



HAL
open science

Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci

Taj Azarian, Patrick Mitchell, Maria Georgieva, Claudette M Thompson, Amel Ghouila, Andrew J Pollard, Anne von Gottberg, Mignon Du Plessis, Martin Antonio, Brenda A Kwambana-Adams, et al.

► To cite this version:

Taj Azarian, Patrick Mitchell, Maria Georgieva, Claudette M Thompson, Amel Ghouila, et al.. Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. PLoS Pathogens, 2018, 14 (11), pp.e1007438. 10.1371/journal.ppat.1007438 . pasteur-02006050

HAL Id: pasteur-02006050

<https://riip.hal.science/pasteur-02006050>

Submitted on 4 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

RESEARCH ARTICLE

Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci

Taj Azarian^{1*}, Patrick K. Mitchell¹, Maria Georgieva¹, Claudette M. Thompson¹, Amel Ghouila², Andrew J. Pollard³, Anne von Gottberg⁴, Mignon du Plessis⁴, Martin Antonio⁵, Brenda A. Kwambana-Adams⁵, Stuart C. Clarke^{6,7}, Dean Everett⁸, Jennifer Cornick⁹, Ewa Sadowy¹⁰, Waleria Hryniewicz¹⁰, Anna Skoczynska¹⁰, Jennifer C. Moisi¹¹, Lesley McGee¹², Bernard Beall¹², Benjamin J. Metcalf¹², Robert F. Breiman¹³, PL Ho¹⁴, Raymond Reid¹⁵, Katherine L. O'Brien¹⁵, Rebecca A. Gladstone¹⁶, Stephen D. Bentley¹⁶, William P. Hanage¹



1 Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States of America, **2** Institut Pasteur de Tunis, LR11IPT02, Laboratory of Transmission, Control and Immunobiology of Infections (LTCII), Tunis-Belvédère, Tunisia, **3** Oxford Vaccine Group, Department of Paediatrics, University of Oxford; NIHR Oxford Biomedical Research Centre, Centre for Clinical Vaccinology and Tropical Medicine (CCVTM), Churchill Hospital, Oxford, United Kingdom, **4** Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa, **5** Medical Research Council Unit The Gambia, Fajara, The Gambia, **6** Faculty of Medicine and Institute for Life Sciences and Global Health Research Institute, University of Southampton, Southampton, United Kingdom, **7** NIHR Southampton Biomedical Research Centre, Southampton General Hospital, Southampton, United Kingdom, **8** Queens Research Institute, University of Edinburgh, Edinburgh, United Kingdom, **9** Institute of Infection and Global Health, University of Liverpool, Liverpool, United Kingdom, **10** National Medicines Institute, Warsaw, Poland, **11** Pfizer Vaccines, Medical Development, Scientific and Clinical Affairs, Paris, France, **12** Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, **13** Global Health Institute, Emory University, Atlanta, Georgia, United States of America, **14** Department of Microbiology, Queen Mary Hospital University of Hong Kong, Hong Kong, People's Republic of China, **15** Center for American Indian Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **16** Wellcome Trust, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

* tazarian@hsph.harvard.edu

OPEN ACCESS

Citation: Azarian T, Mitchell PK, Georgieva M, Thompson CM, Ghouila A, Pollard AJ, et al. (2018) Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci. *PLoS Pathog* 14(11): e1007438. <https://doi.org/10.1371/journal.ppat.1007438>

Editor: Timothy J. Mitchell, University of Birmingham, UNITED KINGDOM

Received: July 5, 2018

Accepted: October 25, 2018

Published: November 26, 2018

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: Whole-genome sequencing data and/or draft genome assemblies are electronically available in public databases and accession numbers listed in supplementary file 1. Supplementary file 1 also contains all available metadata for isolates used in this study. BEAST XML code used in coalescent analysis has been made available at <https://github.com/c2-d2/Projects/tree/master/Serotype%203>.

Funding: The study was funded in part by National Institute of Allergy and Infectious Disease award to

Abstract

Streptococcus pneumoniae serotype 3 remains a significant cause of morbidity and mortality worldwide, despite inclusion in the 13-valent pneumococcal conjugate vaccine (PCV13). Serotype 3 increased in carriage since the implementation of PCV13 in the USA, while invasive disease rates remain unchanged. We investigated the persistence of serotype 3 in carriage and disease, through genomic analyses of a global sample of 301 serotype 3 isolates of the Netherlands^{3–31} (PMEN31) clone CC180, combined with associated patient data and PCV utilization among countries of isolate collection. We assessed phenotypic variation between dominant clades in capsule charge (zeta potential), capsular polysaccharide shedding, and susceptibility to opsonophagocytic killing, which have previously been associated with carriage duration, invasiveness, and vaccine escape. We identified a recent shift in the CC180 population attributed to a lineage termed Clade II, which was estimated by Bayesian coalescent analysis to have first appeared in 1968 [95% HPD: 1939–1989] and increased in prevalence and effective population size thereafter. Clade II isolates are divergent from the pre-PCV13 serotype 3 population in non-capsular antigenic composition, competence, and

WPH (R01 AI106786). The Global Pneumococcal Sequencing project is funded by The Bill and Melinda Gates Foundation (Grant OPP1034556). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. AS has received assistance to attend scientific meetings and honoraria for lecturing funded from Pfizer and from GlaxoSmithKline, and Participated in Advisory Board of GlaxoSmithKline and Pfizer. WPH has consulted for Antigen Discovery Inc. JCM is employed by Pfizer, Inc. This does not alter our adherence to all PLOS Pathogens policies on sharing data and materials. The other authors have declared that no competing interests exist.

antibiotic susceptibility, the last of which resulting from the acquisition of a Tn916-like conjugative transposon. Differences in recombination rates among clades correlated with variations in the ATP-binding subunit of Clp protease, as well as amino acid substitutions in the *comCDE* operon. Opsonophagocytic killing assays elucidated the low observed efficacy of PCV13 against serotype 3. Variation in PCV13 use among sampled countries was not independently correlated with the CC180 population shift; therefore, genotypic and phenotypic differences in protein antigens and, in particular, antibiotic resistance may have contributed to the increase of Clade II. Our analysis emphasizes the need for routine, representative sampling of isolates from diverse geographic regions, including historically under-sampled areas. We also highlight the value of genomics in resolving antigenic and epidemiological variations within a serotype, which may have implications for future vaccine development.

Author summary

Streptococcus pneumoniae is a leading cause of bacterial pneumoniae, meningitis, and otitis media. Despite its inclusion in the most recent pneumococcal conjugate vaccine, PCV13, serotype 3 remains epidemiologically important globally. We investigated the persistence of serotype 3 using whole-genome sequencing data from 301 isolates collected among 24 countries from 1993–2014. Through phylogenetic analysis, we identified three distinct lineages within a single clonal complex, CC180, and found one that has recently emerged and grown in prevalence. We then compared genomic difference among lineages as well as variations in pneumococcal vaccine use among sampled countries. We found that the recently emerged lineage, termed Clade II, has a higher prevalence of antibiotic resistance compared to other lineages, diverse surface protein antigens, and a higher rate of recombination, a process by which bacteria can uptake and incorporate genetic material from its surroundings. Differences in vaccine use among sampled countries did not appear to be associated with the emergence of Clade II. We highlight the need for routine, representative sampling of bacterial isolates from diverse geographic areas and show the utility of genomic data in resolving epidemiological differences within a pathogen population.

Introduction

Pneumococcal disease caused by *Streptococcus pneumoniae* remains a significant cause of morbidity and mortality even in the era of effective conjugate vaccines, which protect against up to 13 of the nearly 100 known serotypes [1]. Serotype is based on biochemical and antigenic properties of the polysaccharide capsule, a key virulence factor encoded by the capsular (*cps*) locus and the target of the pneumococcal conjugate vaccines (PCV). Among those serotypes covered by the 13-valent pneumococcal conjugate vaccine (PCV13), serotype 3 is considered highly invasive and is associated with a high risk of mortality in both observational studies and animal models [2–5]. However, the serotype-specific effectiveness of PCV13 against serotype 3 remains uncertain. There has been little reduction in serotype 3 disease compared to disease due to other vaccine serotypes following implementation of PCV13 in the USA (Fig 1) [6]. In addition, data from a randomized controlled trials on nasopharyngeal carriage [7], post-licensure studies of PCV13 on invasive pneumococcal disease (IPD) [8], and national surveillance

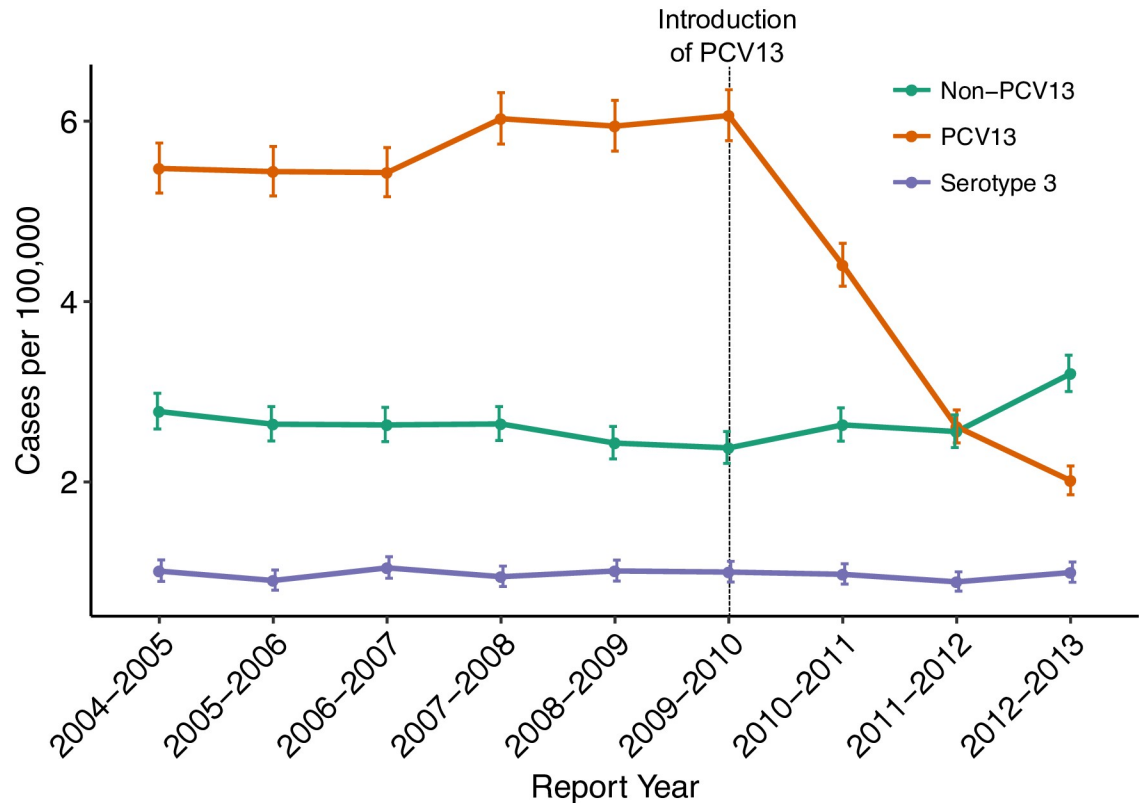


Fig 1. Changes in the incidence of invasive pneumococcal disease (IPD) from 2004 through 2013 among all ages and vaccine status in the United States based on CDC Active Bacterial Core (ABC) surveillance data [6]. Rates of IPD expressed as cases per 100,000 population are on the y-axis, and calendar year of surveillance are on the x-axis. The green line represents the IPD rate for the five most common non-PCV13 serotypes; purple line, the rate for serotype 3 only; orange line, the rate of (PCV13 –serotype 3) serotypes containing (1, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, and 23F). Further information can be obtained from the CDC’s ABC surveillance (www.cdc.gov/abcs/).

<https://doi.org/10.1371/journal.ppat.1007438.g001>

data on IPD from England and Wales [9,10], among other studies have shown little vaccine effect on serotype 3.

Pneumococci can be divided into lineages, often referred to as sequence clusters, based on statistical analysis of nucleotide diversity in the core genome [11–13]. In general, lineages are often comprised of strains possessing the same multi-locus sequence typing (MLST) and serotype. However, whereas MLST profiles are unlikely to appear in multiple lineages, the same serotype may be distributed across the population due to the transfer of the genes encoding the capsule between lineages by homologous recombination [14]. MLSTs can be further grouped into clonal complexes (CCs) [15]. For serotype 3, MLST data have shown that the majority of isolates globally are a single CC of closely related genotypes (CC180) also known as the Netherlands³-31 (PMEN31) clone [16–19], with the exception of Africa where non-CC180 MLST types are prevalent [20].

