nature genetics

Population genomics of post-vaccine changes in pneumococcal epidemiology

Nicholas J Croucher[1,2], Jonathan A Finkelstein[3,4], Stephen I Pelton[5], Patrick K Mitchell[1], Grace M Lee[3,6,7], Julian Parkhill[2], Stephen D Bentley[2,8,10], William P Hanage[1,10] & Marc Lipsitch[1,9,10]

Whole-genome sequencing of 616 asymptomatically carried *Streptococcus pneumoniae* isolates was used to study the impact of the 7-valent pneumococcal conjugate vaccine. Comparison of closely related isolates showed the role of transformation in facilitating capsule switching to non-vaccine serotypes and the emergence of drug resistance. However, such recombination was found to occur at significantly different rates across the species, and the evolution of the population was primarily driven by changes in the frequency of distinct genotypes extant before the introduction of the vaccine. These alterations resulted in little overall effect on accessory genome composition at the population level, contrasting with the decrease in pneumococcal disease rates after the vaccine's introduction.

© 2013 Nature America, Inc. All rights reserved.

S. pneumoniae is a human nasopharyngeal commensal and respiratory pathogen that represents a major cause of pneumonia, bacteremia and meningitis. The bacterium's best-characterized virulence factor is its polysaccharide capsule, of which there are more than 90 serologically distinct variants (serotypes) that are associated with markedly different patterns of carriage and disease[1]. These variants may be categorized into serogroups, defined as sets of serotypes that cross-react with common antisera: for instance, serogroup 23 comprises serotypes 23F, 23A and 23B. Such serology forms the basis of much pneumococcal epidemiology, and the capsule is also the target of all licensed vaccine formulations.

In 2000, the polysaccharide-protein conjugate vaccine (PCV7) was introduced for the routine immunization of children in the United States. This vaccine contained antigens designed to protect against seven pneumococcal serotypes (vaccine types, VTs): 4, 6B, 9V, 14, 18C, 19F and 23F. It was anticipated that it would also provide some cross-protection against vaccine-related serotypes (VRTs) within the same serogroups as these seven targets. Since the introduction of PCV7, the incidence of invasive pneumococcal disease (IPD) caused by VT strains has decreased significantly across the United States[2], accompanied by a smaller increase in disease caused by non-VT, including VRT, pneumococci[3–6]. Surveillance in Massachusetts has monitored contemporaneous changes in the asymptomatically carried pneumococcus population. Surveys of over 600 children in each of the winters of 2000–2001, 2003–2004 and 2006–2007 found that between 23 and 30% of children carried pneumococci, with no evidence of a decrease in carrier prevalence after the introduction of PCV7 (refs. 7–9). This reflected an increase in the prevalence of non-VT pneumococci as the VT strains were gradually eliminated. Nevertheless, by 2003, the rate of IPD occurrence in children under 5 years of age in the state decreased by 69% relative to the pre-vaccine incidence rate[10], with VRT 19A and non-VT 7F emerging as the dominant IPD serotypes after the introduction of PCV7 (refs. 11,12). This reduction in disease incidence, while the levels of pneumococcal carriage remained nearly unchanged, suggests that the non-VT strains that dominated after the introduction of PCV7 were, on average, less invasive than the VT strains that they replaced[11,12]. Patterns of antibiotic resistance were also observed to change[13,14]; for instance, several of the epidemic multidrug-resistant clones identified by the Pneumococcal Molecular Epidemiology Network[15] (PMEN) predominately express VT capsules and have consequently declined in frequency since the introduction of the vaccine.

These substantial alterations to the pneumococcal population resulting from the selective pressure of PCV7 might be expected to affect genetic diversity across the chromosome, akin to the 'periodic selection' expected in a genetically monomorphic bacterium[16]. However, the ability of the pneumococcus to recombine by natural transformation suggests that successful lineages that were mainly vaccine type before PCV7 may persist through the acquisition of non-VT capsules, a process called serotype or capsule switching[17] that has been observed throughout the history of the pneumococcus[18,19].

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [2]Pathogen Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA. [4]Division of General Pediatrics, Boston Children's Hospital, Boston, Massachusetts, USA. [5]Maxwell Finland Laboratory for Infectious Diseases, Boston University Medical Center, Boston, Massachusetts, USA. [6]Department of Laboratory Medicine, Boston Children's Hospital, Boston, Massachusetts, USA. [7]Division of Infectious Diseases, Department of Medicine, Boston Children's Hospital, Boston, Massachusetts, USA. [8]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [9]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, Massachusetts, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to W.P.H. (whanage@hsph.harvard.edu).

Received 12 November 2012; accepted 5 April 2013; published online 5 May 2013; doi:10.1038/ng.2625

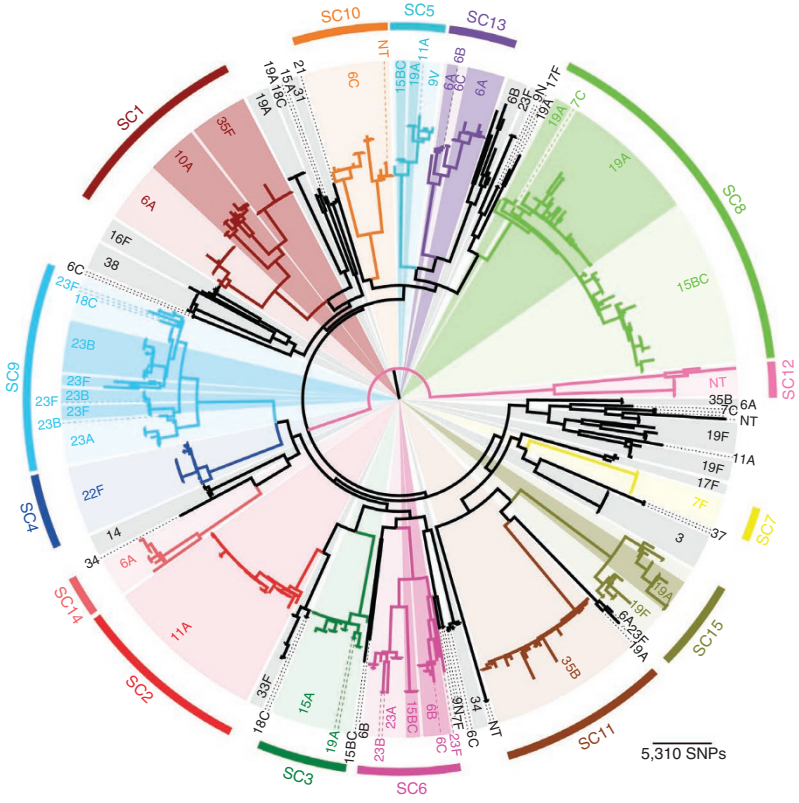

Figure 1 Structure of the pneumococcal population. The maximum-likelihood phylogeny was generated using 106,196 polymorphic sites within a 1.14-Mb codon alignment of 1,194 core COGs. The coloring of this phylogeny delineates sequence clusters defined using the same alignment: 15 monophyletic sequence clusters are labeled, with a sixteenth polyphyletic cluster indicated by the terminal branches of the tree colored black. Within the monophyletic sequence clusters, light background shading indicates one particular serotype, with darker shading and dotted lines used to indicate groups of isolates with alternative serotypes. NT, non-typeable isolates.

To better understand the interplay of clonal replacement and adaptation through recombination in the population's response to vaccination, whole-genome sequencing was performed on isolates obtained from Massachusetts children between 2000 and 2007.

