Global emergence and population dynamics of divergent serotype 3 CC180 pneumococci

Taj Azarian[1], Patrick K Mitchell[1], Maria Georgieva[1], Claudette M Thompson[1], Amel Ghouila[2], Andrew J Pollard[3], Anna von Gottberg[4], Mignon du Plessis[4], Martin Antonio[5], Brenda A Kwambana-Adams[5], Stuart C Clarke[6,7], Dean Everett[8], Jennifer Cornick[9], Ewa Sadowy[10], Waleria Hryniewicz[10], Anna Skoczynska[10], Jennifer C Moïsi[11], Lesley McGee[12], Bernard Beall[12], Benjamin J Metcalf[12], Robert F Breiman[13], PL Ho[14], Raymond Reid[15], Kate L O'Brien[15], Rebecca A Gladstone[16], Stephen D Bentley[16], William P Hanage[1]

1 Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, MA;
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 OX3 7LJ, UK;
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, Atlantic Road, Fajara, The Gambia;
6 Faculty of Medicine and Institute for Life Sciences and Global Health Research Institute, University of Southampton, S016 6YD, United Kingdom;
7 NIHR Southampton Biomedical Research Centre;
8 Queens Research Institute, University of Edinburgh;
9 Institute of Infection and Global Health, University of Liverpool, Liverpool, UK;
10 National Medicines Institute, Warsaw Poland;
11 Agence de Médecine Préventive, Paris, France;
12 Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA;
13 Global Health Institute, Emory University, Atlanta GA
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
16 The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Patrick Mitchell mitchell.patrick.k@gmail.com
Maria Georgieva georgiev@hsph.harvard.edu
Claudette M Thompson cthompso@hsph.harvard.edu
Amel Ghoulia amel.ghouila@pasteur.tn
Andrew J Pollard andrew.pollard@paediatrics.ox.ac.uk
Anne von Gottberg annev@nicd.ac.za
Mignon du Plessis mignond@nicd.ac.za
Martin Antonio mantonio@mrc.gm
Brenda A Kwambana-Adams bkwambana@mrc.gm
Stuart C Clarke s.c.clarke@soton.ac.uk
Dean Everett dean.everett@liverpool.ac.uk
Jennifer Cornick j.cornick@liverpool.ac.uk
Ewa Sadowy ewasadowy@cls.edu.pl
Waleria Hryniewicz waleria@cls.edu.pl
Anna Skoczynska skoczek@cls.edu.pl
Jennifer Moïsi jmoisi@aamp.org
Lesley McGee afi4@cdc.gov
Bernard Beall beb0@cdc.gov
Benjamin J Metcalf ycm6@cdc.gov
Robert F Breiman rfbreiman@emory.edu
PL Ho plho@hku.hk
Raymond Reid rreid2@jhu.edu
Katherine L O'Brien klobrien@jhu.edu
Rebecca Gladstone rg9@sanger.ac.uk
Stephen D Bentley sdb@sanger.ac.uk
William P Hanage whanage@hsph.harvard.edu

Corresponding Author:
Taj Azarian, PhD MPH
Center for Communicable Disease Dynamics,
Harvard T.H. Chan School of Public Health,
677 Huntington Avenue, Suite 506, Boston, MA 02115
tazarian@hsph.harvard.edu

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 United States, 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 related genotypic differences to phenotypic variations through assessment of capsule charge (zeta potential), capsular polysaccharide shedding, and susceptibility to opsonophagocytic killing, which have previously been associated with carriage duration, invasiveness, and vaccine escape. The recent success of CC180 was associated with a globally emerging lineage termed Clade II, which was estimated by Bayesian coalescent analysis to have first appeared in 1968 [95% HPD: 1939-1989]. Clade II isolates were divergent in non-capsular antigenic composition, competence, and antibiotic susceptibility compared with the pre-PCV13 serotype 3 population. Co-resistance to tetracycline, macrolide, and chloramphenicol resulted from the acquisition of a Tn*916*-like conjugative transposon harbouring *tetM, ermB,* and *cat*. 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 correlated with the emergence of Clade II, implicating genotypic and phenotypic differences. Our analysis emphasizes the need for routine, representative sampling of isolates from disperse 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 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 form 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 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 to 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 different serotypes. Among the 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 [1–4]. 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 US (Figure 1) [5]. In addition, data from a randomized controlled trials on nasopharyngeal carriage [6], post-licensure studies of PCV13 on invasive pneumococcal disease (IPD) [7], and national surveillance data on IPD from England and Wales [8,9], among other studies have shown little vaccine effect on serotype 3.

Multi-locus sequence typing (MLST) data has shown that the majority of serotype 3 isolates globally are a single clonal complex (CC) of closely related genotypes (CC180) also known as the Netherlands^3–31 (PMEN31) clone [10–13], with the exception to Africa where non-CC180 MLST types are prevalent [14]. However, more recent genomic studies have shown CC180 contains multiple distinct lineages [12,15], which MLST is not sufficiently discriminating to detect. While a single closely related lineages (termed ‘Clade I’) accounts for the overwhelming majority of previously studied genomes, most 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 [16]. Initial genomic analysis of these carriage isolates showed a change in the serotype 3 population [Mitchell, PK *et al* - unpublished]. 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 and phenotypic analysis of 301 serotype 3 CC180 isolates from carriage and disease collected from 24 countries over 20 years. Further, we assess 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 [17–19].

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) [12], on-going studies of carriage in Massachusetts, USA (n=27) [20], carriage studies in Southwest United States (n=14) [21], the Bacterial Isolate Genome Sequence database (BIGSDB) (n=10), and the Global Pneumococcal Sequencing (GPS) project (http://www.pneumogen.net/gps/) (n=168). Available data included the year and country of isolation, isolation source, and limited patient demographic data. Significance of changes in the proportion of isolates by clade for each sampling year was tested by comparing yearly proportions to 1000 random deviates of a Dirichlet distribution [22]. Current and past PCV 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 custom Velvet pipeline as previously described [23]. The remaining sequencing data were downloaded from the NCBI SRA database (See supplementary table for a list of accession numbers). If only draft assemblies were available, paired-end 150 bp reads were generated using the BBMap's RandomReads script. *De novo* assemblies for non-GPS isolates were generated using SPAdes v3.10 [24]. Assemblies were then annotated using Prokka v1.12 [25] and pangenome analysis was conducted using Roary [26]. 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 [21]. 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 [27], and were filtered requiring QUAL>50, depth of coverage $\geq$5 and a minimum alternate allele frequency $\geq$0.75 [20,28]. Gubbins v2.1 was used to assess recombination, and coding sequences (CDSs) impacted by recombination blocks were annotated and plotted using Circos [29]. Last, we assessed non-synonymous mutations and recombination in the *comCDE* operon, which has previously been implicated for the low competence of CC180 compared to other pneumococcal lineages [12].