However, more recent genomic studies have shown that CC180 contains multiple distinct lineages [18,21], which MLST is not sufficiently discriminating to detect. While one of these lineages (termed ‘Clade I’) accounts for almost all previously studied genomes, the overwhelming majority (92%) of these isolates came from European collections, and might not be representative of the global pneumococcal population.

Following roll-out of PCV13, an unexpected increase in the carriage prevalence of serotype 3 was noted among Massachusetts children less than seven years of age [22]. Initial genomic analysis of these carriage isolates showed a change in the serotype 3 population [23]. Prior to vaccination the majority of isolates fell into the previously described 'Clade I', while following PCV13 introduction, the observed increase in carriage prevalence was largely due to isolates which, while still CC180, were drawn from a more diverse population that has been poorly represented in prior samples. Phenotypic variations within CC180 and how they may relate to vaccine efficacy have not been investigated and remain poorly understood.

To investigate the population structure and evolutionary history of serotype 3 CC180 pneumococci, we conducted a genomic analysis of 301 serotype 3 CC180 isolates from carriage and disease collected from 24 countries over 20 years. Further, we assessed multiple phenotypic features linked to the epidemiology of serotype 3, and which might contribute to vaccine escape including reaction to PCV13 antisera, and how they vary in the CC180 population [24–26].

Methods

Study population and epidemiological data

The sample consisted of 301 CC180 (PMEN31) genomes collected from carriage (n = 68), invasive disease (n = 231), and unknown clinical manifestations (n = 2) between 1993–2014. Whole-genome sequencing (WGS) data for this sample were obtained from a previous analysis of serotype 3 described elsewhere (n = 82) [18], on-going studies of carriage in Massachusetts, USA (n = 27) [12], carriage studies in the Southwest USA (n = 14) [27], the Bacterial Isolate Genome Sequence database (BIGSDB) (n = 10), and the Global Pneumococcal Sequencing (GPS) project (<http://www.pneumogen.net/gps/>) (n = 168). The Harvard IRB approved Massachusetts and Southwest USA carriage studies described elsewhere [12,18,28,29]. The contributions of isolates by GPS sites were approved by the respective institution's IRB. These data as well as publicly available and previously published data were anonymized for our analysis. Available data included the year and country of isolation, isolation source, and limited patient demographic data. To assess the temporal shift in prevalence of the three CC180 clades, we tested the significance of changes in the proportion of isolates by clade for each sampling year by comparing yearly proportions to 1000 random deviates of a Dirichlet distribution [30]. If the temporal change in proportion was found to be significant, then the directionality of the change was assessed. Current and past PCV (PCV7, PCV10, and PCV13) use data for countries where the serotype 3 CC180 sample was collected were queried from International Vaccine Access Center (IVAC) VIEW-hub website (www.view-hub.org, accessed April 5, 2017). Fisher's exact test was used to test the association between countries that introduced PCV and serotype 3 clade emergence based on phylogenetic demography (See phylogenetic analysis methods). Further, among countries that introduced PCV13, the correlation between the date of introduction and changes in serotype 3 demography was assessed using Pearson's correlation coefficient. All statistical analysis and figure generation was performed using Rstudio v1.0.143 with R v3.3.1.

Genomic analysis

For GPS isolates, raw sequencing data and *de novo* assemblies were downloaded from <http://www.pneumogen.net/gps/>. GPS *de novo* assemblies were generated using a Velvet pipeline as previously described [31]. The remaining sequencing data were downloaded from the NCBI SRA database (See [S1 File](#) for a list of accession numbers). If only draft assemblies were available, paired-end 150 bp reads were generated using the BBDMap's RandomReads script. *De*

de novo assemblies for non-GPS isolates were generated using SPAdes v3.10 as previously described [18,32]. Assemblies were then annotated using Prokka v1.12 [33] and pangenome analysis was conducted using Roary [34]. Assemblies constructed using simulated reads were checked for variation in assembly metrics, annotation, and SNP diversity, in order to detect any evidence of sequencing and assembly errors. Variants for 13 polymorphic pneumococcal protein antigens—including pneumococcal surface proteins C and A (*pspC* and *pspA*)—were determined by mapping raw reads to an antigen variant database using SRST2 as previously described [27]. Quality filtered and trimmed reads were mapped to *S. pneumoniae* OXC141 (NCBI Reference Sequence: NC_017592), a Clade I, serotype 3 ST180/CC180 carriage isolate, using SMALT v0.7.6. SNPs were called using SAMtools v1.3.1 [35], and were filtered requiring $QUAL > 50$, depth of coverage ≥ 5 and a minimum alternate allele frequency ≥ 0.75 [12,36]. Gubbins v2.1 was used to assess recombination, and coding sequences (CDSs) impacted by recombination blocks were annotated and plotted using Circos [37,38]. Last, we assessed non-synonymous mutations and recombination in the *comCDE* operon, which has previously been implicated in the low competence of CC180 compared to other pneumococcal lineages [18].

Phylogenetic analysis

A maximum likelihood (ML) phylogeny was inferred from a recombination-censored SNP alignment using RAxML v8.2.1 with an ASC_GTRGAMMA nucleotide substitution model, Lewis ascertainment bias correction, and 100 bootstrap replicates [39]. The ML phylogeny was rooted using strain AP200 (accession # CP002121), a serotype 11A, ST62 *S. pneumoniae* invasive isolate, which was found to be immediately basal to the CC180 clade [40]. After assigning isolates to major clades, Gubbins was run independently on the sequence alignments from each clade to identify putative recombination events. ML phylogenies from recombination-censored alignments were used to test temporal signal by assessing correlation between strain isolation date and root-to-tip distance. To reduce bias in coalescent analysis due to differences in sampling over the study period, isolates were subsampled using a uniform probability from each year and spatial location (country and region), as recommended in [41]. To verify temporal signal, we performed date randomization tests on the down-sampled datasets, whereby the sampling times of sequences are shuffled and evolutionary rates are compared to correctly assigned time estimates (see S1 Methods) [42,43]. For major Clade I (hereon referred to as Clade I- α) and Clade II, coalescent analysis was performed using BEAST v1.8.4 (see S1 Methods) [44]. Parameter estimates for the evolutionary rate, root height, and N_e were obtained from the best-fit model and compared between clades. Last, to infer the ancestral geographic location and migration history of Clade II, we used the structured coalescent implemented in Beast 2.4.4 specifying the region of collection as the geographic location for each tip [45,46]. Additionally, BEAST XML code has been made available (<https://github.com/c2-d2/Projects>).

Phenotypic and genotypic antibiotic resistance

Antimicrobial resistance (AMR)-associated genes and SNPs were identified with ARIBA using ARG-ANNOT and CARD databases [47,48]. Penicillin MICs were predicted using WGS data to type transpeptidase domains of penicillin binding proteins [49]. Genotypic antibiotic resistance was validated among strains with available broth dilution data using published CLSI breakpoints for penicillin, chloramphenicol, erythromycin, clindamycin, and tetracycline. For isolates that possessed multiple AMR-associated genes, annotated *de novo* assemblies were investigated to identify conjugative transposons. Transposons were identified through review of *de novo* assembly annotations and their presence/absence among strains was confirmed by mapping sequencing reads to transposons as described above.

Assessment of capsular polysaccharide variations

First, we assessed variations in the CPS loci by abstracting the region from reference-based genome assemblies and generating an ML phylogeny using RAxML with GTRGAMMA substitution model and 100 bootstrap replicates. The mean pairwise nucleotide difference was also calculated enforcing pairwise deletion of missing sites. CPS loci were manually inspected for recombination events using the output from Gubbins to determine the potential role in clade variation.

To investigate phenotypic variations related to the serotype 3 CC180 capsular polysaccharide between Clades I- α and II and potentially explain the recent emergence of Clade II, we assessed surface charge (zeta potential), capsular release, and opsonophagocytic killing. Zeta potentials among representative Clade I- α ($n = 5$) and Clade II ($n = 3$) serotype 3 strains from Massachusetts were compared to *S. pneumoniae* serotype 3 ST378 laboratory strain WU2 and Δ CPS WU2 as previously described [25]. Capsular release, a mechanism by which type 3 CPS interferes with antibody-mediated killing and gains protection by anti-CPS antibodies, was compared among Clade I- α ($n = 3$) and Clade II ($n = 3$) serotype 3 isolates from Massachusetts to strain WU2 as previously described [50]. To assess the efficiency of antisera against serotype 3 CC180 to opsonize pneumococci for uptake and killing by differentiated polymorphonuclear leukocytes, we used an opsonophagocytic killing assay (OPKA) [51]. The killing assays were performed at multiple dilutions using PCV13 antisera and antisera generated from type 3 polysaccharide (PS) (see [S1 Methods](#)) [52]. All Clade I- α and II strains studied in phenotypic assays are marked on the ML phylogeny.

Results

Population structure

The serotype 3 CC180 isolates ($n = 301$) included in this sample came from 24 countries from North America (38.9%), Western Europe (17.9%), Asia (14.6%), Eastern Europe (14.3%), Africa (7.6%), and South America (6.3%) collected from 1993–2014. The ML phylogeny, inferred from a core genome alignment of 15,327 SNP sites, improved the resolution of the three major lineages identified in previous work. Clade I and the previously described Clade II together form a single monophyletic lineage, distinct from the previously described Clade III ([Fig 2A](#)). Applying an out-group roots the phylogeny on the branch between those strains previously described as Clades II and III, indicating that the previous clade naming scheme was inaccurate. To reflect this, and the fact that we now find two rather than three monophyletic lineages, we refer to Clades I- α and I- β (formerly Clades I and II) and Clade II (formerly Clade III). Clade I- α isolates made up the majority of the sample (68.4%), but a significant proportion is made up of Clade I- β (12.3%) and Clade II (19.3%). The expanded sample now shows a single deep branching lineage containing Clade I- β , which is polyphyletic and subtends Clade I- α . Clade II is further subdivided into three well-supported subclades that are distinct in terms of genome content (see below). Following removal of regions that were inferred to have been introduced by recombination, the nucleotide diversity of Clade II was significantly greater than that of Clade I- α [mean pairwise SNP distance 98.0 (SE 2.1) vs. 112.7 (SE 3.6)].

Phylogeography and country-level vaccine use history

The proportion of isolates belonging to Clade II has significantly increased over time from 1999–2001 (11%) to 2011–2014 (41%) (4.2, 95% CI: 1.9–9.0, $p < 0.0001$) with the largest increase occurring in North America ([Fig 2B](#)). During the same time Clade I- α has