RESULTS

Population structure of isolates

A total of 616 isolates were used in this analysis, including 133 from 2001, 203 from 2004 and 280 from 2007 (**Supplementary Table 1**). In terms of their serotype distribution, these populations represent the pneumococcal community at the point at which PCV7 was introduced, a community with increased diversity after the introduction of the vaccine and a return to equilibrium, respectively[20]. *De novo* genome assembly produced sequences ranging in length from 1.98 to 2.19 Mb (**Supplementary Fig. 1**). The 1,231,516 putative protein-coding sequences extracted from these assemblies were grouped into 5,442 clusters of orthologous genes (COGs). The prevalence of these COGs in the population followed a U-shaped distribution (**Supplementary Fig. 2**), similar to that calculated using a smaller set of pneumococcal genomes generated with older technologies[19].

A total of 1,194 COGs were present in a single copy in all genomes and were therefore used to define a 'core' genome. These were used to generate a 1.14-Mb codon alignment, containing 106,196 polymorphic sites, upon which phylogenetic and population clustering analyses were based (**Fig. 1**). This approach identified 15 sequence clusters corresponding to clades within the tree, ranging in size from 10 to 98 isolates, with a sixteenth group constituting a polyphyletic cluster of the remaining low-frequency genotypes. All of the monophyletic sequence clusters showed a high level of consistency with previous genotyping (multilocus sequence typing, MLST; **Supplementary Fig. 3**)[14,21]. The tree was rooted on the longest branch, separating a diverse lineage of unencapsulated strains (SC12) from the rest of the population. The genotypes in this outgroup, previously associated with outbreaks of conjunctivitis using MLST data[22], are clearly distinct from those of both the other sequence clusters and *Streptococcus pseudopneumoniae* when placed in the context of other mitis-group streptococci (**Supplementary Fig. 4**). In contrast, other unencapsulated strains in the collection appear to reflect sporadic loss of the capsule (**Supplementary Table 2**).

Diversification within sequence clusters

To reconstruct the recent evolutionary history of the population, whole-genome alignments were generated for each sequence cluster through mapping sequence reads for member isolates against a representative reference assembly. Horizontally acquired variation was then identified using previously described methods[23] (**Supplementary Figs. 5–19**). This approach allows the calculation of the net rate at which polymorphisms are accumulated through import of sequence from other lineages via transformation relative to the mutation rate (r/m) for each sequence cluster. The estimates for this r/m parameter ranged from 0.06 (SC7, comprising serotype 7F isolates) to 34.06 (SC15, the multidrug-resistant PMEN14 lineage[15]). When representing the behavior of each sequence cluster by the distribution of r/m statistics derived from each branch of the relevant phylogeny, the differences between sequence clusters were found to be significant, whether comparing per-site r/m values (Kruskal-Wallis test, $\chi^2 = 45.71$, degrees of freedom = 14, $P = 3.97 \times 10^{-5}$) or the ratio of homologous recombination events to point mutations ($\chi^2 = 54.6$, $P = 9.96 \times 10^{-7}$). Correspondingly, the distributions of detected homologous recombination lengths are consistent between sequence clusters (**Supplementary Fig. 20**).

The range of isolation dates was used to establish a molecular clock. Fitting an evolutionary model to four sequence clusters that showed evidence of genetic divergence over time (**Supplementary Fig. 21**) yielded independent estimates of the point mutation rate (**Supplementary Fig. 22**). The consensus rate of 1.0×10^{-6} to 1.5×10^{-6} mutations per base per year is similar to that calculated using a global collection of strains[23]. This indicates that there is little variation in the substitution rate across the species, with isolates accumulating two or three polymorphisms each year. Given the number of observed transformation events and point mutations within each sequence cluster, this implies a variable net rate of sequence import into the sequence clusters: around one transformation event every 6 to 9 years for SC2 isolates compared with one transformation event every 1 to 2 years for SC15 isolates.

Serotype switching and replacement

Transformation has an important role in facilitating the process of serotype switching (**Fig. 2**). SC15 provides a simple example of switching

© 2013 Nature America, Inc. All rights reserved.

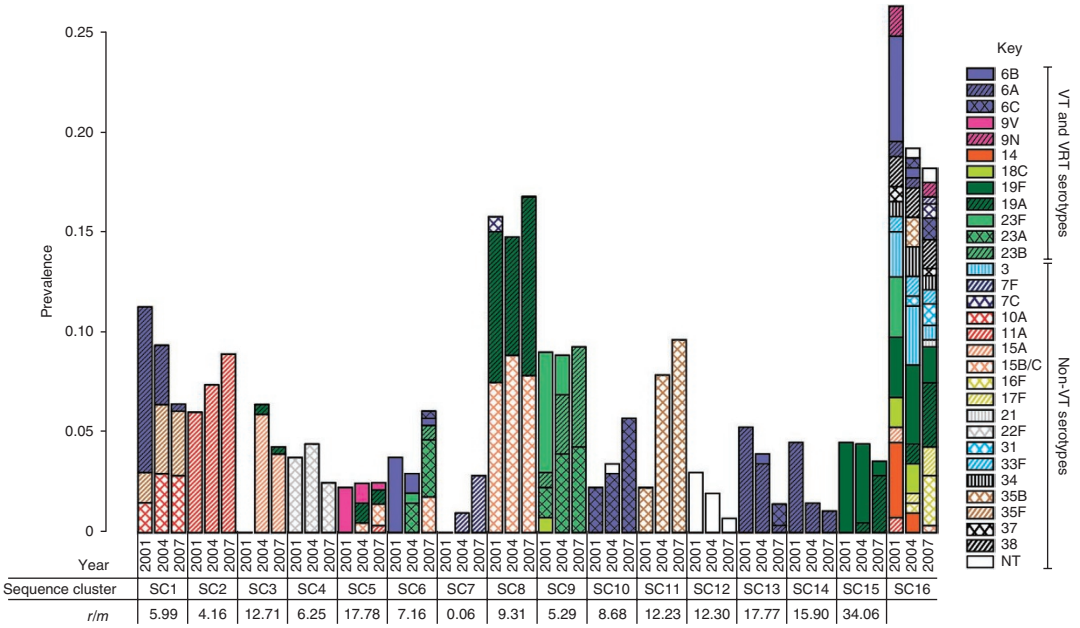

Figure 2 Dynamics of the carried pneumococcal population. The proportion of the population constituted by each of the sequence clusters in the three collection periods is plotted as adjacent bars. Each bar is divided according to the serotype composition of the sequence cluster in each year; VT capsule serotypes are represented by solid fill, VRT capsule serotypes are represented by solid fill (of the color of the vaccine type from the same serogroup) overlaid with black hatching, and non-VT capsule serotypes are represented by colored hatched patterns on a white background. Underneath the graph, per-site *r/m* statistic for the 15 monophyletic clades are shown, as calculated from the analyses used for **Supplementary Figures 5–19**. An equivalent statistic could not be calculated for the polyphyletic SC16.

leading to vaccine escape, previously inferred from MLST[17], with a single transformation event at the capsule biosynthesis (*cps*) locus leading to the emergence of serotype 19A variants (**Supplementary Fig. 19**). These variants are absent in the 2001 sample but account for 80% of the SC15 isolates in 2007. The most recent common ancestor of these serotype 19A strains is predicted to have existed in 1997 (95% credibility interval of 1984–2001). As the acquisition of a serotype 19A capsule by SC15 must have preceded this most recent common ancestor, it seems likely that the emergence of serotype 19A in SC15 represents an increase in the frequency of a variant already extant before the introduction of PCV7.

