

due to a number of factors. It may be that where vaccine pressure was strong enough to drastically change the population size of an SC, it was eliminated (e.g., SC12), so the temporal signal was lost; where changes were more modest, e.g. in SC including both VT and NVT, the method may have been too insensitive to detect a change. While assessment of N_e did not clearly identify consistent changes, we did detect the post-PCV7 emergence of two SCs. By comparing TMRCA and core genome diversity, we infer that that the first, SC10, appears to have been recently introduced among N/WMA, while the second, SC24, appears to have become detectable due to the vaccine [8,65]. It is worth noting that assessing N_e and other population genetic parameters of pneumococcal lineages makes implicit assumptions about defining SCs as populations and a collection of SCs as a metapopulation, which, to varying degrees, may compete or interact with one another through recombination. Indeed, this definition is more complex and requires consideration of competition, gene flow, and niche overlap among lineages [60,66,67]. Here, we statistically define SCs and find that these populations are often

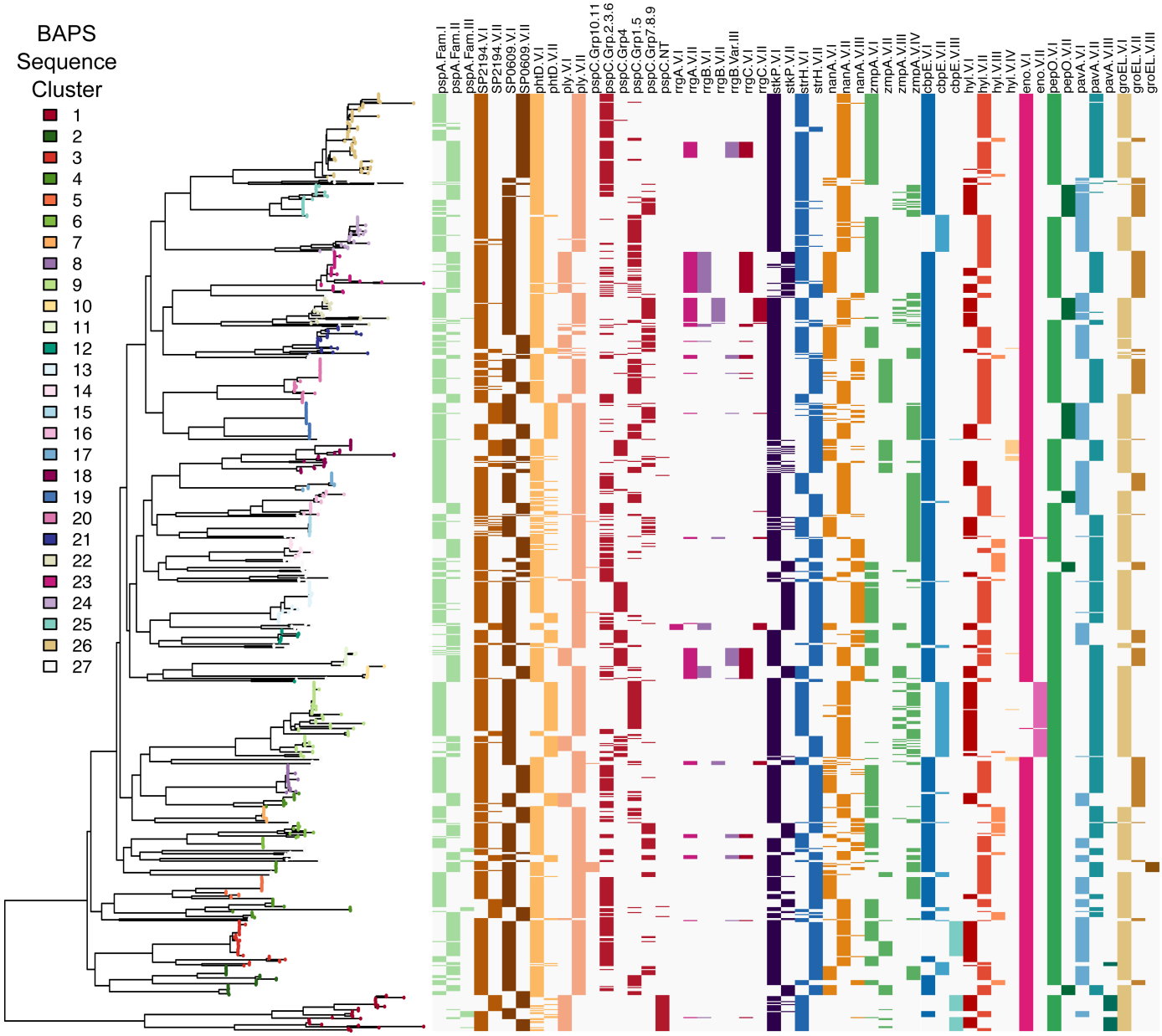

Fig 5. Association of hierBAPS sequence cluster (SC) with polymorphic non-capsular antigens. Maximum likelihood phylogeny of 937 pneumococcal carriage isolates corresponding to Fig 3. Color ramps in legend designate SCs. Antigens include pspA, SP2194, SP0609, phtD, ply, pspC, rrgA/B/C (type 1 pilus islet), stkP, strH, nanA, zmpA, cbpE, hyl, eno, pepO, pavA, and GroEL. Accession numbers for antigen variants have been previously published [15].

https://doi.org/10.1371/journal.ppat.1006966.g005

good predictors of serogroup, metabolic profile, and gene content, thus generally demonstrate genomic coherence consistent with the concept of a bacterial population.