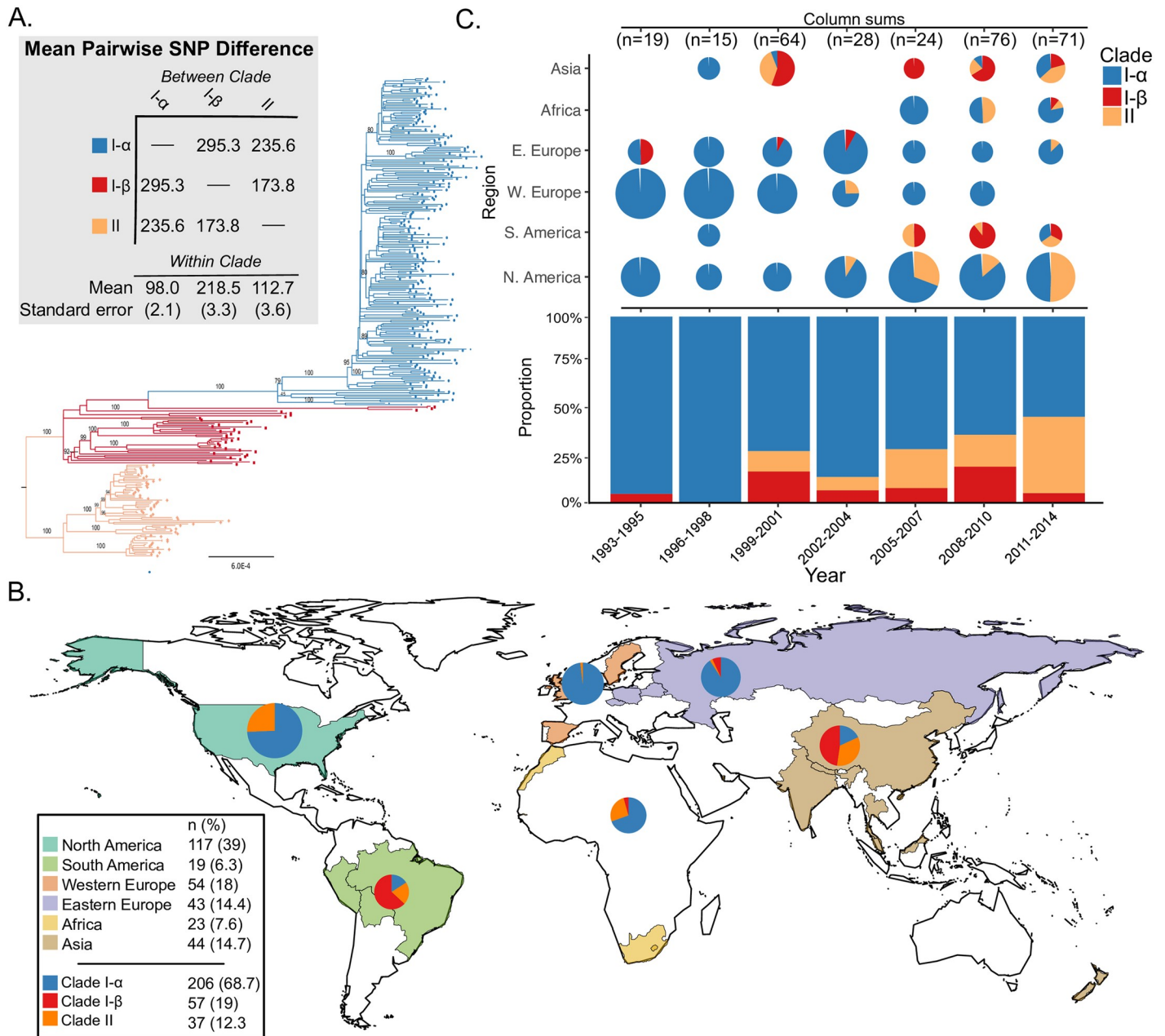


Fig 2. **A)** Rooted maximum likelihood phylogeny of *S. pneumoniae* serotype 3 CC180 isolates ($n = 301$) inferred from an alignment of 15,327 SNPs present in the core genome. Phylogeny was out-group rooted using strain AP200 (accession # CP002121) a serotype 11A, ST62 *S. pneumoniae* invasive isolate, which was found to be immediately basal to the CC180 clade in a global phylogeny of pneumococcal reference genomes and therefore represented the closest out-group. Bootstrap values from 100 replicates are labelled on major clades and epidemiologically relevant sub-clades. Additional values can be found in the supplemental phylogeny. Mean pairwise within- and between-clade SNP difference are presented in the grey shaded box with colors corresponding to the phylogeny. **B.)** World map illustrating sampled countries and regions with respective proportion of isolates belonging to Clade I- α , I- β , and II. Countries are colored according to region and pie charts represent the proportion of isolates belonging to major serotype 3 clades. The size of the pie chart is scaled to the proportion of strains sampled from each region. Mapping was performed using R package *maps* v3.3.0. **C)** Proportion of clade membership by three-year collection window. The proportion of clade membership by region over-time is displayed on the top of the figure. Pie charts are scaled by the number of isolates sampled from a geographic region by time window (i.e., column). The overall proportion of clades for each time window is presented on the bottom of the figure.

<https://doi.org/10.1371/journal.ppat.1007438.g002>

Table 1. Clade composition of CC180 serotype 3 isolates by geographic location. Percentages are based on 295 isolates with available collection date and country.

Clade	Eastern Europe (n = 43)	Western Europe (n = 54)	North America (n = 117)	South America (n = 15)	Africa (n = 23)	Asia (n = 43)	Total
Clade I- α (n = 204)	90.7%	98.1%	74.4%	6.7%	69.6%	18.6%	69.2%
Clade I- β (n = 35)	7.0%	-	-	73.3%	4.3%	46.5%	11.9%
Clade II (n = 56)	2.3%	1.9%	25.6%	20.0%	26.1%	34.9%	19.0%

<https://doi.org/10.1371/journal.ppat.1007438.t001>

significantly decreased, and Clade I- β remained largely unchanged. In the present sample, Clade II was first reported in Asia in 1999, but is now globally distributed, making up a large proportion of samples from Asia, Africa, and North and South America (Fig 2B). However, Clade II was only observed twice among 97 European strains (Table 1), first appearing in 2003. Clade I- α was the most prevalent in the sample and found to be globally distributed, while Clade I- β isolates appear more common in samples from South America and Asia (Fig 2B).

Clades I- α and II had significant temporal signal, identified by root-to-tip date correlation (S1 and S2 Figs). A date randomization test also showed significant temporal signal for Clade II, but date randomizations for Clade I- α failed to reach ESS values >200 despite chain length (S3 Fig). Model comparison using Bayes factors calculated from MLEs identified both Clades I- α and II fit a GMRF SkyGrid demographic model and relaxed molecular clock (S1 Table). In addition, the exponential demographic model was preferred to the constant for Clade II but rejected Clade I- α . Evolutionary rates were not significantly different between the two clades; however, the 95% Highest Posterior Density (HPD) for Clade II was wider (S4 Fig), owing to a greater coefficient of variation ($S = 0.38$ vs 0.21 for Clades II and I- α , respectively) and indicating higher rate variation among branches. Clade II was significantly younger, with an estimated most recent common ancestor (TMRCA) of 1968 [95% HPD: 1939–1989] (S1 Table). The effective population size (N_e) for Clades I- α and II have been increasing, as demonstrated by the SkyGrid N_e plots (Fig 3) and rejection of the constant population size demographic model (S1 Table). Further, the exponential model for Clade II suggested the population was exponentially increasing (mean exponential growth rate = 0.054 [95% HPD: 5.11×10^{-3} , 0.11]).

Phylogeographic migration models of Clade II using the structured coalescent achieved sufficient mixing (i.e., ESS values >200) for all parameters; however, posterior probabilities for migration rates between geographic regions were not significant (<0.30). Therefore, we were unable to infer the ancestral geographic locations of Clade II isolates. We noted, however, that in both ML and Bayesian time-scaled phylogenies, the Asian clade made up of isolates from Hong Kong is proximally basal to the major North American clade (Figs 4 and 5). Further, isolates from Hong Kong possess a unique Tn916-like transposon shared only with isolates found in the dominant North American clade (Fig 5) (see below).

Clade II was found in 11 of 24 countries, 10 of which introduced PCV during our study period (Fig 6). There was no association between countries that introduced PCV13 and the observation of an isolate belonging to Clade II (Fisher's Exact, 4.2 95% CI: 0.3–239.9, $p = 0.3$). There was also no correlation between year of vaccine implementation and the year of first identification of an isolate belonging to Clade II (Pearson's correlation = 0.17, 95% CI: -0.52–0.72, $p = 0.64$). In general, there was no clear pattern among countries in vaccine introduction and Clade II presence. However, in the USA, where annual data was most complete, yet sparse, there was a significant change in the proportion of Clade II and Clade I- α isolates when comparing the four years prior to the introduction of PCV13 to the four years after (2005–2009 vs. 2010–2014, Fisher's Exact, 3.6 95% CI: 1.3–10.7, $p = 0.01$).

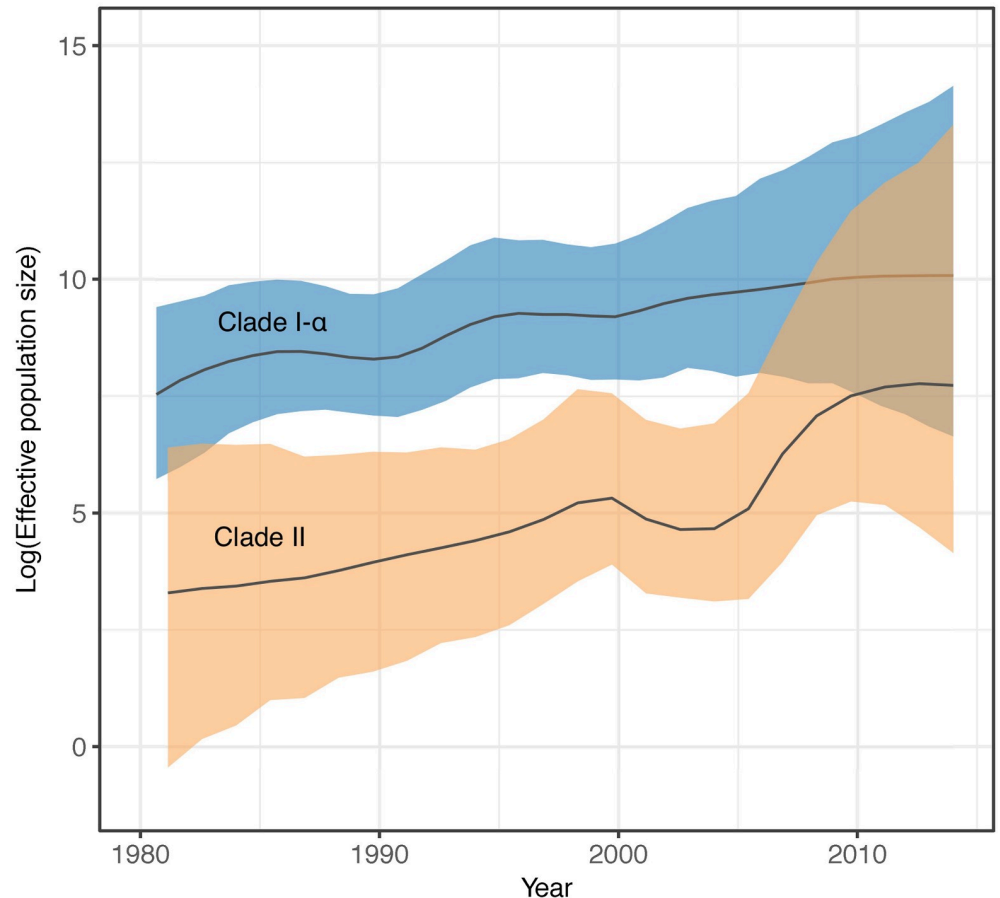


Fig 3. Effective population size (N_e) comparison of serotype 3 CC180 Clades I- α and II. N_e values were estimated using BEAST 1.8.4, enforcing a GMRF SkyGrid demographic model and relaxed molecular clock. The N_e of Clades I- α and II have been increasing with Clade II in particular, increasing exponentially base on phylodynamic model comparison.

<https://doi.org/10.1371/journal.ppat.1007438.g003>

Recombination and genomic variation

We sought to identify whether the clades we identified above showed differences in the historical contribution of recombination, which has been previously considered important in adaptation to clinical interventions [53]. We found that Clades I- α , I- β , and II experienced substantial recombination, impacting 1,179 CDS and diversifying gene content among clades (S5 and S6 Figs). The ratio of polymorphisms introduced through recombination compared to those introduced by mutation (r/m) for the entire sample was estimated at 1.76 with a mean recombination tract length of 12.3 kb. When independently assessed, we found Clade I- α possessed the lowest r/m among the three tested clades as well as the smallest tract length for recombination events (Table 2). Forty-two unique recombination events affecting 582 CDS occurred on the branch segregating Clade II and I- β resulting in an r/m of 4.0 (S5 Fig). Comparison of r/m values between internal and terminal branches of the phylogeny suggests that the great majority of recombination events in CC180 are located on internal branches, although more recent events have occurred especially in Clade I- β (Table 2). We subsequently examined the *comCDE* operon, encoding the competence stimulating peptide and two-component regulatory system, and the associated regulatory genes *comAB* to identify any variations among these loci between the clades. Overall, the *comCDE* operon was markedly

