Dense genomic sampling identifies highways of pneumococcal recombination

Claire Chewapreecha1, Simon R Harris1, Nicholas J Croucher2, Claudia Turner3–5, Pekka Marttinen6, Lu Cheng7, Alberto Pessia7, David M Aanensen2, Alison E Mather1, Andrew J Page1, Susannah J Salter1, David Harris1, Francois Nosten3,5, David Goldblatt8, Jukka Corander7,9,10, Julian Parkhill1, Paul Turner3–5,11 & Stephen D Bentley1,10,11

Evasion of clinical interventions by Streptococcus pneumoniae occurs through selection of non-susceptible genomic variants. We report whole-genome sequencing of 3,085 pneumococcal carriage isolates from a 2.4-km2 refugee camp. This sequencing provides unprecedented resolution of the process of recombination and its impact on population evolution. Genomic recombination hotspots show remarkable consistency between lineages, indicating common selective pressures acting at certain loci, particularly those associated with antibiotic resistance. Temporal changes in antibiotic consumption are reflected in changes in recombination trends, demonstrating rapid spread of resistance when selective pressure is high. The highest frequencies of receipt and donation of recombined DNA fragments were observed in non-encapsulated lineages, implying that this largely overlooked pneumococcal group, which is beyond the reach of current vaccines, may have a major role in genetic exchange and the adaptation of the species as a whole. These findings advance understanding of pneumococcal population dynamics and provide information for the design of future intervention strategies.

S. pneumoniae is a human-restricted nasopharyngeal commensal and respiratory pathogen with a major impact on global health, particularly that of infants in resource-poor settings1. Pneumococcal carriage is regarded as a risk factor for the development of invasive and non-invasive diseases, and carriage rates tend to be higher in poorer countries2. Treatment and prevention of pneumococcal diseases are becoming more challenging owing to the acquisition of antibiotic resistance and vaccine-escape genotypes3–5. The frequent recombination observed in this species allows for rapid acquisition of genetic material with the potential for selective advantage6. Previous studies have noted the relative importance of recombination over single-nucleotide substitution in the rapid evolution of a multidrug-resistant lineage4 and showed a correlation between recombination admixture and the acquisition of antibiotic resistance7. Moreover, recombination has been captured during the course of a single chronic, polyclonal infection8, demonstrating the interaction between lineages9,10 and their coevolution within the nasopharynx of humans. Although previous data have been informative, they are restricted by the limited amount of genotypic information7,11 or relatively sparse samplings4,5. Here we apply very dense sampling and full genomic analysis to analyze pneumococcal evolution and genetic interactions between lineages with unprecedented resolution.

A longitudinal pneumococcal carriage study was conducted in Maela, a rural refugee community located close to the border between Thailand and Myanmar where anti-pneumococcal vaccination has not yet been implemented. The study was performed between 2007 and 2010 on a cohort of almost 1,000 infants and approximately one-quarter of their mothers over a period of 3 years12,13. Whole-genome sequencing was performed on 3,085 isolates randomly selected from this cohort to include about 100 isolates from each consecutive month of the studied period.

After sequencing and identification of SNPs (Online Methods), we determined the structure of the Maela pneumococcal population by clustering genomes on the basis of sequence similarity using BAPS (Bayesian Analysis of Population Structure)14,15. The tool was applied in a hierarchical fashion to generate 33 primary clusters (BCs) subdivided into 183 secondary clusters (sBCs) (Supplementary Table 1). Secondary clusters were mostly clonal, often belonging to the same multi-locus sequence typing (MLST) clonal complex. Non-typeable (NT) pneumococci, which lack genes for capsule biosynthesis, were the most prevalent capsule phenotype group (Fig. 1) and were distributed across multiple...
lineages, including one of the largest BAPS clusters (BC3-NT). Out of 512 isolates classified as NT, 42 appeared to have a deletion of the capsule biosynthesis locus (cps), whereas the remainder harbored typical NT genes, as described in Supplementary Table 2 (ref. 16). We detected 191 plausible capsule-switching events across the population, including 19 switches involving NT status—9 events from encapsulated states to NT, 9 events from NT to encapsulated states and 1 event with an ambiguous direction—indicating that conversion between the encapsulated and non-encapsulated states is not uncommon and may be an important factor in population dynamics.