Pneumococcal genomic data from carriage studies in the US are limited [9]. The N/WMA sample provides an opportunity to assess post-vaccine changes in the pneumococcal populations across demographically and geographically varied regions and, at large, the generalizability of bacterial pathogen population dynamics. Comparable analysis of population structure of 616 carriage isolates from Massachusetts collected between 2001 and 2007 found less structure (15 monophyletic SCs (n = 616)) compared to the N/WMA sample (25 monophyletic SCs

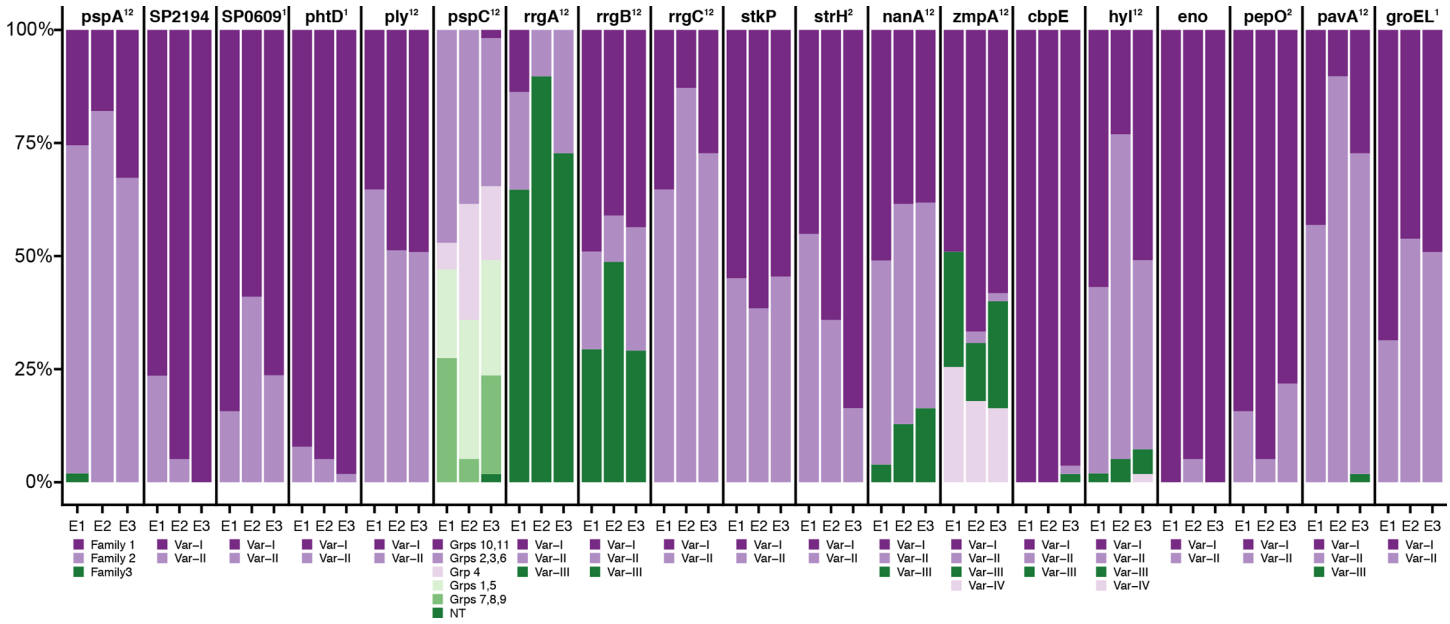

Fig 6. Bar plots comparing proportion of variants among 19 polymorphic antigens between Epochs E1-E3. Each plot represents the change in distribution of polymorphic antigen variants during Epoch 1 (E1), Epoch 2 (E2), and Epoch 3 (E3). Antigens are labeled above each bar plot, and variants of each antigen are colored according to the legend below each plot. Antigen labels are annotated to indicate whether the change in the distribution of frequencies was significant between E1 and E2 ("1"), E2 and E3 ("2"), or both ("12"). All antigens are found among all strains with the exception of rrgA/B/C, which are only present in a subset of taxa. Accession numbers for each variant has been previously published.