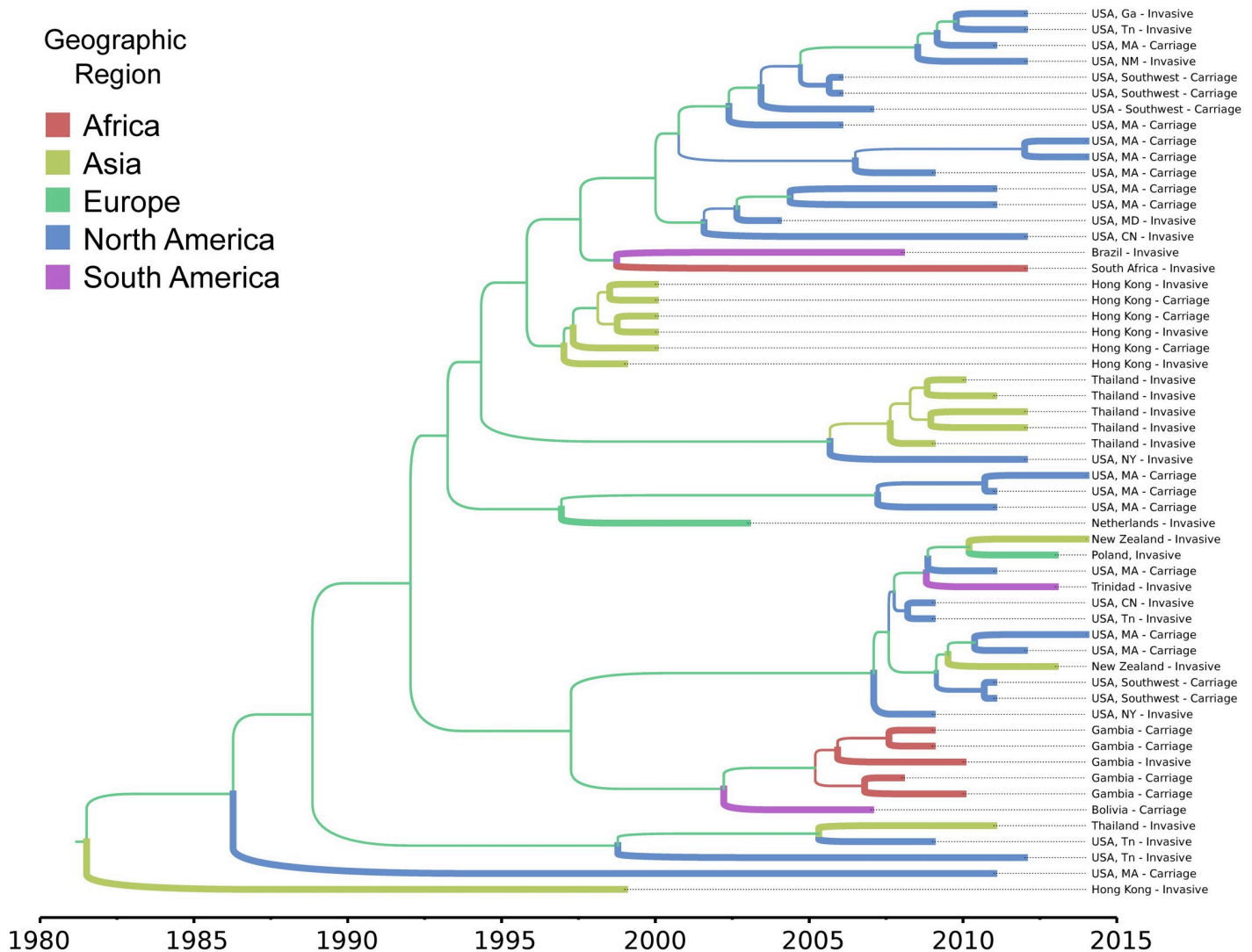


Fig 4. Maximum clade credibility, time-scaled phylogeny of CC180 serotype 3 (PMEN31) Clade II isolates estimated using the structured coalescent model implemented in BEAST2. Branches are colored by geographic region and thickness is scaled values of posterior probabilities for geographic migration. Posterior probabilities for internal branches were all <0.30, precluding assessment of ancestral geographic dispersion.

<https://doi.org/10.1371/journal.ppat.1007438.g004>

uniform (mean pairwise SNP distance = 1 SE 0.5); however, we found a non-synonymous mutation in *comD* (AA104: Pro->Ser) among Clade I-β and II strains, which segregated them from Clade I-α. Further, while competence factor transporting protein *comA* was identical among the three clades, competence factor transport accessory protein *comB* was more diverged between Clades I-β/II and Clade I-α (mean pairwise SNP distance = 1.7 SE 0.7), as a result of a recombination event that affected all strains belonging to Clade II.

Recombination impacted a substantial proportion of the genome, focused in known recombination “hotspots” (Fig 7). As a result, we observed significant variation in mobile genetic elements (MGE) and gene content, including polymorphic protein antigens, evident through comparison of *de novo* assemblies and patterns of recombination. Pangenome comparison identified 1,437 core genes, as well as variations in gene content among clades (S7 Fig). Two notable differences in MGE that resulted in gene content variation among CC180 clades were the presence of a Tn916-like conjugative transposon in Clade II strains (discussed below) and

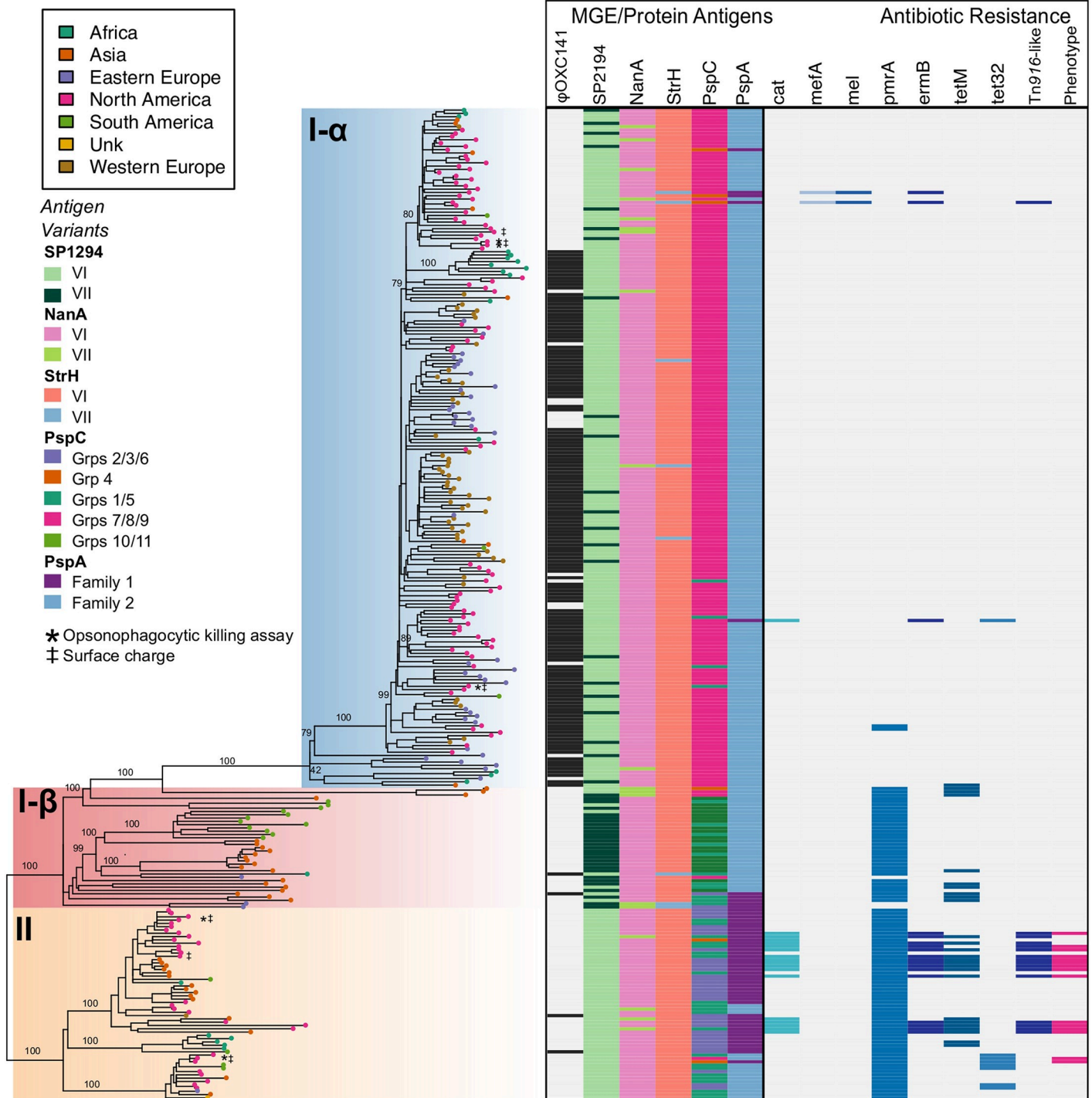


Fig 5. Phylogeny, polymorphic protein antigen variants, and antibiotic resistance. Midpoint rooted maximum-likelihood phylogeny corresponding to Fig 2A with geographic region of isolation represented as colored tip shapes. Clades I- α , I- β , and II are shaded consistent with Fig 2A. Bootstrap values from 100 replicates are provided for major clades and sub-clades (see supplemental phylogeny for all values). Strains used in opsonophagocytic killing assay (asterisk) and surface charge experiments (double dagger) are indicated on the phylogeny. Corresponding protein antigen variants for SP1294, NanA, StrH, PspC, and PspA are illustrated on the left half of the heatmap. Eight other protein antigens are excluded due to a lack of variation in the sample. The presence and absence of AMR-associated genes are illustrated on the right half of the heatmap. The last column of the heatmap indicates genotypic antibiotic resistance that was confirmed by phenotypic testing (broth dilution or disk diffusion).

<https://doi.org/10.1371/journal.ppat.1007438.g005>

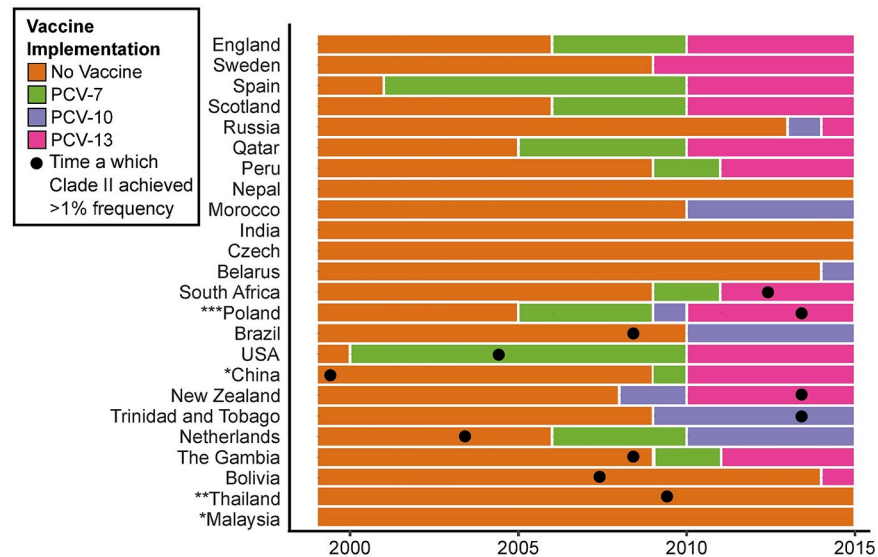


Fig 6. PCV implementation by country and year based on IVAC data from 1999–2014. For each country, the year is shaded based on PCV-7, PCV-10, and PCV-13 vaccine usage. Boxes marked with a dot designate the year in which a Clade II isolate was first observed. *Malaysia and China have yet to introduce PCV as part of their national immunization program. Pneumococcal vaccine use in China varies regionally. Data from China reflects PCV use in Hong Kong where serotype 3 isolates were sampled. **PCV has been available in Thailand as an optional vaccine through the National Vaccine Program since 2006. However, uptake is <5% in children under 5. ***In Poland, PCV was available to parents through private pay. Population based vaccination, without catch-up campaign, was introduced in 2017 using PCV10.

<https://doi.org/10.1371/journal.ppat.1007438.g006>

the absence of the 33.3 kb prophage ϕ OXC141 in Clades I- β and II [54,55]. The ϕ OXC141 prophage, which was putatively acquired by an ancestor of Clade I- α , has been lost multiple times by members of Clade I- α (Fig 5). Notably, it is absent from some Clade I- α isolates from North America, which includes a number of strains from Massachusetts that form a distinct subclade. In all, ϕ OXC141 is present in 71% of Clade I- α isolates and 50% of CC180 serotype 3 strains overall.

Among 13 tested polymorphic protein antigens, five were found to be variable, including membrane associated protein SP2194, cell-wall anchored proteins neuraminidase A (*NanA*) and β -N-acetylhexosaminidase (*StrH*), and surface exposed proteins *pspC* and *pspA* (Fig 5). SP2194 encodes the ATP-binding subunit of Clp protease and is involved in the expression of the *comCDE* operon, which plays an important role in competence, survival, and virulence of pneumococci [56]. Three internal and three terminal branches contain recombination events that have impacted SP2194. Major events occurred on each internal branch leading to Clades

Table 2. Recombination among CC180 lineages. The number of polymorphisms introduced via recombination to mutation (r/m) and the number of recombination events to polymorphisms introduced by mutation (ρ/θ) are reported for internal and terminal branches of dominant clades. Recombination occurring on internal branches is considered ancestral while terminal branches represent recent events. The overall recombination rates are also reported.