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 [30]. 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 [31]. 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. For major Clade I (hereon referred to as Clade I-α) and Clade II, coalescent analysis was performed using BEAST v1.8.4 (see supplemental methods) [32]. 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 [33,34].

Phenotypic and Genotypic Antibiotic Resistance

Antimicrobial resistance (AMR)-associated genes and SNPs were identified with ARIBA using ARG-ANNOT and CARD databases [35,36]. Penicillin MICs were predicted using WGS data to type transpeptidase domains of penicillin binding proteins [37]. 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 variation in the CPS loci by abstracting the region from reference-based genome assemblies and generating a 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. Recombination among CPS loci was assessed 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 ΔCPS WU2 as previously described [18]. 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 [38]. 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) [39]. The killing assays were performed at multiple dilutions using PCV13 antisera and antisera generated from type 3 polysaccharide (PS) (see supplemental methods) [40]. 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 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. Our phylogenetic analysis 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 (Figure 2A). Outgrouping the phylogeny roots the tree on the branch between those strains previously described as clade II and III, indicating that the previous clade naming confused the tree topology. 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 from the time it was first observed in 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 (Figure 2B). During the same time Clade I-α has significantly decreased, and Clade I-β remained largely unchanged. First observed in Asia in 1999, Clade II is now globally distributed, making up a large proportion of samples from Asia, Africa, and North and South America (Figure 2B). However, Clade II was only observed twice among 97 European strains (Table 1), first appearing in 2003. The abundant Clade I-α was also globally distributed, while Clade I-β isolates are more common than in samples from South America and Asia (Figure 2B).

Clades I-α and II had significant temporal signal, identified by root-to-tip date correlation (Supplemental Figures 1 and 2). 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 (Supplemental Figures 3). 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 (Supplemental Table 1). In addition, the exponential demographic model was preferred to the constant for Clade II. Evolutionary rates were not significantly different between the two clades (Supplemental Figure 4); however, Clade II was significantly younger, with an estimated most recent common ancestor (TMRCA) of 1968 [95% Highest Posterior Density (HPD): 1939-1989] (Supplemental Table 1). The effective population size (N_e) for Clades I-α and II have been increasing, as demonstrated by the SkyGrid N_e plots (Figure 3) and rejection of the constant population size demographic model (Supplemental Table 1). 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 (Figures 4 and 5). Further, isolates from Hong Kong possess a unique Tn*916*-like transposon shared only with isolates found in the dominant North American clade (Figure 5) (See below).

Clade II was found in 11 of 24 countries, 10 of which introduced PCV at some time during our study period (Figure 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). Further, there was no clear pattern in vaccine introduction and Clade II presence.

Recombination and genomic variation

We sought to identify whether genomic variation generated by recombination contributed to the emergence of Clade II. We found that Clades I-α, I-β, and II experienced substantial recombination, impacting 1,179 CDS and diversifying gene content among clades (Supplementary Figures 5 and 6). 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 (Supplementary Figure 5). Comparison of *r/m* values between internal and terminal branches of the phylogeny suggests that the great majority of recombination events in CC180 are ancestral, although more recent events have occurred especially in Clade I- β (Table 2). We subsequently assessed the *comCDE* operon, encoding the competence stimulating peptide and two-component regulatory system, and associated regulatory genes *comAB* to determine whether variations in recombination rates between Clades I-α, I-β, and II were associated with recombination or mutation among these loci. Overall, we found little diversity in the *comCDE* operon (mean pairwise SNP distance = 1 SE 0.5); however, *comD* possessed a non-synonymous mutation (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 significantly diverged between Clades I-β/II and Clade I-α (mean pairwise SNP distance = 1.7 SE 0.7), suggesting phenotypic variation may have resulted from mutation or recombination.

Assessment of recombination events identified that *comB* was located within a recombination event that affected all strains belonging to Clade II.

Recombination impacted a substantial proportion of the genome, focused in known recombination “hotspots” (Figure 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 variation in gene content among clades (Supplemental Figure 7). Two notable differences in MGE that resulted in gene content variation among CC180 clades were the presence of a Tn*916*-like conjugative transposon in Clade II strains (discussed below) and the absence of the 33.3 kb prophage □OXC141 in Clades I-β and II [41,42]. Interestingly, □OXC141, which was putatively acquired by an ancestor of Clade I-α, has been lost multiple times by members of Clade I-α (Figure 5). Most notably, it is absent from some Clade I-α isolates from North America, which includes a number of strains from Massachusetts. In all, □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* (Figure 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 [43]. Three internal and three terminal branches contain recombination events that have impacted SP2194. Major events occurred on each internal branch leading to Clades 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” [44], and here we find that recombination generated variation that subsequently become fixed in dominant clades and sub-clades. Family 2 *pspA* variant was largely associated with Clades I-α and I-β, while the Family I variant was associated with Clade II [45]. Using the *pspC* groupings based on structural and nucleotide variation proposed by Iannelli *et al.*, we found that Clade I possessed a *pspC* variant consistent with Groups 7/8/9, whereas Clade II variants include Groups 1/5 and 2/3/6 [46].

Antimicrobial-resistance (AMR) associated genes

We assessed the presence of AMR-associated genes and mutations among CC180 serotype 3 isolates, finding that variations in AMR-associated genes contributed to differences in genome content among clades. Most significant, Clade II isolates from Hong Kong (n=6) and USA (n=9) harbored a 37.6 kb Tn*916*-like conjugative transposon possessing *tetM* and *ermB* (Supplemental Figure 8). These strains also possess chloramphenicol acetyltransferase (*cat*). For 13 of 15 strains with Tn*916*, antibiotic susceptibility testing was available in GPS metadata or from previous publications [15]. In general, Tn*916* conferred 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 [47]. 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 Tn*916* transposons among Clade II isolates (Supplemental Figure 8). 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 Tn*916* transposons. In addition, 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). These four isolates also possessed

Tn*916*. 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 (Supplemental Table 2).