The seven largest primary BAPS clusters (denoted BC1–BC7) were studied further to evaluate evolutionary parameters such as nucleotide substitution and recombination. Each cluster contained more than 100 samples and totaled 1,216 genomes representing 39.4% of the total data set and several common serotypes, BC1-19F, BC2-23F, BC3-NT, BC4-6B, BC5-23A/F, BC6-15B/C and BC7-14 (Fig. 1a). Sequence reads for each cluster were remapped against a closely related reference genome to allow greater sensitivity for detection of variants. Mean estimated substitution rates fell within the range of 1.45–4.81 × 10⁻⁶ substitutions per site per year with overlapping 95% credibility intervals (Fig. 2a and Supplementary Table 3), demonstrating no significant difference in rates between the seven lineages (Kruskal-Wallis test P value = 0.98). It is possible that less prevalent lineages have significantly different substitution rates, but these rates could not be reliably assessed with the current density of sampling.

We calculated the rate of recombination in each cluster as the ratio of the number of homologous recombination events to the number of mutations (r/m); this avoids potential bias from using the number of recombinational polymorphisms, which can be affected by varying genetic distances between the donor and recipient (Supplementary Table 4). The ratio was less than 1 in all studied clusters, indicating that recombination events occurred less frequently than single-nucleotide substitution. Unlike substitution rate, the r/m ratio was significantly different between clusters (Fig. 2b), highlighting different rates of recombination (Kruskal-Wallis test P value = 1.24 × 10⁻⁸). This observation is consistent with Croucher et al.⁵ as well as with other in vitro studies¹⁷, suggesting the potential for a different speed of response to environmental changes within the population. The highest recombination rate was observed in a group dominated by NT isolates (BC3-NT) (Mann-Whitney test P value = 1.76 × 10⁻⁵, ANCOVA test P value = 0.0011), consistent with the idea that the capsule acts as a physical barrier to DNA uptake in natura. Cluster BC3-NT comprised both NT and serotype 14 isolates of the same genetic background, providing an opportunity to test the influence of the capsule on the rate of recombination. The recombination rate was significantly higher in the NT isolates compared to those of serotype 14 in this cluster (Mann-Whitney test P value = 2.44 × 10⁻³), providing further evidence that encapsulation reduces recombination efficiency.

Although we found the amount of recombination to be different between clusters, the genomic loci where we identified recombination events were nonrandom and were remarkably consistent between clusters. These hotspots were defined as sites with recombination frequency above the 95th percentile of the recombination frequency detected for the cluster as a whole (Fig. 3), thus accounting for recombination frequency, population size and diversity. Notably, the highly recombinogenic cluster BC3-NT contained hotspots with the highest frequencies, despite being third in terms of population size. This consistency in the pattern of hotspots across the species indicates that there are a limited number of genes where diversity accumulates as a consequence of recombination occurring at a heightened rate relative to the rest of the genome. It is likely that host immunity and clinical practices are the selective pressures underlying this pattern.

Indeed, the six most prominent hotspots in the Maela pneumococcal population were focused on genes encoding cell surface antigens (pspA and pspC) and genes associated with resistance to antibiotics (pbp1a, pbp2b, pbp2x and folA).
Having found that recombination hotspots coincide with genes associated with antibiotic resistance, we sought to clarify further the relationships between recombination, antibiotic susceptibility and antibiotic consumption in the community.

β-lactam resistances are known to be conferred by allelic forms of the penicillin-binding proteins encoded by pbp1a, pbp2x and pbp2b. The phylogeny of these 3 concatenated genes based on the whole population \((n = 3,085)\) (Fig. 4a) highlights 2 features: (i) admixture of alleles throughout the population, with those of BC3-NT being the most widely distributed, and (ii) an association between β-lactam resistance and long branches, which can be an indication of recombination. The same trend was observed for individual pbp genes on both older (internal) branches and recent (external) branches were associated with resistant phenotypes, implying a continuous selective pressure for β-lactam resistance.