Clade/s	Mean tract length (bp)	r/m			ρ/θ		
		Internal	Terminal	Overall	Internal	Terminal	Overall
Clade I- α	2,590.8	0.11	0.07	0.09	0.00	0.01	0.01
Clade I- β	11,457.6	4.15	2.33	3.35	0.03	0.02	0.03
Clades I- β +II	10,629.1	6.18	1.43	4.00	0.05	0.02	0.04
Clade II	9,446.5	5.34	0.32	2.67	0.05	0.01	0.03
CC180	10,219.9	3.26	0.54	1.76	0.03	0.01	0.02

<https://doi.org/10.1371/journal.ppat.1007438.t002>

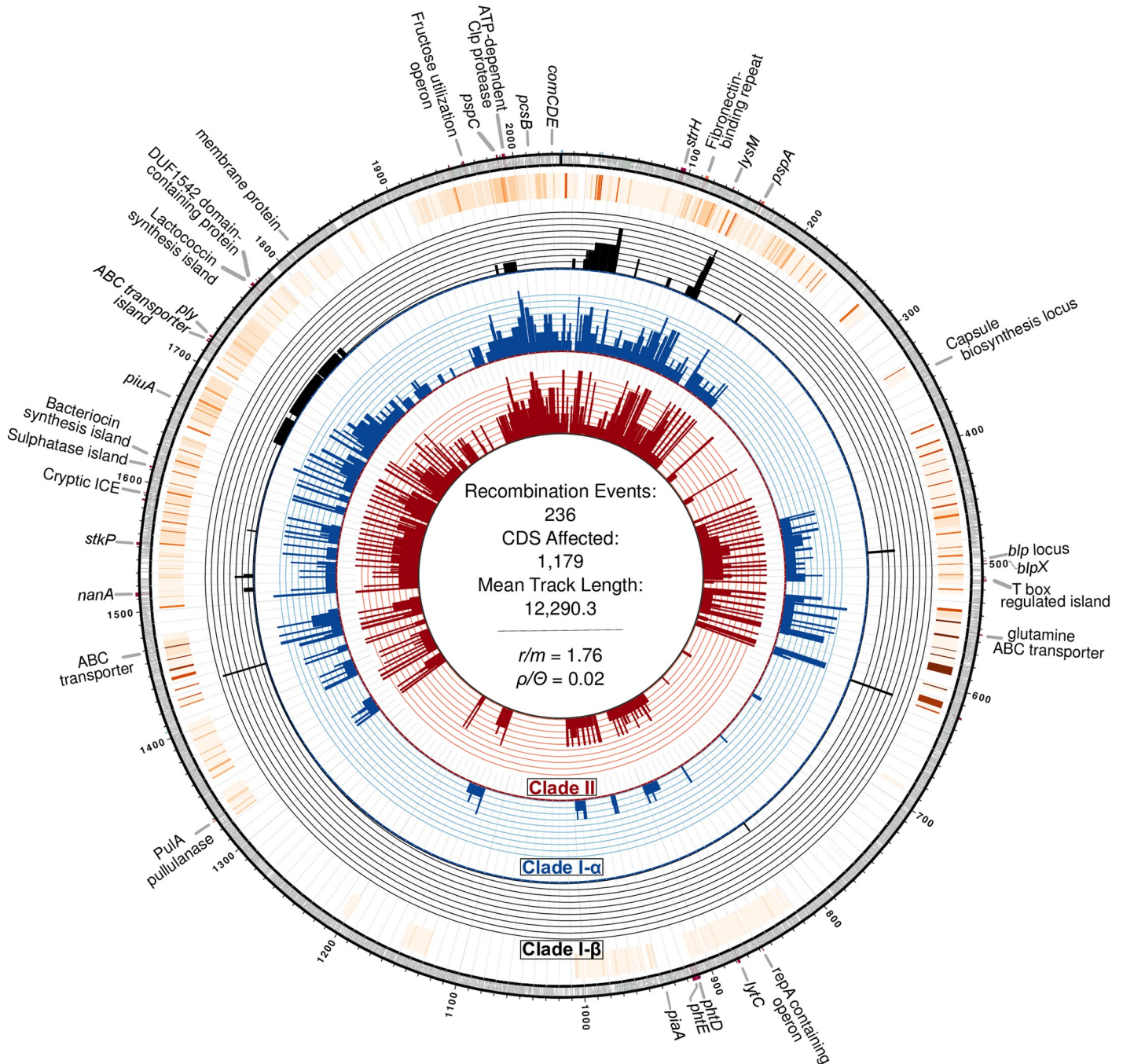


Fig 7. Circos plot of recent and ancestral recombination events inferred among CC180 serotype 3 isolates. Moving from the inner ring outward, rings show a histogram of unique recombination events occurring among isolates belonging to Clades I-β, I-α, and II, respectively, followed by a heatmap of cumulative recombination events. Finally, the outermost ring displays annotations of notable genes and genomic regions corresponding to the location on the OXC141 reference genome.

<https://doi.org/10.1371/journal.ppat.1007438.g007>

I-α, I-β, and II, and as a result, SP2194 is significantly diverged among the three clades (mean between clade SNP difference = 31.1 SE 3.8; p-distance 0.013 SE 0.002). Protein antigen genes *pspA* and *pspC* are known recombination “hotspots” [57], and here we find that recombination generated variation that subsequently become fixed in dominant clades and sub-clades. The Family 2 *pspA* variant was largely associated with Clades I-α and I-β, while the Family I variant

was associated with Clade II [58]. Using the *pspC* groupings—based on structural and nucleotide variation proposed by Iannelli et al.—we found that Clades I and II had non-overlapping sets of *pspC* variation: indeed, Clade I possessed a *pspC* variant consistent with Groups 7/8/9, whereas Clade II variants included Groups 1/5 and 2/3/6 [59]. Hence, we have good evidence that the clades present divergent protein antigenic profiles to the human immune system.

Antimicrobial-resistance (AMR) associated genes

We found variation among clades in genes associated with AMR. Clade II isolates from Hong Kong ($n = 6$) and USA ($n = 9$) harbored a 37.6 kb Tn916-like conjugative transposon possessing *tetM* and *ermB* (S8 Fig). These strains also possess chloramphenicol acetyltransferase (*cat*). For 13 of 15 strains with Tn916, antibiotic susceptibility testing was available in GPS metadata or from previous publications [21]. In general, Tn916 positive isolates exhibited high-level resistance to tetracycline, clindamycin, and erythromycin. However, three USA strains were susceptible to macrolides as the result of a previously described *ermBS* marker, a missense substitution within *ermB* that is associated with erythromycin and clindamycin-susceptibility [60]. Additional phylogenetic analysis of 26 core transposon clusters of orthologous genes (COGs) as well as genealogies of *Int-Tn*, *ermB*, and *tetM* illustrated that there were in fact two acquisitions of genetically similar Tn916 transposons among Clade II isolates (S8 Fig). In both of these instances, the transposon was first observed in a Clade II isolate collected from Hong Kong, which lie basal to a clade of USA isolates possessing the respective variant.

Other notable differences between clades include the presence of *tet32* in eight Clade II isolates, which is distinct from *tetM* carried on Tn916 transposons. In addition, the multi-drug efflux-pump *pmrA* was present in almost all Clade I- β and II isolates. We were able to confidently type PBP transpeptidase domains of 270 isolates to predict penicillin resistance. All were predicted to be penicillin-susceptible; however, four Clade II isolates possessed a PBP profile with first-step β -lactam resistance (transpeptidase domain profile 2-0-111, predicted MIC of 0.06) and were later confirmed by broth dilution to have penicillin MICs ranging from 0.05 ($n = 1$) to 0.06 ($n = 3$). These four isolates also possessed Tn916. PBP profiles further differentiated CC180 clades, with the majority of Clade I- α isolates possessing profile 2-3-2, and Clades I- β and II strains profile 2-0-2 (S2 Table).

Capsular polysaccharide variation in phenotype

Interrogation of inferred recombination events showed that the CPS loci were unaffected by homologous recombination (Fig 7). Furthermore, while the ML phylogeny estimated from the alignment of CPS loci was clearly segregated between the dominant clades, the average pairwise nucleotide difference was only 1.6 (SE 0.4) SNPs (S9 Fig). However the absence of major genotypic variation does not rule out relevant phenotypic variation, so we assessed phenotype of representative Clade I- α ($n = 5$) and Clade II ($n = 3$) serotype 3 strains for three properties previously associated with success: zeta potential, capsule shedding and resistance to neutrophil-mediated opsonophagocytosis. The zeta potential measures surface charge, which varies among capsular type and has previously been associated with differences in prevalence and duration of carriage, with more negatively charged capsules being carried for longer [25], but while the surface charge of clade II isolates was slightly more negative among tested strains this difference was not significant (S10A Fig). In addition, comparison of isolates from Clades I- α and II found no significant difference in capsule shedding (S10B Fig). We did observe differences in neutrophil-mediated opsonophagocytic killing. Clade II isolates were killed in the presence of antisera against both PCV13 and type 3 polysaccharide. In contrast, Clade I- α isolates were not susceptible to killing in the presence of type 3 polysaccharide antisera, and only

one out of the three tested Clade I- α isolate was killed in the presence of high titer PCV13 antisera.

Discussion

Serotype 3 pneumococci belonging to CC180, the most dominant serotype 3 clone globally, remain epidemiologically important. Over the last decade, Clade I- α and Clade II have increased in effective population size, with Clade II becoming particularly prevalent in Asia and North America and emerging in other regions where it has historically not been observed. Phylogenetic analysis demonstrates that Clade II is significantly diverged in both core genome nucleotide diversity and genome content, with multiple recombination events having occurred at the base of the lineage. Temporospatial surveillance data, coalescent analysis, and molecular epidemiology evidence suggests that the diaspora of Clade II may have originated in Southeast Asia in the 1990s and subsequently spread worldwide over two decades, concomitantly growing in prevalence. In North America, Clade II now makes up more than half of the post-PCV13 serotype 3 CC180 population. While neither use of PCV at a country level, nor timing on PCV introduction appears to be clearly associated with the rise of Clade II, we find that Clade II has divergent surface protein antigens and increased prevalence of antibiotic resistance compared to the other CC180 clades. These two observations are likely contributors to the recent increase in prevalence of Clade II.

While phylogeographic analysis was unable to infer confidently the ancestral migration events that disseminated Clade II, the presence of Tn916 among strains from Hong Kong and the USA provides molecular evidence of migration from Asia to North America. In at least two instances, distinct Tn916 transposons are found among isolates from Hong Kong, which are basal to strains from North America. This suggests the acquisition of this transposon predates the emergence of Clade II in North America; thus, strains circulating in Asia were the ancestral populations of a large proportion (~50%) of North American Clade II isolates. Further, multiple introductions of serotype 3 Clade II isolates to the USA from Asia may have occurred. Tn916 confers macrolide resistance, with the rare exception of strains possessing a mutation in *ermB* [61]. Macrolide resistance has previously been described among serotype 3 ST180 isolates from Italy and Japan in 2001 and 2003, respectively, where macrolide-resistant serotype 3 strains are prevalent [19,62,63]. In addition, macrolide and tetracycline resistance ST180 strains have been reported from Taiwan (1997), Spain (2002), and more recently, Canada (2011–14), and Germany (2012) (<https://pubmlst.org/>). Unfortunately, MLST is not discriminating enough to distinguish between the CC180 lineages we discuss here, and as genomic data are not available for these isolates, we cannot determine how they relate to our sample. However, as macrolide resistance among CC180 serotype 3 isolates is relatively rare and largely isolated to Clade II, the early identification of macrolide resistance strains from Hong Kong, Taiwan, and Japan may represent the ancestral population of Clade II putatively possessing Tn916, further supporting a Southeast Asian origin and implicating antibiotic resistance as a possible contributor to the recent success of Clade II.

As PCV use has previously been associated with major changes in pneumococcal population structure including replacement of vaccine serotypes by non-vaccine serotypes (i.e., post-vaccine serotype replacement) [64,65], we considered that serotype 3 may have increased in those settings where PCV was introduced. Moreover, variation in effectiveness of the serotype 3 component of PCV13 among CC180 clades may have disproportionately precipitated an increase in Clade II. Here, we find little evidence that PCV use, albeit at a country level, has shaped the serotype 3 CC180 population. However, there are limitations in these comparisons, as they ignore important factors such as phylogenetic population structure (i.e., whether Clade

II was already present), age of cases, and details of vaccine roll-out (e.g., timing, targeted age groups, and vaccine coverage). Unfortunately, these data are incomplete for a number of cases/countries, precluding an in-depth analysis of the putative association of PCV with emergence of Clade II. Nonetheless, since serotype 3 is a non-vaccine serotype in regions where PCV-7, PCV-10, and, arguably, PCV-13 has been introduced, removal of vaccine serotypes may have precipitated the increase in serotype 3. Further, as we posit below, phenotypic and genomic variation between Clades I- α and II may have led to a competitive advantage to Clade II, which did not exist prior to vaccine introduction.