The emergence of serotype 19A variants within SC15 represents the smallest genetic distance between isolates with different capsule types within a monophyletic sequence cluster; all other such cases of capsule switching involved more distinct genotypes that, by implication of the observed molecular clock, shared a common ancestor further in the past. Hence, there was no evidence for serotype switching events occurring in the monophyletic sequence clusters after the introduction of PCV7. Instead, selection from variation in the pre-vaccine population by PCV7 seems to have been more common: a clear example can be observed within SC9 (**Fig. 3**). Two-thirds of the strains in this sequence cluster expressed the VTs 23F capsule in 2001, with this serotype decreasing in frequency to an undetectable level in 2007. Instead, all representatives of the sequence cluster were VRTs 23A and 23B, which have lower odds ratios (ORs) for causing IPD than serotype 23F in Massachusetts[11], thereby likely contributing to the decrease in IPD rates. However, the original serotype of SC9 was reconstructed as being 23A; in the era before the use of PCV7, the serotype 23F variants, generated through three independent transformation events, seem to have been the most successful. PCV7 reversed this trend, as the serotype 23A isolates, in parallel with the serotype 23B strains formed by a further three separate switches within the phylogeny, seem to have filled the niche vacated by the eliminated serotype 23F isolates. Hence, the vaccine caused no change in the prevalence of SC9, nor did it reduce the diversity of the sequence cluster. Apparent serotype switching that actually reflects the emergence of deep-branching sister taxa was also observed for SC1 and SC6 (**Fig. 1**). By contrast, in the multidrug-resistant SC5 (**Supplementary Fig. 9**) and SC15 (**Supplementary Fig. 19**) lineages, the emergence of serotype 19A variants involved a recent bottleneck constraining post-PCV7 diversity, mirroring what was observed in the PMEN1 clone[23].

Notwithstanding the substantial turnover of lineages within the population, the effect of vaccination on the prevalence of individual COGs was modest (**Fig. 4** and **Supplementary Fig. 23**). The composition of the accessory genome was generally stable throughout the sampling period. Exceptions included four COGs involved in the synthesis of serogroup 6 capsules: one represented the *wciN* gene that is characteristic of serotypes 6A and 6B, both affected by immunity induced by PCV7 (ref. 24). The other three were COGs found in serotype 6C, as well as types 6A and 6B, and therefore did not decrease in frequency to the same extent. As the biosynthetic genes for the VT capsules 14, 19F and 23F were all classified in COGs that are also required for the production of the non-VT capsules 15B/C, 19A and 23B, respectively (with the exception of *lrp*, unique to the capsule locus of serotype 14 (ref. 25), of which only six representatives were assembled), PCV7 had little impact, even on the prevalence of the COGs involved in capsule synthesis. However, an effect could be detected by quantifying the diversity within each COG: four COGs found in both the serotype 14 and serogroup 15 biosynthetic loci showed lower diversity in 2007, as serotype 14 was eliminated (**Supplementary Fig. 24**). By contrast, three COGs shared between serogroups 15 and 23 increased in diversity as serotype 23F was replaced by 23A and 23B and serotype 15A increased in

© 2013 Nature America, Inc. All rights reserved.

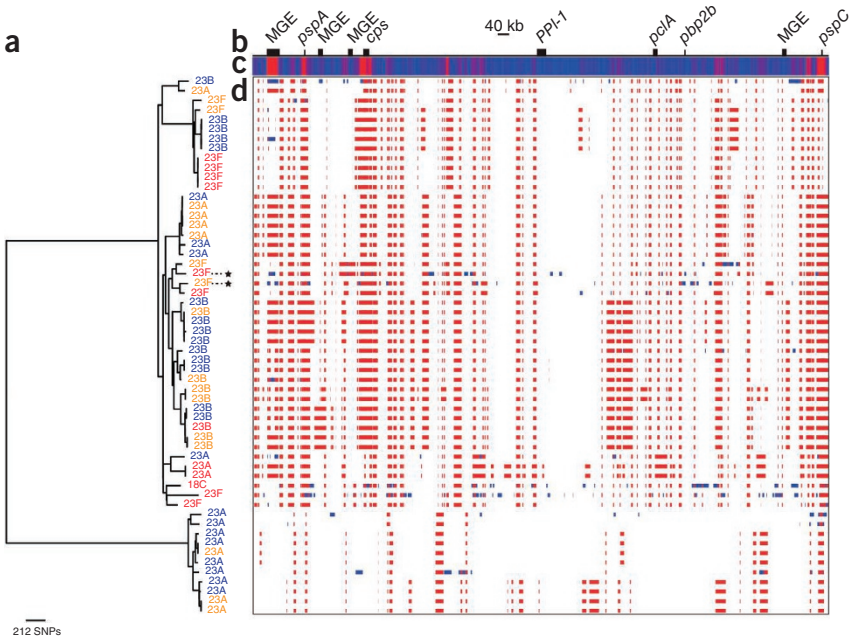

Figure 3 Serotype dynamics of SC9. (**a**) Maximum-likelihood phylogeny of SC9 based on point mutations, excluding polymorphisms introduced through the horizontal import of sequence. Taxa are labeled with their serotype: all have a serogroup 23 capsule, with the exception of a single serotype 18C isolate. Taxa are colored according to their year of isolation: red, 2001; orange, 2004; blue, 2007. Two taxa that developed high-level β-lactam resistance are marked with stars. (**b**–**d**) Putative recombination events detected across the genome alignment. (**b**) Simplified annotation of the SC9 reference genome. (**c**) Heatmap showing the density of recombination events across the genome, with blue indicating regions undergoing few or no recombination events and red indicating regions undergoing high rates of recombination. The highest levels are observed at a putative mobile genetic element (MGE), in the regions encoding the protein antigens PspA and PspC, and at the capsule biosynthesis cluster (*cps*). (**d**) Position of the putative recombination events relative to the genome annotation, including a row for each taxon in the tree. Each detected recombination event is indicated by a red block, if reconstructed as occurring on an internal branch and therefore shared by multiple taxa through common descent, or a blue block, if occurring on a terminal branch and therefore unique to a single taxon.

© 2013 Nature America, Inc. All rights reserved.

prevalence after the introduction of PCV7. Once more, the impact of the vaccine on most other COGs was small.

Alterations in antibiotic resistance

Another set of COGs persisting at an intermediate frequency were antibiotic resistance determinants. Levels of susceptibility to many antibiotics remained stable in Massachusetts after the introduction of PCV7 (refs. 7,9), despite VT serotypes being strongly associated with multidrug resistance in the United States before PCV7 (refs. 26,27) and a decline in the dispensing of antibiotics to children in Massachusetts over the sampled period[28]. For instance, tetracycline resistance is typically conferred by the *tetM* gene carried on Tn*916* family transposons[29] in *S. pneumoniae* (**Fig. 5a**). The persistence of this gene can be partially attributed to its association with SC15, which changed little in prevalence over time. Additionally, the loss of the tetracycline-resistant PMEN1 and PMEN15 lineages by 2007 was offset by the emergence of SC3, which also carries this transposon, from 2004 onwards.

Macrolide resistance was similarly stable, although the causative mechanisms shifted over time. The two resistance determinants common in pneumococci are the *ermB* rRNA methylase gene and the *mel-mef* efflux pump. The former is carried either by Tn*917* or Omega resistance cassettes[23,30], both of which are only found inserted into Tn*916*-type elements in this population. This means that the population of *ermB*-carrying isolates makes up a subset of those that are tetracycline resistant (with the exception of isolate R34-3037, in which the *tetM* gene is disrupted by a frameshift mutation); this mechanism also causes cross-resistance to streptogramin and lincosamide antibiotics. By contrast, the macrolide-specific *mel-mef* pump is carried by the Mega element that can insert into Tn*916*-type elements or be acquired elsewhere in the chromosome through transformation[31]. As a consequence of the distribution of Tn*916*-type elements, *ermB* showed a more stable association with clades, increasing in prevalence as it is carried by SC3 and the serotype 19A isolates of SC15 (**Fig. 5b** and **Supplementary Fig. 25**). The Mega cassette decreased in prevalence owing to its stable association with SC13 and SC14, which became rarer after the introduction of PCV7, and SC15 (**Fig. 5c**).