Allelic forms of dihydrofolate reductase (folA) and dihydropteroate synthase (folP) are known to confer resistance to cotrimoxazole. Similar to the phylogenies for genes involved in β-lactam resistance, a phylogeny of folA showed that this gene had been shuffled within the population, with signals of recombination manifested as alternating short and long branches (Fig. 4b). However, we did not find a similar pattern for folP. Furthermore, there was no overall association between strains undergoing recombination at fol genes and cotrimoxazole resistance. This lack of association might be due to the acquisition of resistance through base substitution or could be a distorted signal due to changes in selective pressure over time. Cotrimoxazole was recommended as a primary treatment for non-severe pneumonia in Maela from 1994 until 2002. However, owing to increasing resistance across the region, its use has since been in decline, contrasting with the increasing trend in β-lactam use. Reflecting this trend in antibiotic usage, we observed a temporal change in the recombination trend for cotrimoxazole resistance. Isolates that had undergone recent recombination (detected on external branches) were phenotypically

**Figure 2** Evolutionary parameters estimated in dominant clusters.
(a) Mutation rates estimated using BEAST\(^\text{**2}2\). Error bars correspond to 95% credibility intervals. The dashed line represents the mutation rate estimated in a previous pneumococcal study of \(1.57 \times 10^{-6}\) substitutions per site per year\(^4\) (95% confidence interval of 1.34–1.79 \(\times 10^{-6}\)).

(b) Recombination events per mutation (r/m) across the investigated clusters quantified by two separate methods: linear regression on each branch of the appropriate phylogeny and the arithmetic mean of r/m on each branch. Error bars represent 95% confidence intervals. BC3-NT (in blue) has the highest r/m ratio, with its subclusters NT and serotype 14 highlighted in red and green, respectively. Please note that the assumptions of the linear regression models were not met for serotype 14.

Having found that recombination hotspots coincide with genes associated with antibiotic resistance, we sought to clarify further the relationships between recombination, antibiotic susceptibility and antibiotic consumption in the community.

β-lactam resistances are known to be conferred by allelic forms of the penicillin-binding proteins encoded by pbp1a, pbp2x and pbp2b. The phylogeny of these 3 concatenated genes based on the whole population \((n = 3,085)\) (Fig. 4a) highlights 2 features: (i) admixture of alleles throughout the population, with those of BC3-NT being the most widely distributed, and (ii) an association between β-lactam resistance and long branches, which can be an indication of recombination. The same trend was observed for individual pbp genes on both older (internal) branches and recent (external) branches were associated with resistant phenotypes, implying a continuous selective pressure for β-lactam resistance.

Allelic forms of dihydrofolate reductase (folA) and dihydropteroate synthase (folP) are known to confer resistance to cotrimoxazole. Similar to the phylogenies for genes involved in β-lactam resistance, a phylogeny of folA showed that this gene had been shuffled within the population, with signals of recombination manifested as alternating short and long branches (Fig. 4b). However, we did not find a similar pattern for folP. Furthermore, there was no overall association between strains undergoing recombination at fol genes and cotrimoxazole resistance. This lack of association might be due to the acquisition of resistance through base substitution or could be a distorted signal due to changes in selective pressure over time. Cotrimoxazole was recommended as a primary treatment for non-severe pneumonia in Maela from 1994 until 2002. However, owing to increasing resistance across the region, its use has since been in decline, contrasting with the increasing trend in β-lactam use. Reflecting this trend in antibiotic usage, we observed a temporal change in the recombination trend for cotrimoxazole resistance. Isolates that had undergone recent recombination (detected on external branches) were phenotypically