Historically, antibiotic resistance among serotype 3 CC180 strains has been low. This is because the previous population was dominated by isolates from Clade I, which is largely devoid of AMR-associated genes and mutations. This may result from the high invasiveness of serotype 3 and its relatively short carriage duration, which in turn reduces bystander antibiotic exposure through treatment of other infections [66]. Here we find chloramphenicol, macrolide, tetracycline, and first step penicillin resistance more frequent among Clade II isolates when compared to Clade I isolates. It is not clear whether these are directly linked to the recent increase in prevalence and emergence in previously unobserved regions. However, increased macrolide usage in North America has previously been associated with increases in macrolide resistant *Staphylococcus aureus* and nonsusceptible *S. pneumoniae* [67,68].

In addition to variations in AMR-associated genes, we also find that MGE and polymorphic protein antigens varied among Clades I- α , I- β , and II. The previously described ϕ OXC141 was generally absent among Clades I- β and II and has been lost by a North American sub-clade of Clade I- α [54,55]. Temperate bacteriophages may be associated with fitness defects among pneumococci or changes in virulence and competence [69,70]. Whether the presence or absence of ϕ OXC141 relates to the relative success of serotype 3 CC180 clades is unclear. However, the absence of ϕ OXC141 from Clade II as well as the loss of the phage by a subclade of North American Clade I- α strains remains notable and warrants further investigation. Contributing to genomic variations among clades, *pspA*, *pspC*, and SP2194 (ATP-dependent Clp protease) were found to vary. Natural immunity to protein antigens is generated through nasopharyngeal colonization, and this immunity may protect against colonization and invasive disease [27,71]. Protein antigen variants were generally conserved among Clade I- α isolates and varied in Clades I- β and II. As protein antigen variants are serologically distinct, variations in antigen profiles among CC180 clades could impact relative transmissibility and virulence, resulting in differences in carriage duration, transmission, and invasive capacity [27,59,72,73]. For example, in previous comparisons of Family 1 and 2 *pspA* variants in isogenic strains of serotype 3 (WU2), Family 2 mutants were slightly less virulent and bound less human lactoferrin, impacting nasopharyngeal colonization [74]. Here, we find that Clade II isolates largely possess the Family 1 *pspA* variant. Further, variants of *pspC* differentially bind factor H and generate distinct antibody responses, promoting immune escape [75]. Most isolates in Clade I- α possess a *pspC* variant corresponding to Groups 7/8/9, which have been grouped here because they share the same structural organization [59]. In contrast, Clades I- β and II possess multiple polymorphic *pspC* variants interspersed throughout the topology of the clades, suggesting that multiple recent recombination events have generated variation in this known recombination hotspot as well as a diversified antigenic profile. Taken together, differences in antigenic profiles add to the immunological variability among CC180 clades. Under a model of immune selection, where novel and beneficial combinations of protein antigens are positively selected, this could have led to an increase in the frequency of Clade II isolates. While the epidemiological implications of these findings require further investigation, together with antibiotic resistance, antigenic variation presents another possible contributor to the success of Clade II.

The population of serotype 3 has been previously thought to be largely clonal, clustering into a group of closely related PFGE patterns corresponding to the Netherlands³-31 clone [76]. Further, evolution of CC180 was thought to be driven primarily through nucleotide substitution and estimates of recombination rates based largely on Clade I- α suggested a comparatively lower rate to other PMEN clones [18], mainly dominated by micro-recombination events [77]. However, the expanded global sample of PMEN31 serotype 3 genomes demonstrates in detail how the prevalent lineages have diverged following significant ancestral recombination events. While comparatively rare, recombination continues to shape the CC180 population, as terminal branches in all three clades show evidence of recent recombination events. This inferred clonality likely resulted from the limited sample of type 3 isolates from restricted geographies. Further, while serotype 3 is generally understood to show lower competence than other serotypes based on *in vitro* studies [18,78], we found that the emergent Clades II and I- β have more evidence of historical recombination events (ρ/Θ) and a greater proportion of macro-recombination events than the previously dominant Clade I- α . The Clade I- α *comCDE* operon generates an antisense transcript, an observation that could be relevant to the low historical recombination rate of Clade I- α [18]. Amino acid changes in the *comCDE*, specifically *comD*, also segregate Clade I- α from I- β , and II. We also found variations in the Clp protease gene, which regulates the *comCDE* operon, and is therefore associated with competence rates among pneumococci and other bacteria [56,79]. Increased recombination rates in Clade II have the potential to generate a more diverse antigenic profile, which would be beneficial under diversifying selection from the immune system. More recombination also facilitates the acquisition of multi-drug resistance [80,81]; indeed, higher rates of recombination have previously been associated with antimicrobial resistance [82]. Negative frequency-dependent selection (NFDS) has been proposed as a mechanism for maintaining intermediate frequency loci among pneumococci [83], and together with host immunity, may also explain serotype coexistence [84] and vaccine-induced metabolic shifts [85]. One possible explanation for the relative success of Clade II, is that it possessed a favorable repertoire of accessory loci, protein antigen variants, or metabolic loci, which were selected for in the post-vaccine landscape. The exact loci which are acted upon by NFDS are yet unknown; however, it could involve immune selection of protein antigens, antibiotic resistance conferring gene, and bacteriophage predation defence [12,83]. Here, we find variation in all three between Clade I and II.

Because serotype 3 pneumococcal strains possess a mucoid capsule and release significantly more capsular polysaccharide during *in vitro* growth and infection in mice, it is hypothesized that this soluble polysaccharide absorbs anti-capsular antibody, effectively impeding antibody-mediated killing *in vitro* and *in vivo* [24,26,52]. We found that Clades I- α and II both shed CPS during growth; however, the amount of CPS released did not significantly differ between the two clades. Another phenotype that has been found to correlate with increased survival due to resistance from phagocytosis by human neutrophils, and contribute to success in nasopharyngeal carriage is low surface charge, measured as zeta potential [25]. We therefore compared this between Clades I- α and II and found no significant difference in charge between clades, suggesting their capsular properties are similar. Where we did find significant phenotypic differences between the clades, it was not easy to relate these differences to the apparent increase in Clade II following PCV13 introduction; we initially hypothesized that antisera against PCV13 and type 3 polysaccharide might be more effective against isolates from Clade I- α , explaining its apparent decline. However, for both Clades I- α and II, neutrophil-mediated opsonophagocytic killing occurred only at high antibody levels. Clade I- α isolates appeared to be resistant to killing by PCV13 antisera, while Clade II isolates were more readily killed. The resistance of Clade I isolates to opsonophagocytic killing is consistent with the low efficacy of PCV13 on serotype 3 invasive disease incidence [10,86], but difficult to link to the recent

emergence of Clade II. However, these results should not be considered a direct representation of vaccine effectiveness, because opsonophagocytic killing may be a poor proxy for the ability to colonize and transmit. Moreover, opsonophagocytic killing only occurred at high titers. Our findings are consistent with those of other groups, suggesting that higher antibody concentrations are needed for killing *in vitro* [52]. They also demonstrate that closely related bacteria with the same capsule do not necessarily behave similarly in the face of vaccination.

Together, our findings further support the involvement of anti-protein antibodies in mediating anti-serotype 3 immunity. Anti-protein antibodies may act in synergy *in vivo* with anticapsular antibodies to mediate pneumococcal killing. This is an intriguing hypothesis in light of our other findings about the differences in protein variant composition among serotype 3 strains. These differences in protein composition underscore the need for better evaluation of the role that anti-protein antibodies have in mediating anti-pneumococcal protection. Overall, the role of these in recent dynamics of serotype 3 epidemiology requires further study.

Serotype 3 remains an epidemiologically important serotype, continuing to cause invasive disease and increasing in carriage prevalence in North America. Here, we explore multiple epidemiological, genotypic, and phenotypic factors associated with the persistence of serotype 3 in carriage and disease. We found that the recent success of CC180 is related to an increasing clade, which is distinct from the previous population's antigenic composition, antibiotic susceptibility, and competence. Based on OPKA, we also found evidence to support the low-observed efficacy of PCV13 against serotype 3, which was previously dominated by CC180 Clade I- α in Europe and North America. Optimistically, the efficacy of PCV13 against the emergent Clade II may be higher; however, more *in vitro* and *in vivo* studies are required to confirm this finding.

Our study is not without limitations. While we base our analysis on the largest collection of extant CC180 genomes, geographic and temporal sampling variations exist. To mitigate coalescent analysis bias, we down-sampled according to published approaches. Additionally, while the clonality and low diversity of Clade I- α may, at first approximation, suggest a more recent origin, HPDs for the TMRCA of Clade I- α and II estimated through coalescent analysis are non-overlapping and correspond to previously published estimates [18]. Further, despite modest sampling from multiple geographies prior to 1999 ($n = 34$), Clade II isolates were not observed. Our results instead indicate that the higher recombination rate and recent population expansion of Clade II has led to its greater diversity. Lastly, our genomic results may be affected by the accuracy of bioinformatic programs. Wherever possible, results were verified using multiple approaches or manual inspection of the data.

Overall, our analysis emphasizes the need for routine, representative sampling of isolates from diverse geographic regions, including historically under-sampled areas. We also highlight the value of genomics in resolving antigenic and epidemiological variations within a serotype, which may have implications for future vaccine development. As PCV usage and antibiotic consumption expands globally, it is imperative to continue genomic and epidemiological surveillance of pneumococci to detect the emergence of new lineages and monitor changes in clinical presentation and severity, as these data have direct implications for prevention and management of pneumococcal disease.

Supporting information

S1 Methods.

(DOCX)

S1 File. File containing accession numbers and associated metadata.

(XLSX)

S1 Fig. Root-to-tip correlation of Clade I- α . The tip dates, assigned using the collection date of the isolate, is regressed upon the distance from the estimated root in the maximum likelihood phylogeny to determine temporal signal.

(TIF)

S2 Fig. Root-to-tip correlation of Clade II. The tip dates, assigned using the collection date of the isolate, is regressed upon the distance from the estimated root in the maximum likelihood phylogeny to determine temporal signal. The corresponding taxa for the points in the upper-right of the regression plot (indicated with a purple oval) are identified on the phylogeny to the left. The root-to-tip correlation was also independently assessed for the sub-clade indicated in purple and found to have a significant correlation.

(TIF)

S3 Fig. Date randomization test of Clade II. The estimated evolutionary rate for the BEAST run with correct date assignment (T:true) appears at the far left of the x-axis, with the realizations for 10 date permutation runs to the right. Evolutionary rates in this figure are not corrected for alignment length.

(TIF)

S4 Fig. Comparison of estimated evolutionary rate (top) and the most recent common ancestor (TMRCA) (bottom) between Clades I- α and II. Evolutionary rate is scaled in SNPs/site/year and TMRCA as the year of the estimated root height. Clade I is significantly “older” than the recently emerged Clade II.

(TIF)

S5 Fig. A.) Distribution of inferred recombination block sizes for the entire sample and B.) the major branch leading to Clade II (shaded and denoted with an asterisk). C.) Phylogeny depicting the population structure of CC180 with Clade II shaded. In total, 236 Recombination events impacted 1,179 coding sequences (CDS), with 42 of those events occurring on the major branch leading to Clade II. The largest event, a 68,164 bp recombination block spanning the region 407,458–475,621 on the OXC141 reference genome, impacting 70 CDS and 13 repeat regions. Pathways including these CDSs involve metabolism ($n = 17$), genetic information processing ($n = 10$), signaling ($n = 7$), and environmental information processing ($n = 5$). Notably, *rpoE* recombination and DNA repair protein (SPNOXC_04590) is included in this region.

(TIF)

S6 Fig. Recombination analysis of clonal complex 180 isolates. The left side of the figure represents a maximum likelihood phylogeny of CC180 isolates inferred from a recombination-censored alignment using Gubbins. The tree has been out-group rooted as described in the methods and clades are ordered Clade I- α , I- β , and II, from top to bottom. On the top of the figure is a linear depiction of the 2,036,867 bp *S. pneumoniae* OXC141 reference genome on which recombination events are mapped. The bottom ribbon is a graph of recombination events distributed linearly across the genome. The center panel indicated recombination events among taxa. Red blocks display events occurring on ancestral branches and are therefore shared by a group of taxa. Darker shades of red indicate multiple events have occurred in a genomic region. Blue blocks signify recombination events that occurred on terminal branches and are unique to only one isolate. Output of Gubbins was visualized in Phandango, which was used to construct this figure.

(TIF)

S7 Fig. Venn diagram of core orthologous gene (COG) content of Clades I- α , I- β , and II. COGs possessed by at least one individual within a clade were counted toward the “gene-pool” for each comparison.

(TIF)

S8 Fig. Annotation of 37.6 kb Tn916-like transposon found in Clade II strains (top) and phylogenies of 26 transposon core genes, *Int-Tn*, *ermB*, and *tetM* (bottom). The core genome phylogeny and genealogies are labeled with length of alignment, mean pairwise difference (mpwd), and the ratio of non-synonymous to synonymous mutations (dN/dS). The phylogenies show a clear delineation between two distinct transposons (blue and black tip labels). In the *ermB* genealogy, two isolates that are macrolide susceptible are colored in green.