https://doi.org/10.1371/journal.ppat.1006966.g006

(n = 937)) [9], and unlike Massachusetts, where the post-PCV7 population emerged largely from the pre-existing serotype diversity, in the N/WMA sample we observed seven previously unidentified serotypes and two entire SCs post-PCV7. Considering carriage data from the larger parent studies, 13 previously unidentified serotypes, excluding 6C, were observed post-PCV7. This difference aside, SC composition and pneumococcal population dynamics were consistent between N/WMA and Massachusetts. For example, SC9 (also SC9 in the Massachusetts study [9]) experienced a near identical population shift post-PCV7 (S15 Fig). This SC, which is comprised of VT 23F and NVTs 23A and 23B, is thought to have arisen through multiple serotype-switching events. In the N/WMA sample, it was one of the most successful in terms of overall prevalence in Epoch 1. As observed in Massachusetts, PCV7 effectively removed 23F isolates from the SC; however, SC9 NVTs subsequently increased 3.5% from Epoch 1 to 3. This shows that these changes were not restricted to the Massachusetts population, but were replicated in a very different setting, and may suggest that SC9 occupies a specific niche. Consistent with this hypothesis, we find that the antigen profiles for VT 23F and the NVT 23B population that replaced it, to be largely consistent with the exception of *zmpA* (S16 Fig). Taken together, we observe similar pneumococcal population dynamics in two geographically and demographically distinct populations that share common vaccine histories, suggesting that response to population shaping processes are relatively consistent.

We find that each SC is defined by a unique profile of metabolic loci, accessory COGs, and antigen variants. These profiles are most resolved at the SC level rather than serotype or serogroup, as the same serotype can be found in multiple SCs due to switching events. Moreover, within an SC, these genomic loci show significant linkage disequilibrium despite appreciable recombination among pneumococci [68]. Consistent with this linkage, we observed a coincident impact of PCV7 on genetic diversity, accessory COG frequencies, polymorphic antigens, and metabolic loci. The population genomic perturbation that resulted from the removal of

PCV7 VT was significantly mitigated by Epoch 3, with frequencies of antigen variants, in particular, returning to pre-vaccine values. A recently proposed model of NFDS provides one putative mechanism for the maintenance of antigen variants and accessory COGs at optimal frequencies [27], and variant-specific host immunity provides a biologically plausible mechanism for NFDS on antigens. Early evidence of balancing selection among pneumococci was the reemergence of strains possessing a type 1 pilus after PCV7 significantly reduced piliated serotypes [69]. In the current study, we also observe the reemergence of type 1 pilus driven by serotype 19A ST320 (SC10). And while the observation with the pilus involved a change in presence-absence frequency, we now see the same dynamic extending to frequencies of antigen variants. Yet, due to linkage it is difficult to untangle which loci are being acted upon by selection and which reflect hitchhiking. Alternatively, balancing selection could be acting upon metabolic loci which are important to niche adaption and have been implicated in post-vaccine metabolic shifts [18]. In an effort to identify which loci may be driving post-vaccine success of SCs, we considered the frequencies of metabolic loci, accessory COG, and antigen variants separately. We find PCV13-era (Epoch 3) frequencies of polymorphic antigens are better predicted by pre-PCV7 (Epoch 1) frequencies than the immediately preceding period. In addition, we observe that overall COG frequencies seemed to trend toward pre-PCV7 norms with increasing time since vaccine introduction, while frequencies of metabolic loci remained disrupted. This does not rule out variation in metabolic loci or other core genes such as GroEL as driving forces for pneumococcal population structure [70]; however, it remains difficult to assign fitness differences based on observed genetic variation. For example, two SCs may be divergent in metabolic loci but capable of exploiting the same metabolic niche.

Previous models have proposed that recombination is the mechanism underlying the post-vaccine shift in metabolic, virulence, and antigenic loci [18]. However, we argue that in our sample, recombination has likely not had enough time to shuffle antigen variants or other COGs into different genomic backgrounds. For example, if we again consider the replacement of VT 23F by NVT 23B belonging to SC9, we observe that both populations possess similar antigenic profiles (S16 Fig). Yet, the TMRCA of the 23B population, and all associated recombination events, predate the introduction of PCV7 (S15 and S16 Figs). This illustrates that at least in this case, an existing population possessing a near identical antigenic profile contributed to the rebalancing of the distribution of antigen variants in the overall pneumococcal population. Overall, the pneumococcal accessory genome is comprised of varying types of MGE (e.g., phages and antigens), and it is likely that their distribution is controlled by many different, yet interconnected, processes [17]. As such, the underlying dynamics maintaining antigenic variant and accessory COG frequencies require further investigation.

Through comprehensive analysis of serotype distribution and population dynamics of *S. pneumoniae* spanning the introduction of PCV7 and PCV13 among N/WMA communities, we gain a broad understanding of the impact of vaccine on population structure, serotype distribution, and pangenome composition. After the introduction of PCV7, we observe clonal replacement of VT by NVT as well as clonal expansion of vaccine-associated serotypes during a period when carriage prevalence remained unchanged. Further, we show PCV7 significantly disrupted accessory COG frequencies, including frequencies of polymorphic antigens important to host-pathogen interactions. This post-PCV7 period of 'flux' in serotype diversity and accessory COG distribution was normalized by Epoch 3, demonstrating rapid adaption to the post-vaccine landscape. Moving forward, continued genomic surveillance will be required to monitor the emergence of new lineages and to investigate the impact on post-PCV13 pneumococcal populations. Last, as balancing selection appears to be an integral component of pneumococcal adaption and considerable serotype-lineage-accessory genome linkage exists, the joint effect of removal of vaccine serotypes and linked antigens on host-susceptibility to extant

lineages merits further study, as it has significant implications for the future of protein-based pneumococcal vaccines. For example, protein-based vaccines should consider the prevalence of polymorphic variants across host populations and either include multiple variants of the same antigen or target those in greatest frequency.