Capsular Polysaccharide Variation

Interrogation of inferred recombination events showed that the CPS loci were unaffected by homologous recombination (Figure 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 (Supplemental Figure 9). Having ruled out significant genotypic variation, we continued to assess phenotype. While the surface charge of clade II isolates were slightly more negative among tested strains this difference was not significant (Supplemental Figure 10A). In addition, comparison of isolates from Clades I-α and II also found no significant difference in capsule shedding (Supplemental Figure 10B). Last, we observed neutrophil-mediated opsonophagocytic killing of Clade II isolates 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

The emergent CC180 Clade II is globally distributed and has recently increased in prevalence and effective population size. 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. When considering prevalence data and geography alone, the diaspora of Clade II appears to 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, which is the dominant serotype 3 clone globally. While neither use of PCV at a country level, nor timing on PCV introduction appears to be associated with the rise of Clade II, we find that Clade II has a higher rate of recombination and increased prevalence of antibiotic resistance compared to the other CC180 clades. These two observations are likely contributors to the emergence of Clade II, notwithstanding absence of definitive evidence of differential vaccine effectiveness among clades.

While phylogeographic analysis was unable to confidently infer the ancestral migration events of Clade II, the presence of Tn*916* among strains from Hong Kong and the United States provides evidence of migration from Asia to North America. In at least two instances, distinct Tn*916* 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 and that strains circulating in Asia were the ancestral population 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. Tn*916* confers macrolide resistance, with the rare exception of strains possessing a mutation in *ermB* [48]. Macrolide resistance has previously been described among serotype 3 ST180 isolates from Japan and Italy in 2003 and 2001, respectively, where macrolide-resistant serotype 3 strains are prevalent [13,49,50]. 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 able 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 Tn*916*, further supporting a Southeast Asian origin.

As PCV use has previously been associated with serotype replacement [51,52], we considered that serotype 3 may have increased in those settings where PCV was introduced. Further, variation in effectiveness of the serotype 3 component of PCV13 among CC180 clades may have disproportionally precipitated an increase in Clade II. Here, we find no evidence that PCV use, albeit at a country level, has shaped the emergence of serotype 3 CC180. However, a limitation is that these comparisons ignore 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 a more in depth analysis of the putative association of PCV with emergence of Clade II. In absence of a clear epidemiological etiology for the increase of Clade II, we investigated phenotypic and genomic variation between Clades I-α and II.

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 low carriage duration, which in turn reduces antibiotic exposure [53]. Here we find chloramphenicol, macrolide, tetracycline, and first step penicillin resistance more frequent among Clade II isolates, providing one possible explanation for its recent emergence and subsequent increase in prevalence. In fact, increased macrolide usage in North America has previously been associated with an increase in macrolide resistant *Staphylococcus aureus* and nonsusceptible *S. pneumoniae* [54,55].

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 □OXC141 was generally absent among Clades I-β and II and has been lost by a North American sub-clade of Clade I-α [41,42]. Temperate bacteriophages may be associated with fitness defects among pneumococci or changes in virulence and competence [56,57]. How the presence or absence of □OXC141 relates to the relative success of serotype 3 CC180 clades is unclear; however, the absence of □OXC141 from Clade II 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 [21,58]. 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 [21,46,59,60]. 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 [61]. Taken together, these differences in antigenic profiles may be contributing factors in the emergence of Clade II. Notably, protein antigen variants were conserved among Clade I-α isolates and polymorphic in Clades I-β and II. For instance, most isolates in Clade I-α possess a *pspC* variant corresponding to Groups 7, 8, and 9, which have been grouped here because they share the same structural organization [46]. 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.

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 Netherlands^{3}–31 clone. 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 [12], mainly dominated by micro-recombination events [62]. 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 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 genomic and *in vitro* studies [12,63], we found that the emergent Clades II and I-β have higher rates of recombination (ρ/Θ) and a greater proportion of macro-recombination events than the previously dominant Clade I-α. Therefore, Clade I-α may possess a genomic defect that reduces its competence, a mutation that is absent in Clade I-β and Clade II. Consistent with this hypothesis, the Clade I-α *comCDE* operon generates an antisense transcript, an observation that likely relates to the low observed recombination rate of Clade I-α [12]. We show amino acid changes in the *comCDE*, specifically *comD*, segregate Clade I-α from I-β, and II. We also found variation in the Clp protease gene, which regulates the *comCDE* operon, and is therefore associated with competence rates among pneumococci and other bacteria [43,64]. Taken together, we suggest that variation in the competence machinery of Clade II likely contributes to increased recombination rates, allowing Clade II isolates to alter their antigenic profile and acquire multi-drug resistance [65,66]. Indeed, higher rates of recombination have been associated with increased likelihood of antimicrobial resistance determinant acquisition [67].

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* [17,19,40] . 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 [18]. We therefore compared this between Clades I-α and II, finding no significant difference in charge between clades, suggesting capsular properties are similar. Where we did find a significant phenotypic difference between the clades, it was not easy to relate 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. 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 [9,68], 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, and moreover 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* [40].

Taken together, our findings point to a likely 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 anti-protein antibodies 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 emerging clade, divergent in antigenic composition, antibiotic susceptibility, and competence. Based on OPKA, we also find 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.

Overall, our analysis emphasizes the need for routine, representative sampling of isolates from disperse 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.

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

Figure Legend

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

Figure 2. A) Rooted maximum likelihood phylogeny of *S. pneumoniae* serotype 3 CC180 isolates (n=301). 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 (data not shown) and therefore represented the closest out-group. Mean pairwise between-clade SNP difference are presented in the shaded box with color corresponding to the phylogeny. Bootstrap values are labeled on major clades. 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. 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.

Figure 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 both Clades I-α and II have been exponentially increasing.

Figure 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.

Figure 5. Phylogeny, polymorphic protein antigen variants, and antibiotic resistance. Midpoint rooted maximum-likelihood phylogeny with geographic region of isolation represented as colored tip shapes. Clade I-α, I-β, and II are shaded consistent with Figure 2A. 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 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).

Figure 6. PCV implementation by county 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.

Figure 7. Circos plot of 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.

Table 1. Clade composition of CC180 serotype 3 isolates by geographic location. Percentages are based on 295 isolates for which the collection date and country are known.