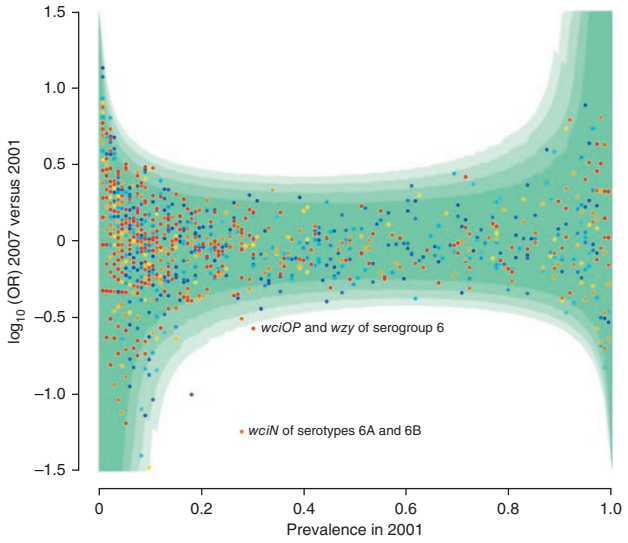

Figure 4 Alteration in COG frequency between 2001 and 2007. The logarithm of ORs indicating the proportion of isolates carrying each COG in 2007 relative to 2001 is shown relative to the prevalence of the COG in 2001. Data points are colored according to the mean length of the proteins encoded by the member orthologs, ranging from red (for longer sequences) to blue (for shorter sequences). The four contours of green shading indicate the positions of the critical values for significance at confidence intervals of 99.999%, 99.99%, 99.9% and 99%, with increasing intensity. The labeled data points lying outside the confidence intervals represent COGs involved in the synthesis of the serogroup 6 capsules; the blue data point is a false positive that seems to have been introduced through differences in the assembly of data from 2001 (75-nt reads) and 2007 (100-nt reads).

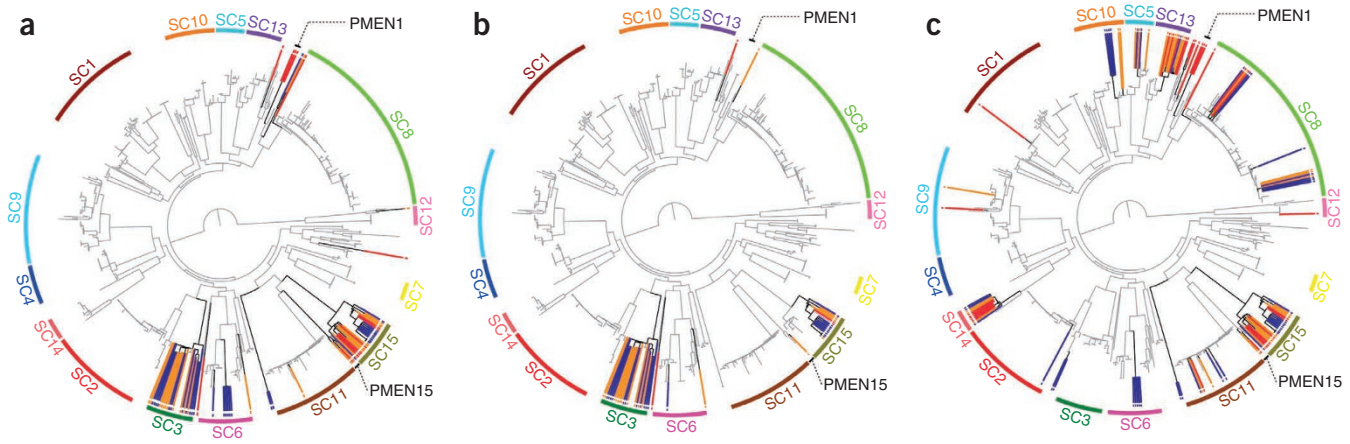

Figure 5 Distribution of antibiotic resistance genes. Maximum-likelihood phylogenies are annotated as in **Figure 1**. The branches of clades containing resistance genes are shown in black, with all other branches shown in gray. Isolates containing the resistance genes are linked to a colored point at the edge of the tree by a radiating line that is red for isolates from 2001, orange for isolates from 2004 and blue for isolates from 2007. In addition to sequence clusters, the multidrug-resistant PMEN1 and PMEN15 lineages are labeled. (**a**–**c**) Shown are the distributions of the *tetM* tetracycline resistance gene (**a**), the *ermB* macrolide resistance gene (**b**) and the *mef* gene (**c**).

© 2013 Nature America, Inc. All rights reserved.

However, it also appeared in other monophyletic sequence clusters: five acquisitions were detected within SC11, and seven were detected within SC8. Such a dispersed distribution ensures that it will be difficult to eliminate this type of resistance-conferring element using partial-coverage vaccines targeting the capsule.

The sequences determining β-lactam resistance also changed after the introduction of PCV7. In *S. pneumoniae*, β-lactam susceptibility is determined by the sequences of three genes encoding penicillin-binding proteins: *pbp2x* and *pbp1a*, found on either side of the *cps* locus, and *pbp2b*, at an unlinked position in the genome[32]. On the basis of pre-2008 definitions of resistance breakpoints[33], post-PCV7 penicillin resistance in Massachusetts was found to be stable, whereas ceftriaxone resistance decreased significantly[8,9]. Minimum inhibitory concentrations (MICs) for both drugs were correlated across the sample, with a strain's MIC for ceftriaxone typically lower than the MIC for benzylpenicillin (**Supplementary Fig. 26**). Hence, the apparent differential response to these antibiotics reflects an underlying trend in which the most resistant strains were replaced by strains with a lower level of resistance, therefore meeting the threshold MIC for penicillin but not for ceftriaxone resistance. To understand these changes, COGs corresponding to *pbp1a*, *pbp2x* and *pbp2b* were extracted from all taxa, and sequences were clustered into alleles, some of which seem to be mosaic (**Fig. 6**).

Across the sampling period, the dominant allele for each *pbp* gene was the ancestral form resulting in drug sensitivity, which altered little in frequency. However, changes in the types of resistance-conferring allele were evident, with a decrease in frequency of the forms of both *pbp2b* and *pbp2x* that were associated with isolates that were highly resistant to β-lactams (**Supplementary Fig. 25**). Allele 3 of *pbp2b*, strongly associated with the multidrug-resistant lineages PMEN1, PMEN3 and PMEN15 (ref. 15), was partially displaced by alleles 2 and 4, which rose in frequency owing to the increase in prevalence of SC11 and a shift toward resistant isolates within SC6, respectively. Allele 2 of *pbp2x*, also associated with the PMEN lineages, was replaced by allele 3, largely as a consequence of the emergence of SC3 alongside the change within SC6. Analogous analysis of the genes encoding dihydrofolate reductase and dihydropteroate synthase, certain alleles of which can cause resistance to sulfa drugs, found a decrease in the prevalence of resistance-conferring alleles (**Supplementary Fig. 27**), in line with the observed decrease in the levels of resistance[9]. This is largely the consequence of a shift in the composition of SC1 and the replacement of the PMEN lineages with SC3 and SC11, which are predominantly sensitive to sulfa drugs.