Supporting information

S1 Fig. Collection sites among Navajo and White Mountain Apache Native American communities in Southwestern United States.
(EPS)

S2 Fig. Samples by epoch and sub-epoch. For temporal comparison, isolates were subdivided into three epochs and six sub-epochs based on the year of collection: pre-PCV7 Epoch 1, sub-epochs A (1998) and B (1999–2001); post-PCV7 Epoch 2, sub-epochs A (2006) and B (2007–08); PCV13-Intermediate Epoch 3, sub-epochs A (2010) and B (2011–2012). An effort was made to balance the number of samples in each sub-epoch.
(EPS)

S3 Fig. PCV13 vaccine coverage, proportion PCV13 VT, and isolate sampling by month and year during Epoch 3 (2010–2012). Reservation-wide PCV13 vaccine coverage (primary y-axis–black) was obtained from Indian Health Services data. Coverage is based on three doses in children 7-<12 months or 1+ dose in children from 12–59 months. The proportion of PCV13 VT carriage in children <5 years of age was calculated using data from the PCV13 parent study (n = 1,603) (primary y-axis–red). The bar-plot represents the collection month of isolates used in Epoch 3 of the present study (secondary y-axis–grey).
(EPS)

S4 Fig. Recombination rate comparison among pneumococcal sequence clusters (SC). The recombination rate (r/m) is the ratio of SNPs predicted to have been imported through recombination events compared to those introduced through mutation. The ratio ρ/θ represent the relative rate of recombination events to mutations on a phylogenetic branch. The asterisks denote SCs that contain *PCV7 vaccine-types and **PCV13 vaccine types.
(EPS)

S5 Fig. Evolutionary rate comparison among pneumococcal sequence clusters (SC). SC and sub-clusters are listed on the x-axis with evolutionary rates (SNPs per site per year) and 95% highest posterior density (HPD) displayed for each. Rates were inferred through population genomic modeling using BEAST v1.8.2. Demographic and molecular clock models selected through model comparison are listed in Table 1. The median evolutionary rate was 1.30E-06 SNPs/sire/year (range: 6.02E-08–7.02E-06).
(EPS)

S6 Fig. Date of the most recent common ancestor (TMRCA) among pneumococcal sequence clusters (SC). TMRCA and 95% HPD were estimated using BEAST v1.8.2. Demographic and molecular clock models selected through model comparison are listed in Table 1. The median TMRCA was 1958 (range: 1839–2000).
(EPS)

S7 Fig. Bayesian maximum clade credibility phylogeny of SC10. The phylogeny was inferred using BEAST v1.8.2 with relaxed molecular clock, Skygrid demographic model, and HKY nucleotide substation model. Node bars represent the 95% Highest Posterior Density (HPD) and the branch are colored based on posterior probability. A branch with a posterior

probability of >0.80 is considered well supported. The density plot represents the marginal probability distribution of the root-height of the tree (i.e., TMRCA).
(EPS)

S8 Fig. Bayesian maximum clade credibility phylogeny of SC24. SC24 is comprised largely of serotype 23A isolates. However, three serotype 23B1 isolates, which were basal to the 23A clade, were excluded for this analysis. Of note, among these isolates there was a serotype switch between a 23A capsule and 15A capsule. A.) The phylogeny was inferred using BEAST v1.8.2 with relaxed molecular clock, Skygrid demographic model, and HKY nucleotide substation model. Node bars represent the 95% Highest Posterior Density (HPD) and the branch are colored based on posterior probability. A branch with a posterior probability of >0.80 is considered well-supported. B.) The density plot represents the marginal probability distribution of the root-height of the tree (i.e., the most recent common ancestor (TMRCA)). C.) Bayesian Skygrid plot of effective population size (log *Ne*) over time. The grey area represents the 95% HPD of the *Ne* estimate. While effective population size appears to be increasing over time, the increase is not significant based on the HPD range.
(EPS)

S9 Fig. Comparison of population stratifications and within-group genetic distance. Within-grouping genetic distance, measured as the patristic distance between pairs of isolates in the maximum likelihood (ML) phylogeny, between three population stratifications: serogroup, serotype, and sequence cluster (SC). ML phylogenies were inferred from 1) presence absence-alignment of 2371 accessory genome clusters of orthologous genes (COGs), 2) 22,424 biallelic polymorphic sites found in 272 metabolic genes present in the core genome, and 3) presence-absence alignment of non-capsular antigen variants. Error bars represent the weighted mean plus or minus the weighted standard error of the estimate. In all instances, within-group distances were significantly less than between-group mean distances.
(EPS)

S10 Fig. Pangenome comparison of epochs and sub-epochs. Pangenome comparison of strains from respective epochs are overlaid, demonstrating variation in accessory clusters of orthologous genes (COG) content after introduction of PCV7. Dotted lines represent the increase in accessory COGs as the number of genomes is increased. Epochs are individually colored. The trajectory of the "total genes" reflects the overall diversity of COGs during each epoch. E1A and E1B represent the pre-PCV7 COG diversity.
(EPS)