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%

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 braches is considered ancestral while terminal braches 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**

References

1. 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. doi:10.1086/374624
2. 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. doi:10.1186/1471-2334-4-21
3. 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.
4. 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.
5. 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. doi:10.1016/S1473-3099(14)71081-3
6. 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. doi:10.1093/cid/cit428
7. 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. doi:10.1016/S1473-3099(14)70822-9
8. 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. doi:10.1097/INF.0b013e318275614b

9. 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. The Lancet. 2014; doi:10.1016/S1473-3099(14)70822-9
10. Mosser JF, Grant LR, Millar E V, 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. doi:10.1371/journal.pone.0079578
11. 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. doi:10.1128/JCM.42.10.4449-4452.2004
12. 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. doi:10.1371/journal.pgen.1003868
13. 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. doi:10.1111/j.1469-0691.2009.02935.x
14. 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. doi:10.1128/JCM.01011-09
15. 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. doi:10.1038/ng.2625

16. 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.
17. 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. doi:10.1128/CVI.00591-15
18. 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. doi:10.1128/IAI.00724-13
19. 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.
20. 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. doi:10.1038/ng.2625
21. 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. doi:10.1093/infdis/jiw628
22. Minka T. Estimating a Dirichlet distribution. Technical report, MIT; 2000.
23. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–9. doi:10.1101/gr.074492.107
24. 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. doi:10.1089/cmb.2012.0021
25. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–9. doi:10.1093/bioinformatics/btu153
26. 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. doi:10.1093/bioinformatics/btv421

27. 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. doi:10.1093/bioinformatics/btp352

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

29. 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-. doi:10.1093/nar/gku1196

30. 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. doi:10.1093/bioinformatics/bti191

31. Camilli R, Del Grosso M, Iannelli F, Pantosti A. New genetic element carrying the erythromycin resistance determinant erm(TR) in *Streptococcus pneumoniae*. Antimicrob Agents Chemother. 2008;52: 619–625. doi:10.1128/AAC.01081-07

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

33. 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. doi:10.1371/journal.pcbi.1003537

34. Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. 2014; btu201-. doi:10.1093/bioinformatics/btu201

35. 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.

36. 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.

37. 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. doi:10.1128/mBio.00756-16

38. 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. doi:10.1128/CVI.00591-15

39. 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. doi:10.1128/CVI.00371-13

40. Zhang F, Lu Y-J, Malley R. Multiple antigen-presenting system (MAPS) to induce comprehensive B- and T-cell immunity. Proc Natl Acad Sci. National Academy of Sciences; 2013;110: 13564–9. doi:10.1073/pnas.1307228110

41. 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. doi:10.1128/JB.01272-08

42. 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. doi:10.3389/fmicb.2015.00026

43. 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. doi:10.1128/JB.183.24.7295-7307.2001

44. 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. doi:10.1186/gb-2010-11-10-r107

45. 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. doi:10.1128/IAI.68.10.5889-5900.2000

46. Iannelli F, Oggioni MR, Pozzi G. Allelic variation in the highly polymorphic locus pspC of *Streptococcus pneumoniae*. Gene. 2002;284: 63–71. doi:10.1016/S0378-1119(01)00896-4

47. 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. doi:10.1016/j.cmi.2016.08.001

48. Rice LB, Carias LL, Marshall SH, Hutton-Thomas R, Rudin S. Characterization of Tn5386, a Tn916-related mobile element. Plasmid. 2007;58: 61–7. doi:10.1016/j.plasmid.2007.01.002

49. 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. doi:10.1007/s10156-008-0600-9

50. 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. doi:10.1128/JCM.01229-06

51. 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. doi:10.1016/j.epidem.2010.03.005

52. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. Lancet. 2011;378: 1962–73. doi:10.1016/S0140-6736(10)62225-8

53. 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. doi:10.1073/pnas.1617849114

54. 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-. doi:10.1073/pnas.1401006111

55. 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. doi:10.1093/cid/cir443

56. 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. doi:10.1128/JB.01556-14

57. 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. doi:10.1093/molbev/mss279

58. 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. doi:10.1371/journal.ppat.1006137

59. 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. doi:10.1111/1469-0691.12286

60. 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. doi:10.1128/IAI.71.1.75-85.2003

61. 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. doi:10.1128/IAI.71.1.75-85.2003

62. 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. doi:10.1371/journal.pgen.1004300

63. 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. doi:10.3201/eid1211.060414

64. Hui FM, Zhou L, Morrison DA. Competence for genetic transformation in *Streptococcus pneumoniae*: organization of a regulatory locus with homology to two lactococcin A secretion genes. Gene. 1995;153: 25–31. doi:10.1016/0378-1119(94)00841-F

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

66. 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. doi:10.1098/rspb.2011.1933

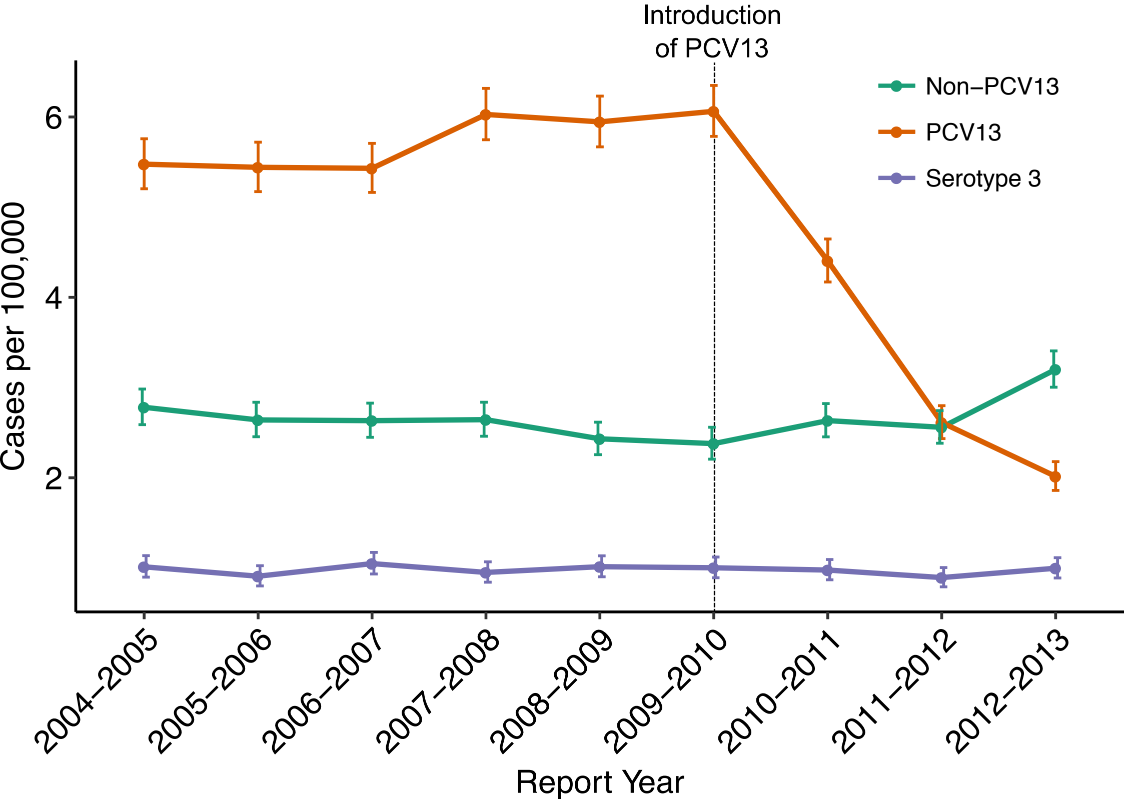
67. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science. 2009;324: 1454–7. doi:10.1126/science.1171908