Although the major changes in β-lactam resistance are associated with shifts in the frequency of different clades of isolates, the emergence of resistant genotypes through recombination events affecting genes encoding penicillin-binding proteins was also observed. This change can be rapid, as demonstrated by the presence of transformation events affecting *pbp1a*, *pbp2x* and *pbp2b* on the same branch of the tree (for example, the isolates with stars in **Fig. 3**). One example of this type of acquisition of resistance, albeit on a long branch, is associated with the emergence of ST3280 isolates within SC6 (**Supplementary Fig. 10**), which are also distinguished from their sister taxa by the acquisition of a Tn*916*-type element carrying a Mega cassette. However, a frameshift mutation disrupts the *mel* gene, which seems to be reflected in an intermediate level of resistance to erythromycin (MIC of erythromycin between 0.25 and 0.5 μg/ml; **Supplementary Table 1**). This phenotype is also observed in an SC8 isolate with a Mega element in which the *mef* gene is disrupted by a frameshift mutation. The persistence of these distinctive genetic traits in the five closely related ST3280 isolates was of particular interest because they were all isolated from a single location in 2007, implying that there may be evidence of transmission chains within the data set.

Population substructuring

To test whether this finding represented a general phenomenon and whether transmission within locations could be detected from the genomic data, the pairwise distances (in point mutations; **Supplementary Figs. 5**–**19**) between all of the taxa within monophyletic sequence clusters were compared with the locations from which the bacteria were recovered (listed in **Supplementary Table 1**). More closely related isolates were much more likely to have been recovered from the same location than would be expected by chance (**Fig. 7**). The probability that the members of the pair would come from the same location decreased approximately exponentially with the distance in point mutations between them. The rate constant calculated from these data implied that, once two isolates were distinguished by approximately 18 point mutations, it would be equally

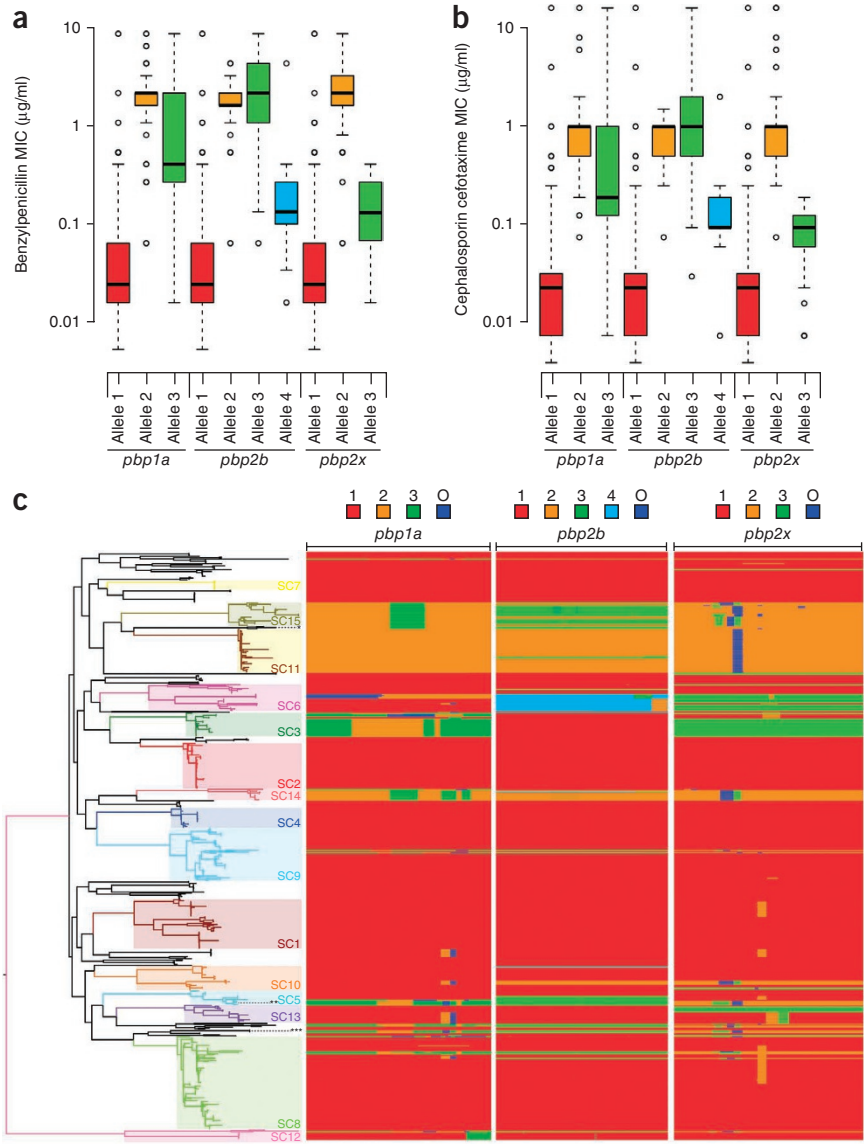

Figure 6 Changes in β-lactam resistance. (**a**,**b**) Variation in resistance. Each of the three genes encoding penicillin-binding proteins involved in resistance was independently clustered on the basis of sequence similarity using BRATnextgen; three alleles each were identified for *pbp1a* and *pbp2x*, and four alleles were identified for *pbp2b*. Box-and-whisker plots show the distribution of benzylpenicillin MICs (**a**) and cephalosporin cefotaxime MICs (**b**) associated with isolates possessing each of these alleles. (**c**) Distribution of *pbp* alleles throughout the pneumococcal population. The phylogeny in **Figure 1** is shown on the left, with the multidrug-resistant lineages PMEN1 (***), PMEN3 (**) and PMEN15 (*) indicated. The three columns on the right represent independent analyses for the three *pbp* genes involved in β-lactam resistance. These comprise one row for each taxon in the tree, with the width of the column representing the length of the gene. Blocks are colored according to the allele group to which the sequence belongs, with colors corresponding to those in **a**,**b**; changes in color indicate the recombination breakpoints characteristic of mosaic genes. The O group represents sequence that comes from an outgroup, that is, from a strain or species not represented within the collection.

likely that they would be isolated from the same or different communities in the data set. Given the mutation rate of 2–3 mutations per year, this represents around 4 years of divergence between 2 isolates. Such evidence of transmission chains indicates that the spread of bacteria within communities is significantly faster than the dissemination of bacteria between them, which occurs sufficiently slowly to be detectable via the analysis of point mutations.

Population substructuring may also result from the immune status of the host, which changes with age after successive exposures. Previous work found that strains carrying the type 1 (*rlr*) pneumococcal pilus, hypothesized to be involved in adhesion to the nasopharyngeal surface[34], were most prevalent in children under 5 years of age[35]. This was attributed to the development of an adaptive immune response to this surface structure, which precludes bacteria with this antigen from colonizing older children while still providing a functional benefit in naive hosts. The resulting niche differentiation would result in strains adapted to transmission between hosts of different ages. To identify antigens distributed in a similar pattern, logistic regression against host age (excluding individuals under 6 months of age who might be influenced by maternal immunity) was performed for COGs present at a range of intermediate frequencies.

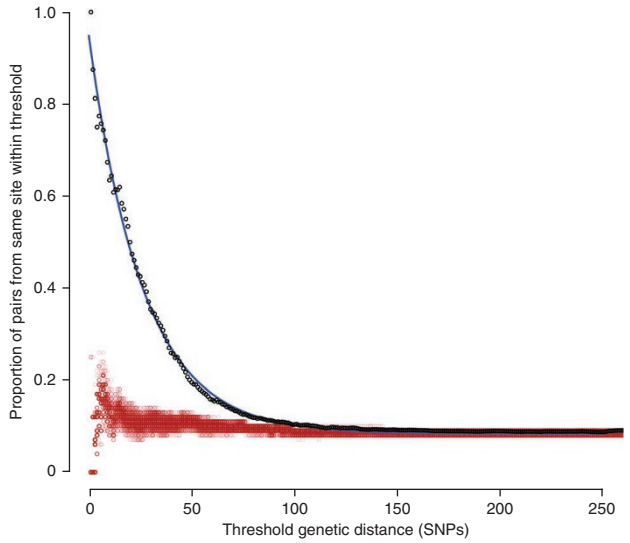

Figure 7 Geographic structure within the population. Pairwise genetic distances, with separation defined in terms of point mutations, between all strains within the same monophyletic sequence cluster were calculated and combined into a single data set. For a series of maximum genetic distance thresholds, the proportion of all pairwise comparisons meeting the condition that both isolates originate within the same location was calculated. The resulting values are plotted as the black data points, which appear to fit a curve with approximately exponential decay, indicated by the blue line. The red data points represent the outcome of 100 permutations in which the same statistic was calculated when the locations of the isolates were randomized.