S11 Fig. Scatterplots comparing accessory genome clusters of orthologous genes (COG) frequencies among epochs. Each plot represents a comparison of COG frequencies during Epoch 1 compared to subsequent epochs. As a control, pre-PCV7 Epochs 1A and 1B are compared to each other. The sum of residuals was obtained from regressing the COG frequencies of one epoch on another using a linear regression model. Pearson's correlation coefficients are also presented for each comparison.
(EPS)

S12 Fig. Scatterplots comparing metabolic loci frequencies among epochs. Each plot represents a comparison of frequencies of 22,434 biallelic SNPs found among 256 metabolic genes present in the core genome during Epoch 1 compared to subsequent epochs. As a control, pre-PCV7 Epochs 1A and 1B are compared to each other. The sum of residuals was obtained from regressing the frequencies of one epoch on another using a linear regression model. Pearson's

correlation coefficients are also presented for each comparison.
(EPS)

S13 Fig. Scatterplots comparing polymorphic protein antigen variant frequencies (n = 19) among epochs. Each plot represents a comparison of protein antigen variants frequencies during Epoch 1 compared to subsequent epochs. As a control, pre-PCV7 Epochs 1A and 1B are compared to each other. Variants of each polymorphic protein antigen are colored similarly. The sum of residuals was obtained from regressing the variant frequencies of one epoch on another using a linear regression model. Pearson's correlation coefficients are also presented for each comparison.
(EPS)

S14 Fig. Predictability of Epoch 3 genomic loci frequencies by previous sub-epochs. Overlaid histograms of mean squared errors (MSEs) from 1,000 bootstrap replicates comparing three sets of genomic loci frequencies between Epoch 3A and preceding sub-epochs 1B, 2A, and 2B. Histograms are colored according to the epoch comparison. Dashed lines represent median MSE values for each comparison. MSE values closer to zero indicated better prediction of Epoch 3A frequencies. A.) MSEs for sub-epoch comparison of frequencies of 22,434 biallelic SNP sites found among 256 metabolic genes. B.) MSEs for sub-epoch comparison of frequencies of 53 variants of 19 polymorphic, non-capsular antigens. For protein antigens, Epoch 3A vs. 1B frequencies had lower MSEs [$5.78x10^{-3}$ (SE $3.92x10^{-5}$)] than Epoch 3A vs. 2B [$6.06x10^{-3}$ (SE $5.45x10^{-}$)], indicating E1B pre-PCV7 frequencies were better predictors of post-PCV7 frequencies that than the immediately preceding period (E2B). C.) MSEs for sub-epoch comparison of frequencies of 2370 COGs found from 5–95% among all taxa (n = 937).
(EPS)

S15 Fig. Bayesian maximum clade credibility phylogeny of SC9. SC9 is comprised of serotypes 23B, 23F, and 23A. A.) The phylogeny was inferred using BEAST v1.8.2 with relaxed molecular clock, Skygrid demographic model, and HKY nucleotide substation model. Node bars represent the 95% Highest Posterior Density (HPD) and the branch are colored based on posterior probability. A branch with a posterior probability of >0.80 is considered well supported. B.) The density plot represents the marginal probability distribution of the root-height of the tree (i.e., the most recent common ancestor (TMRCA)). C.) Bayesian Skygrid plot of effective population size (log *Ne*) over time, and the grey area represents the 95% HPD of the *Ne* estimate. The effective population size remained constant over the study period.
(EPS)

S16 Fig. Comparison of polymorphic protein antigens and recombination events among VT and NVT serotypes comprising heirBAPS sequence cluster (SC) 9. Maximum likelihood phylogeny of 75 pneumococcal serogroup 23 carriage isolates. Heatmap illustrates the serotype and protein antigens profile including pspA, SP2194, phtD, pspC, stkP, strH, and nanA. Accession numbers for protein variants have been previously published [47]. Recombination events are presented in the right half of the figure, visualized linearly in reference to the D39 genome. VT serotype 23F was replaced by NVT serotype 23B. While 23A isolates have a varying protein antigen profile, 23F and 23B are largely comparable. Of note, multiple recombination events within this SC resulted in variations in protein antigen profile composition; however, recombination events on the major 23B clade predate the introduction of PCV7.
(EPS)

S1 Table. Table of accession numbers and associated metadata.
(XLSX)

Acknowledgments

We would like to acknowledge the core informatics, library-making, and sequencing teams at the Wellcome Trust Sanger Institute. Furthermore, the authors express their appreciation to the individuals in the Navajo and White Mountain Apache communities who participated in the studies. We also gratefully acknowledge the dedicated efforts of the Center for American Indian Health field staff who collected these data over many years.

Author Contributions

Conceptualization: Taj Azarian, Lindsay R. Grant, Laura L. Hammitt, Raymond Reid, Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

Data curation: Taj Azarian, Lindsay R. Grant, Laura L. Hammitt, Claudette M. Thompson, Stephen D. Bentley.

Formal analysis: Taj Azarian, Brian J. Arnold, Claudette M. Thompson, Stephen D. Bentley.