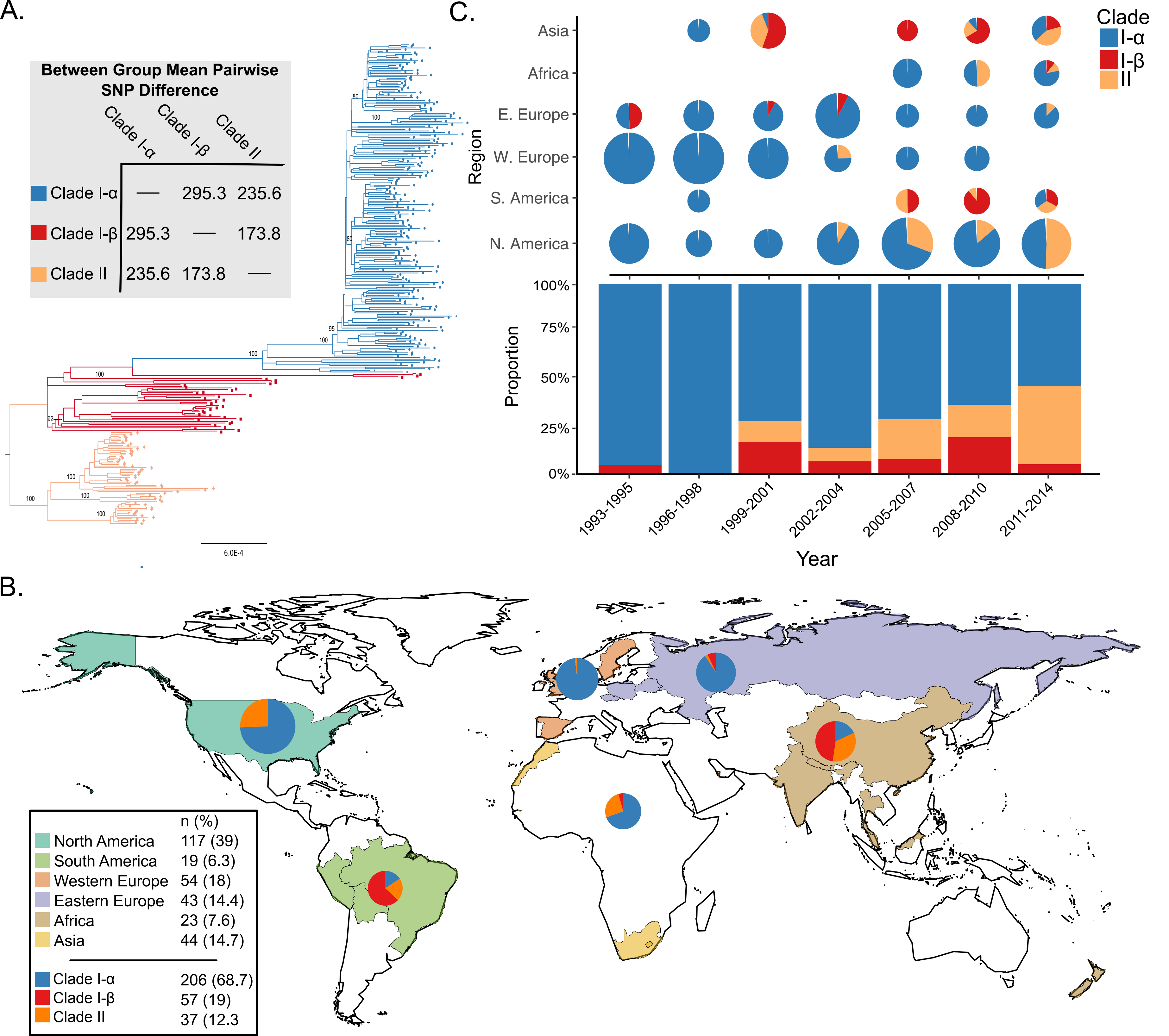
68. Moore MR, Link-Gelles R, Schaffner W, Lynfield R, Lexau C, Bennett NM, et al. Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance. Lancet Infect Dis. Elsevier; 2015;15: 301–9. doi:10.1016/S1473-3099(14)71081-3

Supporting Information Legends

S1 – Supplemental methods, tables (2), and figures (11)