© 2013 Nature America, Inc. All rights reserved.

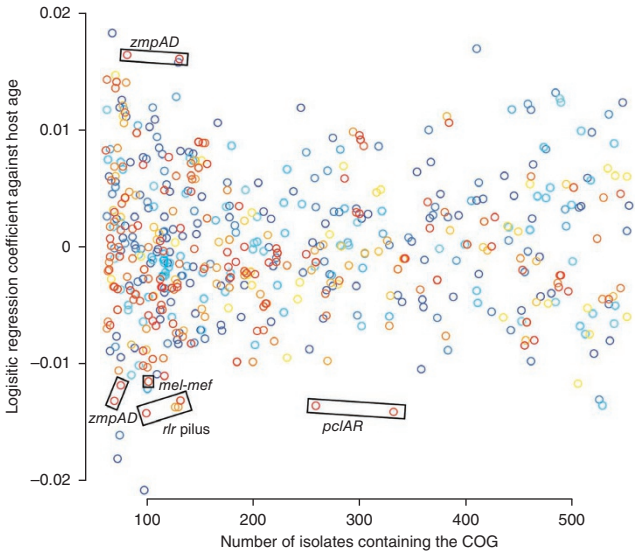

Figure 8 Effect of host age on pneumococcal genotype. A logistic regression coefficient of each COG against the host child's age (in months) was calculated. Terms relating to the year of isolation were included in the regression to account for differences in age distribution between samples, and children under 6 months of age were excluded to avoid the confounding effects of maternal immunity. The coefficient relating to host age is plotted against the prevalence of the COG within the bacterial population; only those COGs present in between 10 and 90% of the population, suggesting that they may be under balancing selection, are shown. Data points are colored according to the mean length of the proteins encoded by the member genes, ranging from red (for longer sequences) to blue (for shorter sequences). COGs described in the text are annotated.

Aside from three rare, short COGs, the functions of which are difficult to interpret, the *rlr* pilus was found to have one of the strongest relationships with host age, confirming the previous finding (**Fig. 8**). An association of a similar magnitude was shown for the *pclAR* gene cluster[36], encoding a large protein antigen that was present in about half of the sampled bacterial isolates, and for two linked alleles of the immunoglobulin A protease *zmpA* and paralogous zinc metalloprotease *zmpD*, both of which encode large surface-displayed proteins; by contrast, two alternative alleles of *zmpAD* showed association with carriage by older children. As these antigen COGs were stably associated with particular genotypes, sequence clusters showed varying patterns of association with different host ages (**Supplementary Fig. 28**). The effect of antigens frequently affected by recombination events, such as *pspA* and *pspC* (**Supplementary Figs. 5–19**), is unclear, as these genes are difficult to assemble and fall into multiple COGs owing to their extensive sequence diversity. The apparently weaker association of the *mel-mef* efflux pump with age was, in fact, most pronounced in infants younger than 6 months, in agreement with the findings in an independent sample[37] (**Supplementary Fig. 29**).

DISCUSSION

The pneumococcal population was heavily disrupted by PCV7, with almost complete loss of the seven vaccine serotypes and their rapid replacement with non-VT strains. Whole-genome sequencing provides a more detailed view of this process, identifying distant relationships between VT and non-VT pneumococci that are difficult to establish using MLST in many cases, while also indicating that capsule-switching events between closely related taxa are likely to be old relative to the introduction of PCV7. Hence, aside from the rise of lineages such as SC3 and SC7—examples of novel genotypes distantly related to those of previously observed isolates—population dynamics are largely the consequence of VT isolates being replaced by non-VT relatives that were present at low frequencies before the vaccine. The clearest example is SC9, where the serotype 23F isolates predominated before PCV7 and were replaced after the introduction of the vaccine by serotype 23A and 23B isolates. Nevertheless, the sequence cluster did not decrease in prevalence, suggesting that its average fitness has not declined, especially in the context of stable pneumococcal carriage rates. As it seems unlikely that the distribution before PCV7 resulted from genetic drift, the data indicate that this sequence cluster might occupy a specific niche that was vacated by serotype 23F isolates after the introduction of PCV7. Such a niche might be defined by cross-immunity against serogroup 23 acting as the major constraint on the prevalence of strains within this group or by the functional implications of other components of the accessory genome. Relating such considerations to the pneumococcal population structure will be important for understanding the impact of partial-coverage vaccines.

Replacement of VT strains with non-VT relatives partially explains the stable prevalence of most accessory loci. Comparison of the pooled data from 2001 and 2007 identifies a detectable vaccine effect on only a few genes associated with one of the vaccine serotypes (**Fig. 4**), demonstrating the importance of assigning genomic data to specific taxa rather than using pooled shotgun reads in studying strain dynamics. That both populations are composed of distinct strains, themselves representing different combinations of similar loci, while the IPD rate decreases indicates that subtle differences within COGs or the consequences of interaction between them are crucial in determining the rate at which pneumococci cause disease within a carrier. As the detectable alterations generally concern the *cps* locus, changes in serotype are implicated as the most important factor underlying the decrease in the incidence of IPD in Massachusetts.

The details of bacterial population restructuring after the introduction of PCV7 may well differ in locations where the initial bacterial populations or host characteristics are different from those in this study. Nor are the observed dynamics necessarily representative of the response over all timescales. Whereas levels of resistance to some antibiotics seem to have been stable in the samples of this study, on the basis of the surveillance of IPD isolates from young children across the United States, it seems that levels of resistance decreased immediately after the introduction of PCV7 before rebounding by 2004 (ref. 38). This delayed response may be a consequence of how quickly non-VT resistant pneumococcal lineages could increase in prevalence or, on the basis of the geographic structuring observed in this study, the speed at which they could spread from sources where they were present before the use of PCV7. Whether the 2007 sample represents a final equilibrium is unclear. There is little evidence of beneficial transformation events after the introduction of PCV7 in this sample, which may partially relate to the slow rate at which recombination events accumulate in some lineages. Therefore, it seems likely that the bacterial population's response to vaccine introduction will continue through adaptation via horizontal sequence exchange. On this point, it may be instructive to note that SC15, associated with the highest *r*/*m* value, recently seems to have been highly successful in the post-PCV7 environment[13]. Hence, whereas the serotype distribution may have reached equilibrium by 2007 (ref. 17), the broader response of the pneumococcal population to the conjugate vaccines is likely to continue.

© 2013 Nature America, Inc. All rights reserved.
npg

URLs. SMALT, http://www.sanger.ac.uk/resources/software/smalt/; Path-O-Gen, http://tree.bio.ed.ac.uk/software/pathogen/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. Sequence data have been deposited in the European Nucleotide Archive (ENA) under project ERP000809 as listed in **Supplementary Table 1**.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

N.J.C. is funded by an AX A postdoctoral fellow award. Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the US National Institutes of Health (NIH) under award R01AI066304 and by Wellcome Trust grant 098051. We acknowledge the support of the Sanger Institute core sequencing and informatics teams and productive discussions at PERMAFROST workshops. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US NIH.