(TIF)

S9 Fig. Maximum likelihood phylogeny inferred from 13.2 kb region comprising the capsular polysaccharide (CPS) coding loci. Tips are labeled and colored by clade. Despite the clear divergence of Clades I- α and II+I- β , very little nucleotide diversity exists.

(TIF)

S10 Fig. Comparison of surface charge (zeta potential measured in mV) (top) and capsule shedding (bottom) between strains belonging to Clades I- α and II. ST3 Serotype 3 strain WU2 is used as a control for both experiments and WU2 Δ CPS is included in the analysis of zeta potentials.

(TIF)

S11 Fig. Opsonophagocytic killing in the presence of antisera against type 3 polysaccharide (A) and PCV13 (B) between isolates from Clades I- α and II. Antisera dilutions are labeled on the x-axis and percent survival on the y-axis. Clade I- α isolates are on the left half of the x-axis and Clade II appear on the right. Opsonophagocytic killing appears as a reduction in percent survival with increasing concentrations of antisera.

(TIF)

S1 Table. Results of Bayesian phylogenetic analysis using Beast 1.8.4 for Clades I- α and II. For each clade, strict (SC) and relaxed (RC) molecular clock models were compared for constant, exponential, and GMRF Skygrid demographic models. Log marginal likelihood estimates (MLE) from path-sampling (PS) and stepping-stone (SS) analysis were used to calculate Bayes Factors for model comparison. Log Bayes Factors (BF) are specified for each molecular clock and demographic model comparison. The date of the most recent common ancestor (TMRCA) and evolutionary rate, scaled in SNPs/site/year are presented for the final model with corresponding highest posterior densities (HPD).

(DOCX)

S2 Table. PBP transpeptidase domain variants (PBP1a-PBP2b-PBP2x) of 270 serotype 3 CC180 isolates. 2-0-111 possesses first step PBP-2X 111 allele that is associated with a slightly elevated penicillin MIC of 0.06. All other profiles are associated with a penicillin MIC ($\mu\text{g/ml}$) ≤ 0.03 .

(DOCX)

Acknowledgments

We acknowledge the Wellcome Trust Sanger Institute sequencing facility. We would also like to thank Richard Malley’s Laboratory for providing the antisera for the OPKA assays.

Author Contributions

Conceptualization: Taj Azarian, Maria Georgieva, Rebecca A. Gladstone, William P. Hanage.

Data curation: Taj Azarian, Maria Georgieva, Claudette M. Thompson, Bernard Beall.

Formal analysis: Taj Azarian, Maria Georgieva, Claudette M. Thompson, Amel Ghouila, Bernard Beall, Benjamin J. Metcalf.

Funding acquisition: William P. Hanage.

Investigation: Patrick K. Mitchell, Maria Georgieva, Claudette M. Thompson, Anne von Gottberg, Mignon du Plessis, Martin Antonio, Brenda A. Kwambana-Adams, Stuart C. Clarke, Dean Everett, Jennifer Cornick, Ewa Sadowy, Waleria Hryniewicz, Anna Skoczynska, Jennifer C. Moisi, Lesley McGee, Benjamin J. Metcalf, Robert F. Breiman, PL Ho, Raymond Reid, Katherine L. O'Brien, Rebecca A. Gladstone, Stephen D. Bentley.

Methodology: Taj Azarian, Maria Georgieva, Claudette M. Thompson, Andrew J. Pollard, Anne von Gottberg, Martin Antonio, Stuart C. Clarke, Jennifer Cornick, Ewa Sadowy, Bernard Beall, PL Ho, William P. Hanage.

Project administration: William P. Hanage.

Resources: Katherine L. O'Brien, Rebecca A. Gladstone, Stephen D. Bentley, William P. Hanage.

Software: Benjamin J. Metcalf, Stephen D. Bentley.

Supervision: Stephen D. Bentley, William P. Hanage.

Validation: Taj Azarian, William P. Hanage.

Visualization: Taj Azarian.

Writing – original draft: Taj Azarian, Maria Georgieva, Anna Skoczynska, William P. Hanage.

Writing – review & editing: Taj Azarian, Andrew J. Pollard, Anne von Gottberg, Mignon du Plessis, Brenda A. Kwambana-Adams, Stuart C. Clarke, Dean Everett, Jennifer Cornick, Ewa Sadowy, Waleria Hryniewicz, Anna Skoczynska, Jennifer C. Moisi, Lesley McGee, Bernard Beall, Benjamin J. Metcalf, Robert F. Breiman, PL Ho, Raymond Reid, Katherine L. O'Brien, Rebecca A. Gladstone, Stephen D. Bentley, William P. Hanage.

References

1. van Tonder AJ, Bray JE, Quirk SJ, Haraldsson G, Jolley KA, Maiden MCJ, et al. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. *Microb Genomics*. 2016; 2: 000090. <https://doi.org/10.1099/mgen.0.000090> PMID: 28133541
2. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J Infect Dis*. 2003; 187: 1424–32. <https://doi.org/10.1086/374624> PMID: 12717624
3. Martens P, Worm SW, Lundgren B, Konradsen HB, Benfield T. Serotype-specific mortality from invasive *Streptococcus pneumoniae* disease revisited. *BMC Infect Dis*. BioMed Central; 2004; 4: 21. <https://doi.org/10.1186/1471-2334-4-21> PMID: 15228629
4. Täuber MG, Burroughs M, Niemöller UM, Kuster H, Borschberg U, Tuomanen E. Differences of pathophysiology in experimental meningitis caused by three strains of *Streptococcus pneumoniae*. *J Infect Dis*. 1991; 163: 806–11. Available: <http://www.ncbi.nlm.nih.gov/pubmed/1901331> PMID: 1901331

5. Briles DE, Crain MJ, Gray BM, Forman C, Yother J. Strong association between capsular type and virulence for mice among human isolates of *Streptococcus pneumoniae*. *Infect Immun. American Society for Microbiology (ASM)*; 1992; 60: 111–116. Available: <http://www.ncbi.nlm.nih.gov/pubmed/1729176>
6. Moore MR, Link-Gelles R, Schaffner W, Lynfield R, Lexau C, Bennett NM, et al. Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance. *Lancet Infect Dis.* 2015; 15: 301–309. [https://doi.org/10.1016/S1473-3099\(14\)71081-3](https://doi.org/10.1016/S1473-3099(14)71081-3) PMID: 25656600
7. Dagan R, Patterson S, Juergens C, Greenberg D, Givon-Lavi N, Porat N, et al. Comparative Immunogenicity and Efficacy of 13-Valent and 7-Valent Pneumococcal Conjugate Vaccines in Reducing Nasopharyngeal Colonization: A Randomized Double-Blind Trial. *Clin Infect Dis.* 2013; 57: 952–962. <https://doi.org/10.1093/cid/cit428> PMID: 23804191
8. Andrews NJ, Waight PA, Burbidge P, Pearce E, Roalfe L, Zancolli M, et al. Serotype-specific effectiveness and correlates of protection for the 13-valent pneumococcal conjugate vaccine: a postlicensure indirect cohort study. *Lancet Infect Dis.* 2014; 14: 839–46. [https://doi.org/10.1016/S1473-3099\(14\)70822-9](https://doi.org/10.1016/S1473-3099(14)70822-9) PMID: 25042756
9. Kaplan SL, Barson WJ, Lin PL, Romero JR, Bradley JS, Tan TQ, et al. Early Trends for Invasive Pneumococcal Infections in Children After the Introduction of the 13-valent Pneumococcal Conjugate Vaccine. *Pediatr Infect Dis J.* 2013; 32: 203–207. <https://doi.org/10.1097/INF.0b013e318275614b> PMID: 23558320
10. Andrews NJ, Waight PA, Burbidge P, Pearce E, Roalfe L, Zancolli M, et al. Serotype-specific effectiveness and correlates of protection for the 13-valent pneumococcal conjugate vaccine: a postlicensure indirect cohort study. www.thelancet.com/Infect. 2014;
11. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, 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
12. 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
13. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol.* 2013; 30: 1224–1228. <https://doi.org/10.1093/molbev/mst028> PMID: 23408797
14. Coffey TJ, Enright MC, Daniels M, Morona JK, Morona R, Hryniewicz W, et al. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. *Mol Microbiol.* 1998; 27: 73–83. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9466257> PMID: 9466257
15. Spratt BG, Hanage WP, Li B, Aanensen DM, Feil EJ. Displaying the relatedness among isolates of bacterial species—the eBURST approach. *FEMS Microbiol Lett.* 2004; 241: 129–34. <https://doi.org/10.1016/j.femsle.2004.11.015> PMID: 15598523
16. Mosser JF, Grant LR, Millar EV, Weatherholtz RC, Jackson DM, Beall B, et al. Nasopharyngeal carriage and transmission of *Streptococcus pneumoniae* in American Indian households after a decade of pneumococcal conjugate vaccine use. *PLoS One. Public Library of Science*; 2014; 9: e79578. <https://doi.org/10.1371/journal.pone.0079578> PMID: 24465365
17. Clarke SC, Scott KJ, McChlery SM. Serotypes and sequence types of pneumococci causing invasive disease in Scotland prior to the introduction of pneumococcal conjugate polysaccharide vaccines. *J Clin Microbiol. American Society for Microbiology*; 2004; 42: 4449–4452. <https://doi.org/10.1128/JCM.42.10.4449-4452.2004>
18. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, et al. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. *PLoS Genet.* 2013; 9: e1003868. <https://doi.org/10.1371/journal.pgen.1003868> PMID: 24130509
19. Imai S, Ito Y, Ishida T, Hirai T, Ito I, Maekawa K, et al. High prevalence of multidrug-resistant Pneumococcal molecular epidemiology network clones among *Streptococcus pneumoniae* isolates from adult patients with community-acquired pneumonia in Japan. *Clin Microbiol Infect.* 2009; 15: 1039–1045. <https://doi.org/10.1111/j.1469-0691.2009.02935.x> PMID: 19694764
20. Mothibeli KM, Du Plessis M, Von Gottberg A, De Gouveia L, Adrian P, Madhi SA, et al. An unusual pneumococcal sequence type is the predominant cause of serotype 3 invasive disease in South Africa. *J Clin Microbiol. American Society for Microbiology*; 2010; 48: 184–191. <https://doi.org/10.1128/JCM.01011-09> PMID: 19889905
21. 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
22. Lapidot R, Shea K, Little B, Yildirim I, Pelton S. No Title. Impact of PCV13 on Serotype 3 Invasive Pneumococcal Disease and Nasopharyngeal Colonization in Massachusetts' children. San Diego, California: IDWeek; 2017.
 23. Mitchell P, Azarian T, Croucher NJ, Callendrello A, Thompson CM, Pelton SI, et al. Population genomics of pneumococcal carriage in Massachusetts children following PCV-13 introduction. bioRxiv. Cold Spring Harbor Laboratory; 2017; 235192. <https://doi.org/10.1101/235192>
 24. Choi EH, Zhang F, Lu YJ, Malley R. Capsular Polysaccharide (CPS) Release by Serotype 3 Pneumococcal Strains Reduces the Protective Effect of Anti-Type 3 CPS Antibodies. Clin Vaccine Immunol. American Society for Microbiology; 2016; 23: 162–167. <https://doi.org/10.1128/CVI.00591-15> PMID: 26677201
 25. Li Y, Weinberger DM, Thompson CM, Trzciński K, Lipsitch M. Surface charge of *Streptococcus pneumoniae* predicts serotype distribution. Infect Immun. American Society for Microbiology (ASM); 2013; 81: 4519–24. <https://doi.org/10.1128/IAI.00724-13> PMID: 24082068
 26. WOOD WB, SMITH MR. The inhibition of surface phagocytosis by the capsular slime layer of pneumococcus type III. J Exp Med. 1949; 90: 85–96. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18152341> PMID: 18152341
 27. 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.
 28. Mitchell P, Azarian T, Croucher NJ, Callendrello A, Thompson CM, Pelton SI, et al. Population genomics of pneumococcal carriage in Massachusetts children following PCV-13 introduction. bioRxiv. 2017; <http://biorxiv.org/content/early/2017/12/16/235192.abstract>
 29. Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M, et al. The impact of serotype-specific vaccination on phylodynamic parameters of *Streptococcus pneumoniae* and the pneumococcal pan-genome. Tang C, editor. PLOS Pathog. Public Library of Science; 2018; 14: e1006966. <https://doi.org/10.1371/journal.ppat.1006966> PMID: 29617440
 30. Minka T. Estimating a Dirichlet distribution. Technical report, MIT; 2000.
 31. 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
 32. 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
 33. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30: 2068–9. <https://doi.org/10.1093/bioinformatics/btu153> PMID: 24642063
 34. 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
 35. 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
 36. 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
 37. 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
 38. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. Cold Spring Harbor Lab; 2009; 19: 1639–1645.
 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. Camilli R, Bonnal RJP, Del Grosso M, Iacono M, Corti G, Rizzi E, et al. Complete genome sequence of a serotype 11A, ST62 *Streptococcus pneumoniae* invasive isolate. BMC Microbiol. BioMed Central; 2011; 11: 25. <https://doi.org/10.1186/1471-2180-11-25> PMID: 21284853
 41. Hall MD, Woolhouse MEJ, Rambaut A. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: A simulation study. Virus Evol. 2016; 2: vew003. <https://doi.org/10.1093/ve/vew003> PMID: 27774296