S2 – File containing accession numbers and associated metadata

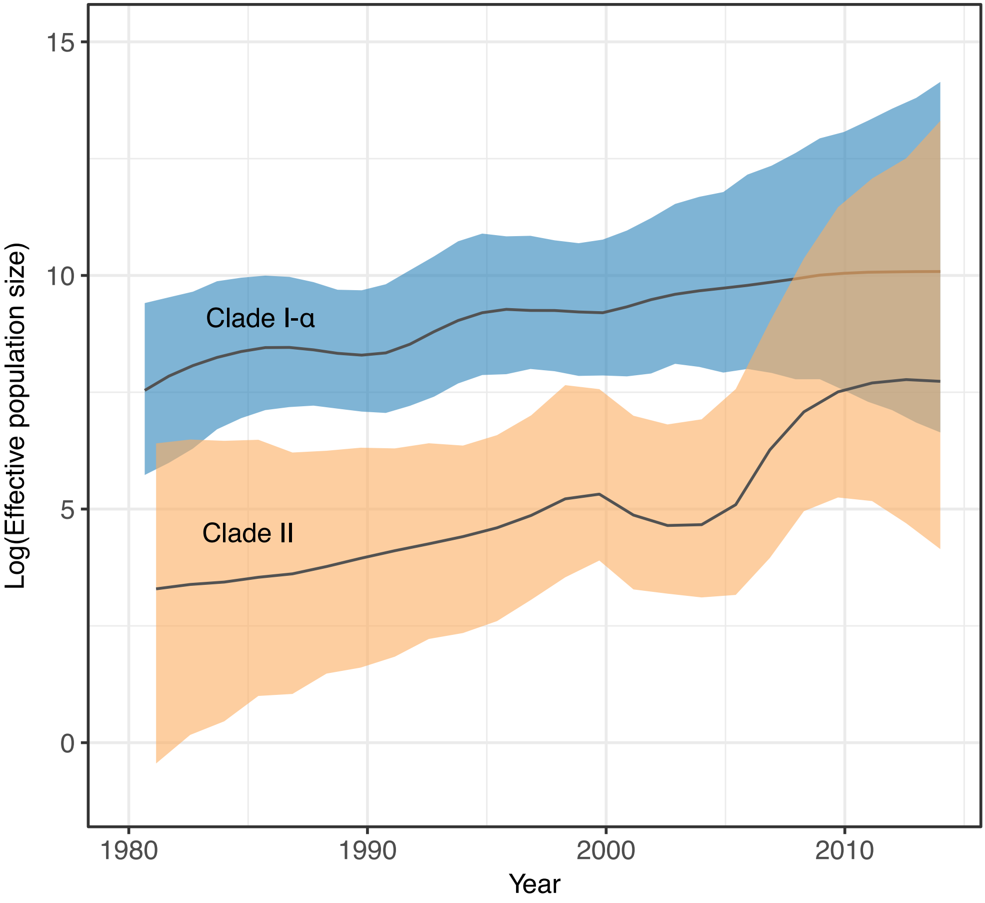

Introduction
of PCV13
Non−PCV13
PCV13
Serotype 3
Cases per 100,000
6
4
2
2004−2005
2005−2006
2006−2007
2007−2008
2008−2009
2009−2010
2010−2011
2011−2012
2012−2013
Report Year