AUTHOR CONTRIBUTIONS

M.L. and W.P.H. conceived the project. M.L., W.P.H., S.D.B., J.P. and J.A.F. supervised the project. N.J.C. and P.K.M. analyzed the data. S.I.P. and G.M.L. helped with interpretation of the data. All authors contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Weinberger, D.M. *et al.* Pneumococcal capsular polysaccharide structure predicts serotype prevalence. *PLoS Pathog.* **5**, e1000476 (2009).
2. Whitney, C.G. *et al.* Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine. *N. Engl. J. Med.* **348**, 1737–1746 (2003).
3. Steenhoff, A.P., Shah, S.S., Ratner, A.J., Patil, S.M. & McGowan, K.L. Emergence of vaccine-related pneumococcal serotypes as a cause of bacteremia. *Clin. Infect. Dis.* **42**, 907–914 (2006).
4. Hicks, L.A. *et al.* Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998–2004. *J. Infect. Dis.* **196**, 1346–1354 (2007).
5. Pai, R. *et al.* Postvaccine genetic structure of *Streptococcus pneumoniae* serotype 19A from children in the United States. *J. Infect. Dis.* **192**, 1988–1995 (2005).
6. Pelton, S.I. *et al.* Emergence of 19A as virulent and multidrug resistant Pneumococcus in Massachusetts following universal immunization of infants with pneumococcal conjugate vaccine. *Pediatr. Infect. Dis. J.* **26**, 468–472 (2007).
7. Finkelstein, J.A. *et al.* Antibiotic-resistant *Streptococcus pneumoniae* in the heptavalent pneumococcal conjugate vaccine era: predictors of carriage in a multicommunity sample. *Pediatrics* **112**, 862–869 (2003).
8. Huang, S.S. *et al.* Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004. *Pediatrics* **116**, e408–e413 (2005).
9. Huang, S.S. *et al.* Continued impact of pneumococcal conjugate vaccine on carriage in young children. *Pediatrics* **124**, e1–e11 (2009).
10. Hsu, K., Pelton, S., Karumuri, S., Heisey-Grove, D. & Klein, J. Population-based surveillance for childhood invasive pneumococcal disease in the era of conjugate vaccine. *Pediatr. Infect. Dis. J.* **24**, 17–23 (2005).
11. Yildirim, I. *et al.* Serotype specific invasive capacity and persistent reduction in invasive pneumococcal disease. *Vaccine* **29**, 283–288 (2010).
12. Yildirim, I., Stevenson, A., Hsu, K.K. & Pelton, S.I. Evolving picture of invasive pneumococcal disease in Massachusetts children: a comparison of disease in 2007–2009 with earlier periods. *Pediatr. Infect. Dis. J.* **31**, 1018–1021 (2012).
13. Hanage, W.P. *et al.* Clonal replacement among 19A *Streptococcus pneumoniae* in Massachusetts, prior to 13 valent conjugate vaccination. *Vaccine* **29**, 8877–8881 (2011).
14. Hanage, W.P. *et al.* Diversity and antibiotic resistance among nonvaccine serotypes of *Streptococcus pneumoniae* carriage isolates in the post-heptavalent conjugate vaccine era. *J. Infect. Dis.* **195**, 347–352 (2007).
15. McGee, L. *et al.* Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. *J. Clin. Microbiol.* **39**, 2565–2571 (2001).
16. Atwood, K.C., Schneider, L.K. & Ryan, F.J. Periodic selection in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **37**, 146–155 (1951).
17. Moore, M.R. *et al.* Population snapshot of emergent *Streptococcus pneumoniae* serotype 19A in the United States, 2005. *J. Infect. Dis.* **197**, 1016–1027 (2008).
18. Wyres, K.L. *et al.* Pneumococcal capsular switching: a historical perspective. *J. Infect. Dis.* **207**, 439–449 (2013).
19. Donati, C. *et al.* Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol.* **11**, R107 (2010).
20. Hanage, W.P. *et al.* Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics* **2**, 80–84 (2010).
21. Hanage, W.P. *et al.* Carried pneumococci in Massachusetts children: the contribution of clonal expansion and serotype switching. *Pediatr. Infect. Dis. J.* **30**, 302–308 (2011).
22. Hanage, W.P., Kaijalainen, T., Saukkoriipi, A., Rickcord, J.L. & Spratt, B.G. A successful, diverse disease-associated lineage of nontypeable pneumococci that has lost the capsular biosynthesis locus. *J. Clin. Microbiol.* **44**, 743–749 (2006).
23. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
24. Park, I.H. *et al.* Differential effects of pneumococcal vaccines against serotypes 6A and 6C. *J. Infect. Dis.* **198**, 1818–1822 (2008).
25. Bentley, S.D. *et al.* Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.* **2**, e31 (2006).
26. Gherardi, G., Whitney, C.G., Facklam, R.R. & Beall, B. Major related sets of antibiotic-resistant Pneumococci in the United States as determined by pulsed-field gel electrophoresis and *pbp1a-pbp2b-pbp2x-dhf* restriction profiles. *J. Infect. Dis.* **181**, 216–229 (2000).
27. Whitney, C.G. *et al.* Increasing prevalence of multidrug-resistant *Streptococcus pneumoniae* in the United States. *N. Engl. J. Med.* **343**, 1917–1924 (2000).
28. Greene, S.K. *et al.* Trends in antibiotic use in Massachusetts children, 2000–2009. *Pediatrics* **130**, 15–22 (2012).
29. Franke, A.E. & Clewell, D.B. Evidence for conjugal transfer of a *Streptococcus faecalis* transposon (Tn*916*) from a chromosomal site in the absence of plasmid DNA. *Cold Spring Harb. Symp. Quant. Biol.* **45**, 77–80 (1981).
30. Shaw, J.H. & Clewell, D.B. Complete nucleotide sequence of macrolide-lincosamide-streptogramin B-resistance transposon Tn*917* in *Streptococcus faecalis*. *J. Bacteriol.* **164**, 782–796 (1985).
31. Del Grosso, M., Camilli, R., Iannelli, F., Pozzi, G. & Pantosti, A. The *mef*(E)-carrying genetic element (mega) of *Streptococcus pneumoniae*: insertion sites and association with other genetic elements. *Antimicrob. Agents Chemother.* **50**, 3361–3366 (2006).
32. Hakenbeck, R., Tarpay, M. & Tomasz, A. Multiple changes of penicillin-binding proteins in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Antimicrob. Agents Chemother.* **17**, 364–371 (1980).
33. Cavalieri, S. *Manual of Antimicrobial Susceptibility Testing* (American Society for Microbiology, Washington, D.C., 2005).
34. Barocchi, M.A. *et al.* A pneumococcal pilus influences virulence and host inflammatory responses. *Proc. Natl. Acad. Sci. USA* **103**, 2857–2862 (2006).
35. Regev-Yochay, G. *et al.* Re-emergence of the type 1 pilus among *Streptococcus pneumoniae* isolates in Massachusetts, USA. *Vaccine* **28**, 4842–4846 (2010).
36. Paterson, G.K., Nieminen, L., Jefferies, J.M. & Mitchell, T.J. PclA, a pneumococcal collagen-like protein with selected strain distribution, contributes to adherence and invasion of host cells. *FEMS Microbiol. Lett.* **285**, 170–176 (2008).
37. Hyde, T.B. *et al.* Macrolide resistance among invasive *Streptococcus pneumoniae* isolates. *J. Am. Med. Assoc.* **286**, 1857–1862 (2001).
38. Kyaw, M.H. *et al.* Effect of introduction of the pneumococcal conjugate vaccine on drug-resistant *Streptococcus pneumoniae*. *N. Engl. J. Med.* **354**, 1455–1463 (2006).