Funding acquisition: Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

Investigation: Taj Azarian, Laura L. Hammitt, Raymond Reid, Mathuram Santosham, Robert Weatherholtz, Novalene Goklish, Claudette M. Thompson, Stephen D. Bentley, Katherine L. O'Brien, Marc Lipsitch.

Methodology: Taj Azarian, Lindsay R. Grant, Laura L. Hammitt, Stephen D. Bentley, Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

Project administration: Mathuram Santosham, Robert Weatherholtz, Novalene Goklish, Claudette M. Thompson, Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

Resources: Laura L. Hammitt, Raymond Reid, Mathuram Santosham, Novalene Goklish, Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

Supervision: William P. Hanage, Marc Lipsitch.

Validation: Taj Azarian, William P. Hanage, Marc Lipsitch.

Visualization: Taj Azarian.

Writing – original draft: Taj Azarian, Marc Lipsitch.

Writing – review & editing: Lindsay R. Grant, Brian J. Arnold, Laura L. Hammitt, Raymond Reid, Mathuram Santosham, Robert Weatherholtz, Novalene Goklish, Claudette M. Thompson, Stephen D. Bentley, Katherine L. O'Brien, William P. Hanage, Marc Lipsitch.

References

1. O'Brien KL, Moulton LH, Reid R, Weatherholtz R, Oski J, Brown L, et al. Efficacy and safety of seven-valent conjugate pneumococcal vaccine in American Indian children: group randomised trial. Lancet (London, England). 2003; 362: 355–61. https://doi.org/10.1016/S0140-6736(03)14022-6 PMID: 12907008
2. Black S, Shinefield H, Fireman B, Lewis E, Hansen JR, Elvin L, et al. Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Pediatr Infect Dis J. 2000; 19: 187–95. PMID: 10749457
3. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013; 45: 656–63. https://doi.org/10.1038/ng.2625 PMID: 23644493
4. Waight PA, Andrews NJ, Ladhani SN, Sheppard CL, Slack MPE, Miller E. Effect of the 13-valent pneumococcal conjugate vaccine on invasive pneumococcal disease in England and Wales 4 years after its

introduction: an observational cohort study. Lancet Infect Dis. 2015; 15: 535–543. https://doi.org/10.1016/S1473-3099(15)70044-7 PMID: 25801458

5. Scott JR, Millar EV, Lipsitch M, Moulton LH, Weatherholtz R, Perilla MJ, et al. Impact of more than a decade of pneumococcal conjugate vaccine use on carriage and invasive potential in Native American communities. J Infect Dis. 2012; 205: 280–8. https://doi.org/10.1093/infdis/jir730 PMID: 22128315
6. Grant LR, Hammitt LL, O'Brien SE, Jacobs MR, Donaldson C, Weatherholtz RC, et al. Impact of the 13-Valent Pneumococcal Conjugate Vaccine on Pneumococcal Carriage Among American Indians. Pediatr Infect Dis J. 2016; 35: 907–914. https://doi.org/10.1097/INF.0000000000001207 PMID: 27171679
7. Hanage WP, Bishop CJ, Huang SS, Stevenson AE, Pelton SI, Lipsitch M, et al. Carried pneumococci in Massachusetts children: the contribution of clonal expansion and serotype switching. Pediatr Infect Dis J. 2011; 30: 302–8. https://doi.org/10.1097/INF.0b013e318201a154 PMID: 21085049
8. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. Lancet. 2011; 378: 1962–73. https://doi.org/10.1016/S0140-6736(10)62225-8 PMID: 21492929
9. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013; 45: 656–63. https://doi.org/10.1038/ng.2625 PMID: 23644493
10. Pilishvili T, Lexau C, Farley MM, Hadler J, Harrison LH, Bennett NM, et al. Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. J Infect Dis. 2010; 201: 32–41. https://doi.org/10.1086/648593 PMID: 19947881
11. Whitney CG, Farley MM, Hadler J, Harrison LH, Bennett NM, Lynfield R, et al. Decline in Invasive Pneumococcal Disease after the Introduction of Protein–Polysaccharide Conjugate Vaccine. N Engl J Med. Massachusetts Medical Society; 2003; 348: 1737–1746. https://doi.org/10.1056/NEJMoa022823 PMID: 12724479
12. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. Public Library of Science; 2006; 2: e31. https://doi.org/10.1371/journal.pgen.0020031 PMID: 16532061
13. Weinberger DM, Trzciński K, Lu Y-J, Bogaert D, Brandes A, Galagan J, et al. Pneumococcal capsular polysaccharide structure predicts serotype prevalence. PLoS Pathog. Public Library of Science; 2009; 5: e1000476. https://doi.org/10.1371/journal.ppat.1000476 PMID: 19521509
14. Darrieux M, Goulart C, Briles D, Leite LC de C. Current status and perspectives on protein-based pneumococcal vaccines. Crit Rev Microbiol. Informa Healthcare; 2015; 41: 190–200. https://doi.org/10.3109/1040841X.2013.813902 PMID: 23895377
15. Azarian T, Grant L, Georgieva M, Hammitt L, Reid R, Bentley S, et al. Pneumococcal protein antigen serology varies with age and may predict antigenic profile of colonizing isolates. J Infect Dis. 2016; jiw628. https://doi.org/10.1093/infdis/jiw628
16. Wilson R, Cohen JM, Reglinski M, Jose RJ, Chan WY, Marshall H, et al. Naturally Acquired Human Immunity to Pneumococcus Is Dependent on Antibody to Protein Antigens. Mitchell TJ, editor. PLOS Pathog. Saunders; 2017; 13: e1006137. https://doi.org/10.1371/journal.ppat.1006137 PMID: 28135322
17. Croucher NJ, Campo JJ, Le TQ, Liang X, Bentley SD, Hanage WP, et al. Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. Proc Natl Acad Sci. National Academy of Sciences; 2017; 114: E357–E366. https://doi.org/10.1073/pnas.1613937114 PMID: 28053228
18. Watkins ER, Penman BS, Lourenço J, Buckee CO, Maiden MCJ, Gupta S. Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. PLoS Pathog. Public Library of Science; 2015; 11: e1005034. https://doi.org/10.1371/journal.ppat.1005034 PMID: 26181911
19. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, et al. Metabolic Resource Allocation in Individual Microbes Determines Ecosystem Interactions and Spatial Dynamics. Cell Rep. Cell Press; 2014; 7: 1104–1115. https://doi.org/10.1016/J.CELREP.2014.03.070 PMID: 24794435
20. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 2010; 11: R107. https://doi.org/10.1186/gb-2010-11-10-r107 PMID: 21034474
21. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun. Nature Publishing Group; 2014; 5: 1–12. https://doi.org/10.1038/ncomms6471 PMID: 25407023
22. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. Hughes D, editor. PLoS Genet. Public Library of Science; 2016; 12: e1006280. https://doi.org/10.1371/journal.pgen.1006280 PMID: 27618184

23. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. Nature Publishing Group; 2017; 2: 17040. https://doi.org/10.1038/nmicrobiol.2017.40 PMID: 28350002

24. Croucher NJ, Mostowy R, Wymant C, Turner P, Bentley SD, Fraser C. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. PLoS Biol. 2016;14. https://doi.org/10.1371/journal.pbio.1002394 PMID: 26934590

25. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, et al. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81. J Bacteriol. 2009; 191: 1480–9. https://doi.org/10.1128/JB.01343-08 PMID: 19114491

26. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. bioRxiv. 2017;6. https://doi.org/10.7554/eLife.26255 PMID: 28742023

27. Corander Jukka, Fraser Christophe, Gutmann Michael U., Arnold Brian, Hanage William P., Bentley Stephen D., Marc Lipsitch NJC. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nat Ecol Evol. Nature Publishing Group; 2017; In press. https://doi.org/10.1038/s41559-017-0337-x

28. Cortese MM, Wolff M, Almeido-Hill J, Reid R, Ketcham J, Santosham M. High Incidence Rates of Invasive Pneumococcal Disease in the White Mountain Apache Population. Arch Intern Med. American Medical Association; 1992; 152: 2277. https://doi.org/10.1001/archinte.1992.00400230087015 PMID: 1444688

29. O'Brien KL, Shaw J, Weatherholtz R, Reid R, Watt J, Croll J, et al. Epidemiology of Invasive *Streptococcus pneumoniae* among Navajo Children in the Era before Use of Conjugate Pneumococcal Vaccines, 1989–1996. Am J Epidemiol. Oxford University Press; 2004; 160: 270–278. https://doi.org/10.1093/aje/kwh191 PMID: 15258000

30. Millar E V, O'Brien KL, Zell ER, Bronsdon MA, Reid R, Santosham M. Nasopharyngeal carriage of *Streptococcus pneumoniae* in Navajo and White Mountain Apache children before the introduction of pneumococcal conjugate vaccine. Pediatr Infect Dis J. 2009; 28: 711–6. https://doi.org/10.1097/INF.0b013e3181a06303 PMID: 19593248

31. Weatherholtz R, Millar EV, Moulton LH, Reid R, Rudolph K, Santosham M, et al. Invasive pneumococcal disease a decade after pneumococcal conjugate vaccine use in an American Indian population at high risk for disease. Clin Infect Dis. 2010; 50: 1238–46. https://doi.org/10.1086/651680 PMID: 20367225

32. Mosso KL, Grant LR, Weatherholtz RC, Campbell J, Donaldson C, Dallas J. Impact of the 13-valent pneumococcal conjugate vaccine on a population at high risk for invasive pneumococcal disease. Program and abstracts of the 9th International Symposium on Pneumococci and Pneumococcal Disease. 2014. pp. 9–13.

33. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. BioMed Central; 2014; 6: 90. https://doi.org/10.1186/s13073-014-0090-6 PMID: 25422674

34. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, et al. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ. PeerJ Inc.; 2016; 4: e2477. https://doi.org/10.7717/peerj.2477 PMID: 27672516

35. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18: 821–9. https://doi.org/10.1101/gr.074492.107 PMID: 18349386

36. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30: 2068–9. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

37. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31: btv421. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

38. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. Mol Biol Evol. Oxford University Press; 2013; 30: 1224–1228. https://doi.org/10.1093/molbev/mst028 PMID: 23408797

39. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005; 21: 456–63. https://doi.org/10.1093/bioinformatics/bti191 PMID: 15608047

40. Minka T. Estimating a Dirichlet distribution. Technical report, MIT; 2000.

41. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014; 15: 524. https://doi.org/10.1186/s13059-014-0524-x PMID: 25410596

42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19: 455–77. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

43. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011; 27: 578–9. https://doi.org/10.1093/bioinformatics/btq683 PMID: 21149342

44. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012; 13: R56. https://doi.org/10.1186/gb-2012-13-6-r56 PMID: 22731987

45. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics. 2009; 25: 2071–3. https://doi.org/10.1093/bioinformatics/btp356 PMID: 19515959

46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

47. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog. 2012; 8: e1002745. https://doi.org/10.1371/journal.ppat.1002745 PMID: 22719250

48. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2014; gku1196-. https://doi.org/10.1093/nar/gku1196 PMID: 25414349

49. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomicsbioRxiv. 2017;

50. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123: 585–95. PMID: 2513255

51. Watterson GA. The homozygosity test of neutrality. Genetics. 1978; 88: 405–17. PMID: 17248803

52. Drummond AJ, Suchard M a, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 1–5. https://doi.org/10.1093/molbev/mss075

53. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard M a. Improving bayesian population dynamics inference: a coalescent-based model for multiple Loci. Mol Biol Evol. 2013; 30: 713–24. https://doi.org/10.1093/molbev/mss265 PMID: 23180580

54. Gray RR, Tatem AJ, Johnson J a, Alekseyenko A V, Pybus OG, Suchard M a, et al. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant *Staphylococcus aureus* ST239 genome-wide data within a bayesian framework. Mol Biol Evol. 2011; 28: 1593–603. https://doi.org/10.1093/molbev/msq319 PMID: 21112962

55. Harris SR, Feil EJ, Holden MTG, Quail M a, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327: 469–74. https://doi.org/10.1126/science.1182395 PMID: 20093474

56. Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 2010; 8: 114. https://doi.org/10.1186/1741-7007-8-114 PMID: 20807414

57. Baele G, Lemey P, Vansteelandt S. Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. BMC Bioinformatics. 2013; 14: 85. https://doi.org/10.1186/1471-2105-14-85 PMID: 23497171

58. Kass R. Bayes factors. J Am Stat Assoc. 1995; 90: 773–795.

59. Polz MF, Alm EJ, Hanage WP. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 2013; 29: 170–5. https://doi.org/10.1016/j.tig.2012.12.006 PMID: 23332119

60. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014; 22: 235–47. https://doi.org/10.1016/j.tim.2014.02.006 PMID: 24630527

61. Väkeväinen M, Eklund C, Eskola J, Käyhty H. Cross-Reactivity of Antibodies to Type 6B and 6A Polysaccharides of *Streptococcus pneumoniae*, Evoked by Pneumococcal Conjugate Vaccines, in Infants. J Infect Dis. Oxford University Press; 2001; 184: 789–793. https://doi.org/10.1086/322984

62. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Mol Biol Evol. 2016; msw048. https://doi.org/10.1093/molbev/msw048 PMID: 26931140

63. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014; 46: 305–9. https://doi.org/10.1038/ng.2895 PMID: 24509479

64. Cobey S, Lipsitch M. Niche and Neutral Effects of Acquired Immunity Permit Coexistence of Pneumococcal Serotypes. Science (80-). 2012; 335: 1376–1380. https://doi.org/10.1126/science.1215947 PMID: 22383809

65. Lipsitch M. Interpreting results from trials of pneumococcal conjugate vaccines: a statistical test for detecting vaccine-induced increases in carriage of nonvaccine serotypes. Am J Epidemiol. 2001; 154: 85–92. PMID: 11427408

66. Andam CP, Challagundla L, Azarian T, Hanage WP, Robinson DA. Population Structure of Pathogenic Bacteria. Genetics and Evolution of Infectious Diseases. 2nd ed. Elsevier; 2017. p. 51.

67. Lawson DJ. Populations in statistical genetic modelling and inference. Population in the human sciences: Concepts, models, evidence. OUP Oxford; 2015. pp. 108–130.

68. Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, et al. Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. Genetics; 2018; genetics.300662.2017. https://doi.org/10.1534/genetics.117.300662 PMID: 29330348

69. Regev-Yochay G, Hanage WP, Trzcinski K, Rifas-Shiman SL, Lee G, Bessolo A, et al. Re-emergence of the type 1 pilus among *Streptococcus pneumoniae* isolates in Massachusetts, USA. Vaccine. NIH Public Access; 2010; 28: 4842–4846. https://doi.org/10.1016/j.vaccine.2010.04.042 PMID: 20434550

70. Lourenço J, Watkins ER, Obolski U, Peacock SJ, Morris C, Maiden MCJ, et al. Lineage structure of *Streptococcus pneumoniae* may be driven by immune selection on the groEL heat-shock protein. Sci Rep. Nature Publishing Group; 2017; 7: 9023. https://doi.org/10.1038/s41598-017-08990-z PMID: 28831154