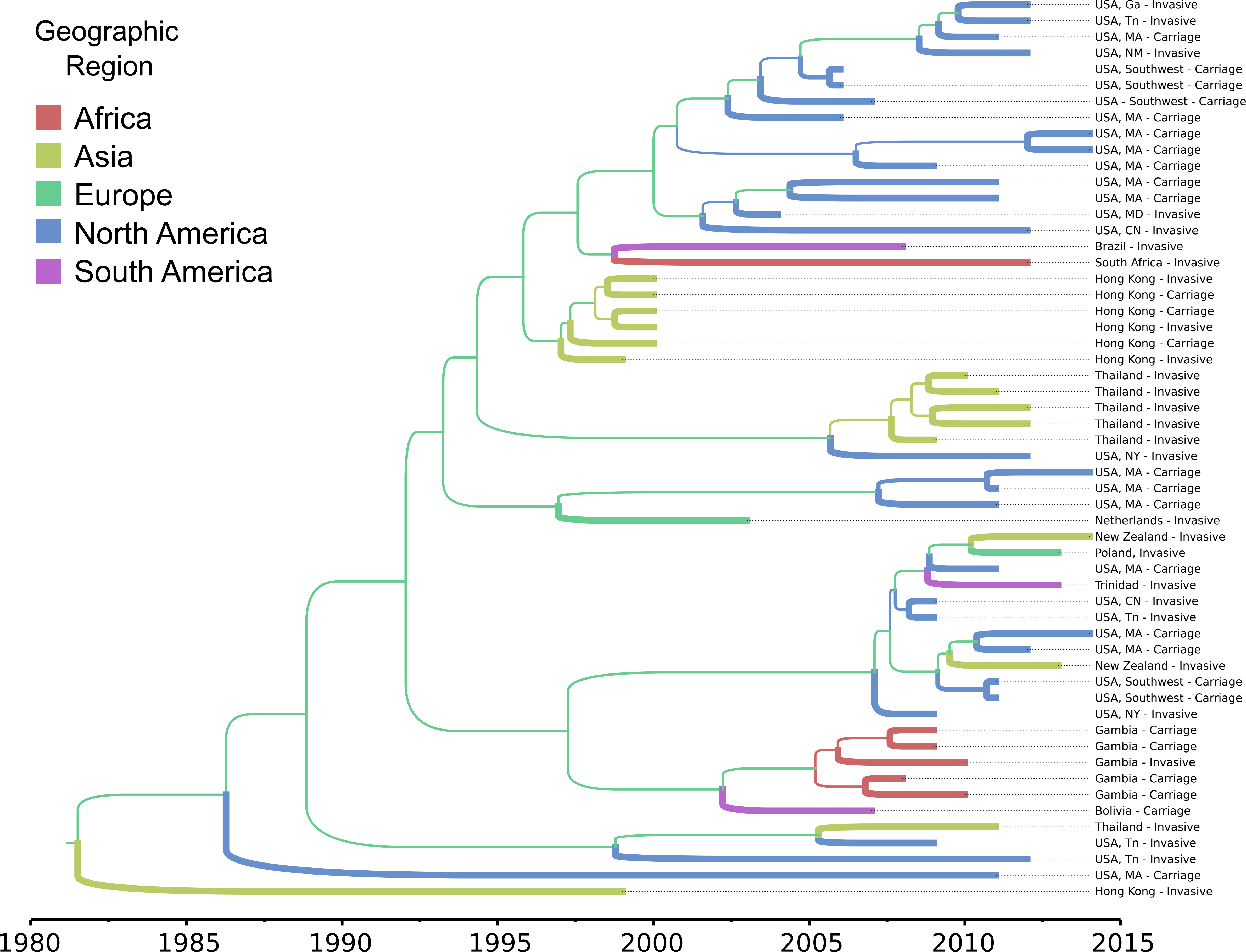


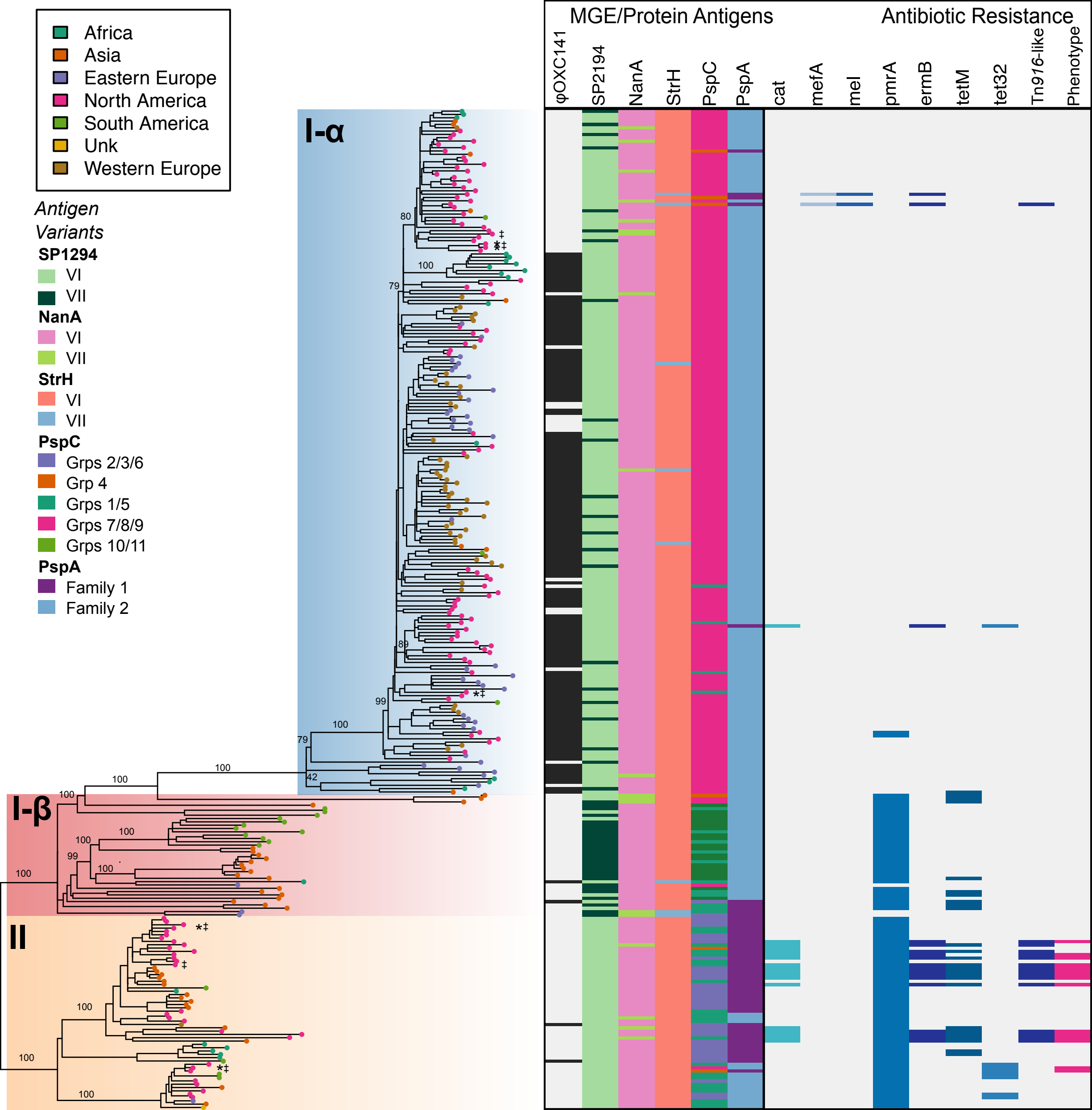
Between Group Mean Pairwise SNP Difference

	Clade I-α	Clade I-β	Clade II
🟦 Clade I-α	—	295.3	235.6
🟥 Clade I-β	295.3	—	173.8
🟧 Clade II	235.6	173.8	—

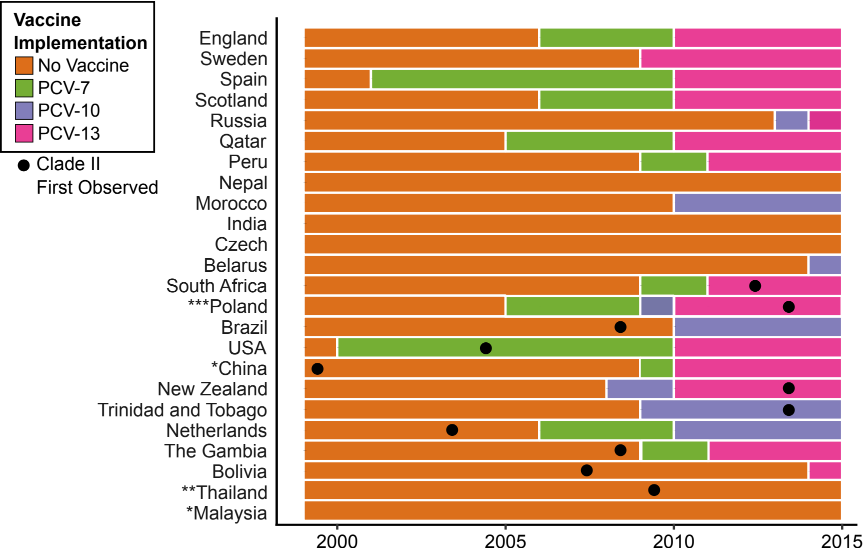
	n (%)
North America	117 (39)
South America	19 (6.3)
Western Europe	54 (18)
Eastern Europe	43 (14.4)
Africa	23 (7.6)
Asia	44 (14.7)
Clade I-α	206 (68.7)
Clade I-β	57 (19)
Clade II	37 (12.3