**Figure 3** Recombination hotspots in seven prevalent clusters. Panels (top to bottom) are ordered by decreasing cluster population size. For each cluster, recombination hotspots were identified as sites with recombination frequency above the 95th percentile of the homologous recombination frequency detected in that cluster. The 95th percentile levels are represented as horizontal dashed lines. Recombination hotspots detected in at least four of the seven studied clusters are shaded in different colors. These common hotspots (in order by genomic coordinates) encode pneumococcal surface protein A (pspA, purple), penicillin-binding protein 2x (pbp2x, blue), penicillin-binding protein 1a (pbp1a, green), dihydrofolate reductase (folA, orange), penicillin-binding protein 2b (pbp2b, red) and pneumococcal surface protein C (pspC, gray). The figure includes 2,077 recombination events; the 132 events associated with mobile genetic elements are not displayed.
less resistant than those with older recombination events (internal branches) (Fisher’s exact test \( P \) value = \( 4.21 \times 10^{-5} \); Table 1), which is consistent with a reduction in selection pressure. Having shown two cases where recombination allowed the bacteria to respond to fluctuating selection pressure, we next asked which genotypes were the major source of recombinant DNA that might allow such adaptation.

The highly saturated sampling frame allowed detection of the sources of recombinant DNA fragments (referred to as ‘donor blocks’) by searching the entire assembled genome data set for identity with the detected recombined regions (referred to as ‘recipient blocks’) using strict detection criteria to reduce the number of false positives (Supplementary Note). Only recipient blocks uniquely detected in single isolates (at the tips of the phylogenetic tree) of BC1–BC7 were considered, thus restricting analysis to recent events where donor detection was less likely to be confounded by subsequent recombination or substitution. Of the 928 unique recipient blocks identified in BC1–BC7, 443 were found to have identical matches (donor blocks) elsewhere in the data set (Fig. 1a). When classified by sBCs, potential donor blocks were identified in single or multiple genetic backgrounds. For example, isolate SMRU1452 had nine recipient blocks with identical hits detected in eight different clusters, each with a different serotype (Supplementary Fig. 2 and Supplementary Table 6). Eight of the nine recipient blocks were detected in one of eight clusters (sBC145–sBC152, serotype 6B), and the remaining fragment was detected in a second cluster (sBC140–sBC143, serotype NT). Taken together, these observations suggest that the recent ancestor of isolate SMRU1452 had recombined with members of sBC145–sBC152 (serotype 6B) and sBC140–sBC143 (serotype NT), resulting in the import of eight and one DNA regions of diversity, respectively.

Mapping all such interactions across the wider population allowed us to ask whether there was heterogeneity in the donation frequency between each cluster within the population. When considering

![Image](https://fakeimage.com/figure4.png)

**Figure 4** Associations between recombining genes and resistant phenotypes. (a) \( \beta \)-lactam resistance. The center shows a SNP-based phylogeny for the concatenated \( pbp1a, pbp2b \) and \( pbp2x \) genes from 3,085 strains rooted on *Streptococcus mitis*. The inner ring is colored according to membership in the seven dominant population clusters (BC1–BC7). The outer ring is colored on the basis of \( \beta \)-lactam resistance phenotypes (white, susceptible; black, non-susceptible). (b) Cotrimoxazole resistance. Center plots show SNP-based phylogenies of \( folA \) and \( folP \). The color scheme for the inner ring is the same as for the concatenated \( pbp \) genes, whereas the outer ring is colored according to cotrimoxazole resistance phenotype (white, sensitive; gray, intermediate; black, resistant).

| Observed phenotypes | No recombination at loci of interest | Recombination at loci of interest | Recent recombination (external node) at loci of interest | Older recombination (internal node) at loci of interest |
|---------------------|-----------------------------------|-------------------------------|----------------------------------|---------------------|
| \( \beta \)-lactam resistance: resistant/sensitive (ratio)  | 120/146\((0.82)\) | 795/150\((5.30)\) | 25/6\((4.17)\) | 770/144\((5.35)\) |
| Cotrimoxazole resistance: resistant/sensitive + intermediate (ratio) | 210/28\((7.50)\) | 873/100\((8.73)\) | 10\((9.0)\) | 863/91\((9.48)\) |