© 2013 Nature America, Inc. All rights reserved.

ONLINE METHODS

DNA sequencing. Samples were sequenced as multiplexed libraries on the Illumina HiSeq platform to produce paired-end reads of either 75 nucleotides (for samples from 2001 or 2004) or 100 nucleotides (for samples from 2007) in length. Samples were only used where more than 60 Mb of data (equivalent to approximately 30-fold coverage of a pneumococcal genome) was available. Serotypes and sequence types were extracted as described previously[23] and compared to those recorded in earlier studies to verify the integrity of sample handling.

Genome assembly and annotation. Genomes were assembled *de novo* using Velvet[39] through optimizing the *k*-mer and expected coverage values as described previously[23]. Short contigs less than twice the length of the *k*-mer value used in the final assembly were discarded. Both Glimmer3 (ref. 40) and Prodigal[41] were trained on the complete reference sequence of *S. pneumoniae* ATCC 700669 (ref. 42), and these models were applied to each draft assembly modified through the addition of a three-frame stop codon sequence to each end of the genome. A consensus of these two methods was then derived through only using putative protein-coding sequences where the central halves of the Glimmer3 and Prodigal gene models overlapped on the same strand of the assembly. Finally, protein-coding sequences were trimmed wherever they overlapped breaks in the assembly, such that they did not span multiple contigs.

Clustering and analysis of orthologous gene products. All putative protein-coding sequences were translated, and an all-against-all comparison was performed using BLAT[43] with default settings. An initial clustering based on these alignments was then produced using COGtriangles[44]. COGnitor[45] was then used to extract a unique best-hit cluster for each protein-coding sequence. As COGtriangles requires the presence of at least three orthologs in the data set to produce a COG, it left simple sequences, singletons and pairs of orthologs unclustered. To resolve clustering for these elements, a BLAT *E*-value threshold corresponding to a *P* value of 0.05 that was Bonferroni corrected for the number of reported BLAT comparisons was established. Unclustered gene products with reciprocal best-BLAT hits surpassing this threshold were paired into COGs with two members. Remaining gene products were either deemed insufficiently complex to cluster, if their self-BLAT *E* value was greater than the threshold, or assigned as singleton COGs.

Each COG was characterized through alignment of the protein sequences with MUSCLE[46] and was reverse translated into a codon alignment using the DNA sequences. On the basis of this alignment, diversity was quantified as the mean pairwise Kimura distance between members. The codon alignments of the COGs associated with β-lactam and sulfa drug resistance were analyzed using BRATnextgen[47] with a window size of 100 bp. Recombinant segments were identified using a *P* value of 0.05, as calculated through 100 replicates of 10 iterations. Where MICs were specified as being less than a given value, the value was halved for the purposes of associating alleles with different levels of resistance; correspondingly, where MICs were described as being greater than a specified value, the value was doubled.

Analysis of population structure. The codon alignments of each core COG (those with a single representative in each genome assembly) were concatenated to give a single core alignment. A maximum-likelihood phylogeny was then generated with RAxML[48] using a general time-reversible nucleotide substitution model with four gamma categories for rate heterogeneity. This alignment was also analyzed using BAPS[49]. Three runs, starting from a maximum cluster number of 20, 40 or 60, each converged on the 15 monophyletic sequence clusters described, with the exception that 1 run added serotype 3 isolates to SC7.

Analysis of individual lineages. Fifteen BAPS clusters were monophyletic within the maximum-likelihood tree and therefore seemed to constitute groups of isolates where not all phylogenetic information had been abrogated by recombination (that is, a clonal frame remained). One isolate within each of these groups was selected for reassembly with both SGA[50] and Velvet[40]; the resulting contigs were merged using Zorro[51] and then arranged into scaffolds using SSPACE[52]. Reference assemblies were manually curated using ACT[53]. Illumina read pairs from all members of the clade were then mapped against this reference using SMALT (see URLs), the output of which was processed as described previously[54,55] to produce a whole-genome alignment. This was analyzed using an algorithm that iteratively generates a maximum-likelihood phylogeny on the basis of point mutations while identifying recombination events[23]. When comparing the level of recombination observed in each sequence cluster, branches on which a total of 5,000 substitutions or more occurred were excluded from the data set, as the accuracy of recombination detection was liable to decrease for these branches. The difference in r/m values was still significant, as calculated by a Kruskal-Wallis test, when all branches were included, as well as at all length cutoff values between 1,000 and 10,000 SNPs at 500-SNP intervals. Recombination events occurring in regions annotated as MGEs in **Supplementary Figures 5–19** were also excluded, as these may represent the transfer of autonomously mobile elements rather than homologous recombination events. Exponential distributions were fitted to the lengths of homologous recombinations as described previously[55].

Identification of spatiotemporal signals. Path-O-Gen (see URLs) was used to examine these lineage phylogenies for signs of a temporal signal. Where there was a significant positive correlation between the dates of isolation and root-to-tip distance, the alignment of polymorphisms caused by point mutations was analyzed using BEAST[56]. Tree topology was fixed to maintain consistency with the prediction of recombination events, while a general time-reversible substitution model was allowed to occur with a relaxed lognormal mutation rate[57]. A skyline plot was used as the population size prior[58].

When analyzing the geographic distribution of isolates, pairwise distances were extracted from the phylogenies shown in **Supplementary Figures 5–19** using Bioperl[59]. When plotted as shown in **Figure 7**, an exponential relationship was fitted of the form

$$y = Ae^{-Bx} + C$$

where y represents the proportion of pairs of isolates originating from the same location and x is the threshold maximum genetic distance, in terms of point mutations, between a pair. C represents the probability of two isolates originating from the same location by chance (estimated as 0.083, 95% confidence interval of 0.083–0.084). The sum of A and C represents the probability that two identical isolates come from the same location; A was estimated as 0.86 (95% confidence interval of 0.86–0.87). B is the rate at which pairs become discordant for location per point mutation; this was estimated as 0.038 per mutation (95% confidence interval of 0.038–0.039).

Logistic regressions. Logistic regressions were performed using R[60]. When identifying COGs associated with different host ages, the binary-encoded presence-absence pattern for each COG was regressed against the age of the host child in months and the year of isolation as categories. These latter confounding variables were included to guard against different population structures based on age for children sampled in different years. Children under 6 months of age were excluded from the analysis to avoid confounding effects from maternal immunity.

39. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
40. Delcher, A.L., Bratke, K.A., Powers, E.C. & Salzberg, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673–679 (2007).
41. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
42. Croucher, N.J. *et al.* Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*^{Spain23F} ST81. *J. Bacteriol.* **191**, 1480–1489 (2009).
43. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
44. Kristensen, D.M. *et al.* A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26**, 1481–1487 (2010).
45. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
46. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

© 2013 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2625

47. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).
48. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood–based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463 (2005).
49. Tang, J., Hanage, W.P., Fraser, C. & Corander, J. Identifying currents in the gene pool for bacterial populations using an integrative approach. *PLoS Comput. Biol.* **5**, e1000455 (2009).
50. Simpson, J.T. & Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* **26**, i367–i373 (2010).
51. Argueso, J.L. *et al.* Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome Res.* **19**, 2258–2270 (2009).
52. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
53. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672–2676 (2008).
54. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
55. Croucher, N.J., Harris, S.R., Barquist, L., Parkhill, J. & Bentley, S.D. A high-resolution view of genome-wide pneumococcal transformation. *PLoS Pathog.* **8**, e1002745 (2012).
56. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
57. Drummond, A.J., Ho, S.Y., Phillips, M.J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
58. Drummond, A.J., Rambaut, A., Shapiro, B. & Pybus, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
59. Stajich, J.E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).
60. R Core Development Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2011).

© 2013 Nature America, Inc. All rights reserved.