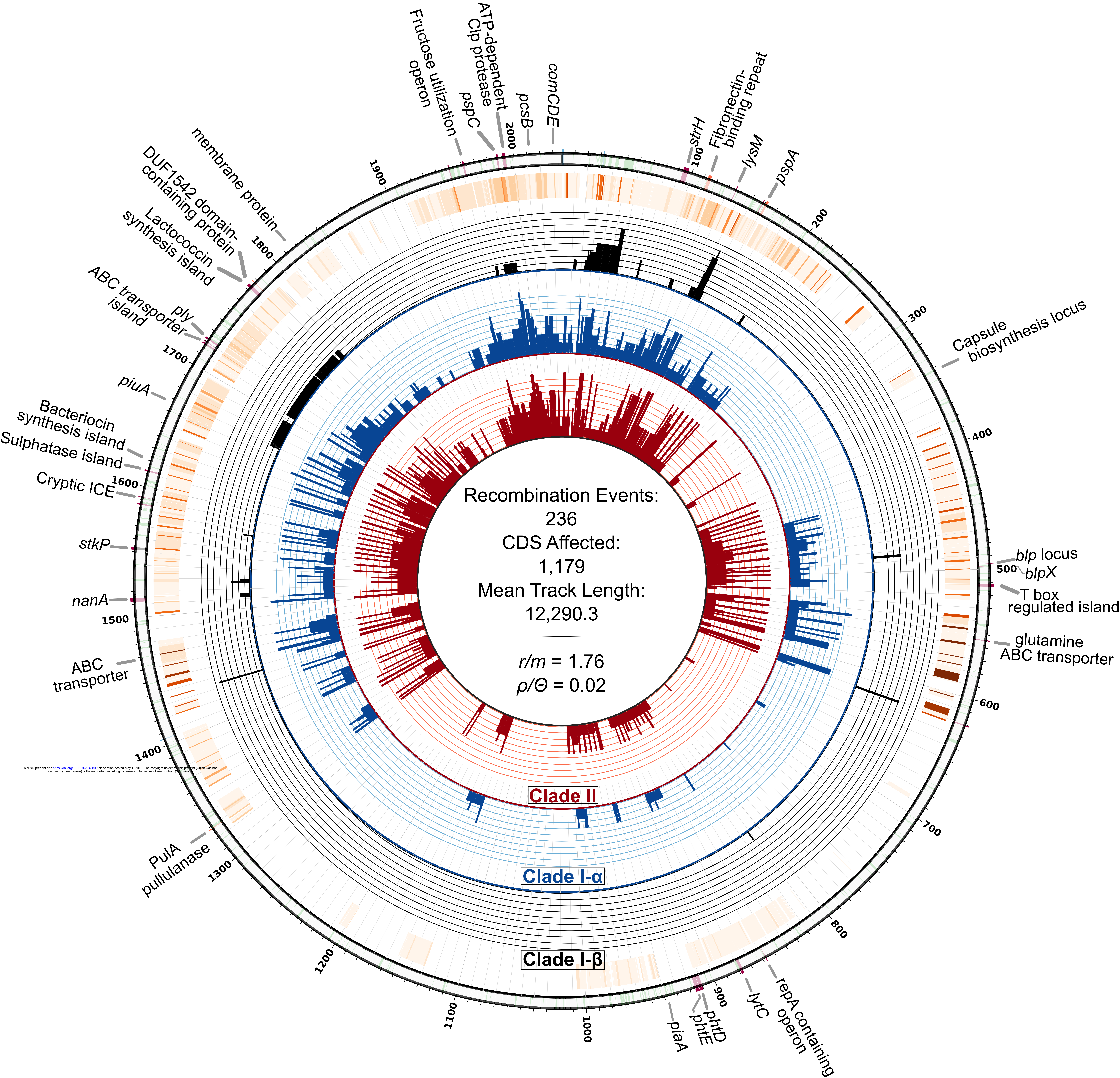

Clade I-α
Clade II
15
10
5
0
Log(Effective population size)
1980
1990
2000
2010
Year


Geographic Region
Africa
Asia
Europe
North America
South America
USA, Ga - Invasive
USA, Tn - Invasive
USA, MA - Carriage
USA, NM - Invasive
USA, Southwest - Carriage
USA, Southwest - Carriage
USA - Southwest - Carriage
USA, MA - Carriage
USA, MA - Carriage
USA, MA - Carriage
USA, MA - Carriage
USA, MA - Carriage
USA, MA - Carriage
USA, MD - Invasive
USA, CN - Invasive
Brazil - Invasive
South Africa - Invasive
Hong Kong - Invasive
Hong Kong - Carriage
Hong Kong - Carriage
Hong Kong - Invasive
Hong Kong - Carriage
Hong Kong - Invasive
Thailand - Invasive
Thailand - Invasive
Thailand - Invasive
Thailand - Invasive
Thailand - Invasive
USA, NY - Invasive
USA, MA - Carriage
USA, MA - Carriage
USA, MA - Carriage
Netherlands - Invasive
New Zealand - Invasive
Poland, Invasive
USA, MA - Carriage
Trinidad - Invasive
USA, CN - Invasive
USA, Tn - Invasive
USA, MA - Carriage
USA, MA - Carriage
New Zealand - Invasive
USA, Southwest - Carriage
USA, Southwest - Carriage
USA, NY - Invasive
Gambia - Carriage
Gambia - Carriage
Gambia - Invasive
Gambia - Carriage
Gambia - Carriage
Bolivia - Carriage
Thailand - Invasive
USA, Tn - Invasive
USA, Tn - Invasive
USA, MA - Carriage
Hong Kong - Invasive
1980
1985
1990
1995
2000
2005
2010
2015



Africa
Asia
Eastern Europe
North America
South America
Unk
Western Europe
Antigen Variants
SP1294
VI
VII
NanA
VI
VII
StrH
VI
VII
PspC
Grps 2/3/6
Grp 4
Grps 1/5
Grps 7/8/9
Grps 10/11
PspA
Family 1
Family 2
I-α
I-β
II
80
100
79
89
99
100
79
42
100
100
100
100
100
99
100
100
100
100
100
*‡
‡
MGE/Protein Antigens
Antibiotic Resistance
φOXC141
SP2194
NanA
StrH
PspC
PspA
cat
mefA
mel
pmrA
ermB
tetM
tet32
Tn916-like
Phenotype


Vaccine Implementation
No Vaccine
PCV-7
PCV-10
PCV-13
● Clade II First Observed
England
Sweden
Spain
Scotland
Russia
Qatar
Peru
Nepal
Morocco
India
Czech
Belarus
South Africa
***Poland
Brazil
USA
*China
New Zealand
Trinidad and Tobago
Netherlands
The Gambia
Bolivia
**Thailand
*Malaysia
2000
2005
2010
2015


Recombination Events:
236
CDS Affected:
1,179
Mean Track Length:
12,290.3
r/m = 1.76
ρ/Θ = 0.02
Clade II
Clade I-α
Clade I-β
comCDE
pcsB
ATP-dependent Clp protease
pspC
Fructose utilization operon
2000
strH
100
Fibronectin-binding repeat
lysM
pspA
200
300
Capsule biosynthesis locus
400
blp locus
500
blpX
T box regulated island
glutamine ABC transporter
600
700
800
repA containing operon
lytC
900
phtD
phtE
piaA
1000
1100
1200
1300
PulA pullulanase
1400
ABC transporter
1500
nanA
stkP
Cryptic ICE
1600
Sulphatase island
Bacteriocin synthesis island
piuA
1700
ply
ABC transporter island
Lactococcin synthesis island
DUF1542 domain-containing protein
1800
membrane protein
1900