42. Duchêne S, Duchêne D, Holmes EC, Ho SYW. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol.* Oxford University Press; 2015; 32: 1895–1906. <https://doi.org/10.1093/molbev/msv056> PMID: 25771196
43. Ramsden C, Melo FL, Figueiredo LM, Holmes EC, Zanotto PMA, Consortium V. High rates of molecular evolution in hantaviruses. *Mol Biol Evol.* Oxford University Press; 2008; 25: 1488–1492.
44. 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> PMID: 22367748
45. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 2014; 10: e1003537. <https://doi.org/10.1371/journal.pcbi.1003537> PMID: 24722319
46. Vaughan TG, Kühnert D, Poppinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. *Bioinformatics.* 2014; btu201-. <https://doi.org/10.1093/bioinformatics/btu201> PMID: 24753484
47. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother.* Am Soc Microbiol; 2013; 57: 3348–3357.
48. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother.* Am Soc Microbiol; 2014; 58: 212–220.
49. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, et al. Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β -Lactam Resistance Levels in *Streptococcus pneumoniae*. *MBio.* American Society for Microbiology; 2016; 7: e00756–16. <https://doi.org/10.1128/mBio.00756-16> PMID: 27302760
50. Choi EH, Zhang F, Lu YJ, Malley R. Capsular Polysaccharide (CPS) Release by Serotype 3 Pneumococcal Strains Reduces the Protective Effect of Anti-Type 3 CPS Antibodies. *Clin Vaccine Immunol.* American Society for Microbiology (ASM); 2016; 23: 162–167. <https://doi.org/10.1128/CVI.00591-15> PMID: 26677201
51. Daniels CC, Kim KH, Burton RL, Mirza S, Walker M, King J, et al. Modified opsonization, phagocytosis, and killing assays to measure potentially protective antibodies against pneumococcal surface protein A. *Clin Vaccine Immunol.* 2013; 20: 1549–1558. <https://doi.org/10.1128/CVI.00371-13> PMID: 23925886
52. Zhang F, Lu Y-J, Malley R. Multiple antigen-presenting system (MAPS) to induce comprehensive B- and T-cell immunity. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2013; 110: 13564–9. <https://doi.org/10.1073/pnas.1307228110> PMID: 23898212
53. Croucher NJ, Harris SR, Fraser C, Quail M a, Burton J, van der Linden M, et al. Rapid pneumococcal evolution in response to clinical interventions. *Science.* 2011; 331: 430–4. <https://doi.org/10.1126/science.1198545> PMID: 21273480
54. Romero P, Croucher NJ, Hiller NL, Hu FZ, Ehrlich GD, Bentley SD, et al. Comparative genomic analysis of ten *Streptococcus pneumoniae* temperate bacteriophages. *J Bacteriol.* 2009; 191: 4854–62. <https://doi.org/10.1128/JB.01272-08> PMID: 19502408
55. Chancey ST, Agrawal S, Schroeder MR, Farley MM, Tettelin H, Stephens DS. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. *Front Microbiol.* Frontiers Media SA; 2015; 6: 26. <https://doi.org/10.3389/fmicb.2015.00026> PMID: 25709602
56. Chastanet A, Prudhomme M, Claverys JP, Msadek T. Regulation of *Streptococcus pneumoniae* clp genes and their role in competence development and stress survival. *Journal of Bacteriology.* American Society for Microbiology; 2001. pp. 7295–7307. <https://doi.org/10.1128/JB.183.24.7295-7307.2001>
57. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, 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
58. Hollingshead SK, Becker R, Briles DE. Diversity of PspA: Mosaic Genes and Evidence for Past Recombination in *Streptococcus pneumoniae*. *Infect Immun.* 2000; 68: 5889–5900. <https://doi.org/10.1128/IAI.68.10.5889-5900.2000> PMID: 10992499
59. Iannelli F, Oggioni MR, Pozzi G. Allelic variation in the highly polymorphic locus pspC of *Streptococcus pneumoniae*. *Gene.* 2002; 284: 63–71. [https://doi.org/10.1016/S0378-1119\(01\)00896-4](https://doi.org/10.1016/S0378-1119(01)00896-4) PMID: 11891047
60. Metcalf BJ, Chochua S, Gertz RE, Li Z, Walker H, Tran T, et al. Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. *Clin Microbiol Infect.* 2016; 22: 1002.e1–1002.e8. <https://doi.org/10.1016/j.cmi.2016.08.001> PMID: 27542334

61. Rice LB, Carias LL, Marshall SH, Hutton-Thomas R, Rudin S. Characterization of Tn5386, a Tn916-related mobile element. *Plasmid*. 2007; 58: 61–7. <https://doi.org/10.1016/j.plasmid.2007.01.002> PMID: [17408741](https://pubmed.ncbi.nlm.nih.gov/17408741/)
62. Isozumi R, Ito Y, Ishida T, Hirai T, Ito I, Maniwa K, et al. Molecular characteristics of serotype 3 *Streptococcus pneumoniae* isolates among community-acquired pneumonia patients in Japan. *J Infect Chemother*. 2008; 14: 258–61. <https://doi.org/10.1007/s10156-008-0600-9> PMID: [18574666](https://pubmed.ncbi.nlm.nih.gov/18574666/)
63. Gherardi G, Fallico L, Del Grosso M, Bonanni F, D'Ambrosio F, Manganelli R, et al. Antibiotic-resistant invasive pneumococcal clones in Italy. *J Clin Microbiol*. American Society for Microbiology; 2007; 45: 306–12. <https://doi.org/10.1128/JCM.01229-06> PMID: [17122014](https://pubmed.ncbi.nlm.nih.gov/17122014/)
64. Hanage WP, Finkelstein JA, Huang SS, Pelton SI, Stevenson AE, Kleinman K, et al. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics*. 2010; 2: 80–84. <https://doi.org/10.1016/j.epidem.2010.03.005> PMID: [21031138](https://pubmed.ncbi.nlm.nih.gov/21031138/)
65. 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](https://doi.org/10.1016/S0140-6736(10)62225-8) PMID: [21492929](https://pubmed.ncbi.nlm.nih.gov/21492929/)
66. Lehtinen S, Blanquart F, Croucher NJ, Turner P, Lipsitch M, Fraser C. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc Natl Acad Sci*. National Academy of Sciences; 2017; 114: 1075–1080. <https://doi.org/10.1073/pnas.1617849114> PMID: [28096340](https://pubmed.ncbi.nlm.nih.gov/28096340/)
67. Uhlemann A-C, Dordel J, Knox JR, Raven KE, Parkhill J, Holden MTG, et al. Molecular tracing of the emergence, diversification, and transmission of *S. aureus* sequence type 8 in a New York community. *Proc Natl Acad Sci*. 2014; 1401006111-. <https://doi.org/10.1073/pnas.1401006111> PMID: [24753569](https://pubmed.ncbi.nlm.nih.gov/24753569/)
68. Hicks L a., Chien Y-W, Taylor TH, Haber M, Klugman KP. Outpatient Antibiotic Prescribing and Nonsusceptible *Streptococcus pneumoniae* in the United States, 1996–2003. *Clin Infect Dis*. 2011; 53: 631–639. <https://doi.org/10.1093/cid/cir443> PMID: [21890767](https://pubmed.ncbi.nlm.nih.gov/21890767/)
69. DeBardeleben HK, Lysenko ES, Dalia AB, Weisera JN. Tolerance of a phage element by *Streptococcus pneumoniae* leads to a fitness defect during colonization. *J Bacteriol*. American Society for Microbiology (ASM); 2014; 196: 2670–2680. <https://doi.org/10.1128/JB.01556-14> PMID: [24816604](https://pubmed.ncbi.nlm.nih.gov/24816604/)
70. Bobay L-M, Rocha EPC, Touchon M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Mol Biol Evol*. Oxford University Press; 2013; 30: 737–751. <https://doi.org/10.1093/molbev/mss279> PMID: [23243039](https://pubmed.ncbi.nlm.nih.gov/23243039/)
71. 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](https://pubmed.ncbi.nlm.nih.gov/28135322/)
72. Turner P, Turner C, Green N, Ashton L, Lwe E, Jankhot A, et al. Serum antibody responses to pneumococcal colonization in the first 2 years of life: results from an SE Asian longitudinal cohort study. *Clin Microbiol Infect*. 2013; 19: E551–8.
73. Ren B, Szalai AJ, Thomas O, Hollingshead SK, Briles DE. Both family 1 and family 2 PspA proteins can inhibit complement deposition and confer virulence to a capsular serotype 3 strain of *Streptococcus pneumoniae*. *Infect Immun*. American Society for Microbiology; 2003; 71: 75–85. <https://doi.org/10.1128/IAI.71.1.75-85.2003>
74. Ren B, Szalai AJ, Thomas O, Hollingshead SK, Briles DE. Both family 1 and family 2 PspA proteins can inhibit complement deposition and confer virulence to a capsular serotype 3 strain of *Streptococcus pneumoniae*. *Infect Immun*. American Society for Microbiology; 2003; 71: 75–85. <https://doi.org/10.1128/IAI.71.1.75-85.2003>
75. Georgieva M, Kagedan L, Lu Y-J, Thompson CM, Lipsitch M. Antigenic Variation in *Streptococcus pneumoniae* PspC Promotes Immune Escape in the Presence of Variant-Specific Immunity. *MBio*. Am Soc Microbiol; 2018; 9: e00264–18.
76. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, et al. Nomenclature of Major Antimicrobial-Resistant Clones of *Streptococcus pneumoniae* Defined by the Pneumococcal Molecular Epidemiology Network. *J Clin Microbiol*. 2001; 39: 2565–2571. <https://doi.org/10.1128/JCM.39.7.2565-2571.2001> PMID: [11427569](https://pubmed.ncbi.nlm.nih.gov/11427569/)
77. Mostowy R, Croucher NJ, Hanage WP, Harris SR, Bentley S, Fraser C. Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution. Cooper VS, editor. *PLoS Genet*. Public Library of Science; 2014; 10: e1004300. <https://doi.org/10.1371/journal.pgen.1004300> PMID: [24786281](https://pubmed.ncbi.nlm.nih.gov/24786281/)
78. Hsieh Y-C, Wang J-T, Lee W-S, Hsueh P-R, Shao P-L, Chang L-Y, et al. Serotype competence and penicillin resistance in *Streptococcus pneumoniae*. *Emerg Infect Dis*. 2006; 12: 1709–14. <https://doi.org/10.3201/eid1211.060414> PMID: [17283621](https://pubmed.ncbi.nlm.nih.gov/17283621/)

79. Hui FM, Zhou L, Morrison DA. Competence for genetic transformation in *Streptococcus pneumoniae*: organization of a regulatory locus with homology to two lactococcal A secretion genes. *Gene*. 1995; 153: 25–31. [https://doi.org/10.1016/0378-1119\(94\)00841-F](https://doi.org/10.1016/0378-1119(94)00841-F) PMID: 7883181
80. 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
81. Perron GG, Lee AEG, Wang Y, Huang WE, Barraclough TG. Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations. *Proc Biol Sci*. 2011; 1477–1484. <https://doi.org/10.1098/rspb.2011.1933> PMID: 22048956
82. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science*. 2009; 324: 1454–7. <https://doi.org/10.1126/science.1171908> PMID: 19520963
83. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat Ecol Evol*. Nature Publishing Group; 2017; 1: 1950–1960. <https://doi.org/10.1038/s41559-017-0337-x> PMID: 29038424
84. 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
85. 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
86. Moore MR, Link-Gelles R, Schaffner W, Lynfield R, Lexau C, Bennett NM, et al. Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance. *Lancet Infect Dis*. Elsevier; 2015; 15: 301–9. [https://doi.org/10.1016/S1473-3099\(14\)71081-3](https://doi.org/10.1016/S1473-3099(14)71081-3) PMID: 25656600