\(^{a}\)Significant difference between \( \beta \)-lactam resistance phenotypes observed in strains with recombination at \( pbp \) genes and those without recombination (\( P \) value < \( 2.2 \times 10^{-16} \)). \(^{b}\)Significant difference in cotrimoxazole resistance phenotypes between recent recombination and older recombination at \( fol \) genes (\( P \) value = \( 3.49 \times 10^{-5} \)). Note that the difference is still significant when ratios are grouped by resistant + intermediate/sensitive phenotype (\( P \) value = \( 0.00931 \)).
primary BAPS clusters, there was a higher probability of NT isolates being the donor than the rest of the population (Mann-Whitney U test between NT isolates and other clusters, P value < 2.2 × 10^{-16}; Supplementary Fig. 3a). However, this finding should be interpreted with caution, as NT strains are known to be efficient recombination recipients, meaning that the results here may be confounded by NT strains being co-recipients as well as recombination donors. This hypothesis is consistent with the positive correlations observed between cluster population size and, separately, cluster diversity with the probability of becoming a potential donor (Spearman’s rank correlation ρ = 0.592 and 0.773, P value = 2.69 × 10^{-4} and 1.45 × 10^{-6}; Supplementary Fig. 3b,c). The clusters of NT isolates demonstrate both these characteristics, large population size and high diversity and, correspondingly, were identified as having a large number of potential donors.

In summary, lineage-specific rates of recombination, both for donation and receipt of DNA, suggest a structure to the genetic flux within this population in which specific lineages function as hubs of gene flow. The most frequently exchanged genes are those associated with antibiotic resistance and immune interaction, with trends in the former sensitive to the level of antibiotic consumption. Together, these data imply that within the bacterial population there are differential rates of response to environmental selection pressure between lineages and, further, that geographically or temporally distinct pneumococcal populations have different capacities for adaptation. The high rates of receipt and donation for recombiant DNA observed in NT pneumococci make them a potential major reservoir of genetic diversity for the wider population. Although increased recombination rate could bring transient benefit, there are potential long-term disadvantages due to increasing genomic instability. It is therefore notable that we observe sporadic switching between NT and encapsulated states, which may serve as a mechanism to modulate the tradeoff between benefit and cost for recombination rates. As NT pneumococci are rarely associated with disease, they are excluded from the targets of currently available polysaccharide vaccines and will therefore act as a reservoir for antibiotic resistance determinants.

**URLs.** Assemblies generated in this study can be accessed at ftp://ftp.sanger.ac.uk/pub/pathogens/Streptococcus_pneumoniae/Maela_assemblies.tgz. Software can be found at the following web pages: 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 codes.** Sequencing reads for all 3,085 isolates have been deposited in the NCBI Sequencing Read Archive (SRA) under study accessions ERP000435, ERP000483, ERP000485, ERP000487, ERP000598 and ERP000599. A full list is provided in Supplementary Table 1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank the microbiology and clinical team in the Shoklo Malaria Research Unit, part of the Mahidol Oxford University Research Unit, Faculty of Tropical Medicine, Mahidol University, Thailand, and the core informatics, library-generating and sequencing teams at the Wellcome Trust Sanger Institute. Attending authors are grateful for the opportunity for discussion at the Permafrost workshop. C.C. was funded by a Royal Thai Government scholarship and a Wellcome Trust PhD studentship. J.C., P.M., A.P. and L.C. were funded by Academy of Finland grant 251770 and European Research Council grant 239784. S.D.B. is partly funded by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. P.T. was funded by Wellcome Trust grant 083735/Z/07Z. This work is sponsored by Wellcome Trust grant 098051.

**AUTHOR CONTRIBUTIONS**

S.D.B., P.T., J.P., D.G. and F.N. conceived the study. P.T. and C.T. collected and provided the samples for the study. J.P., S.D.B., C.C., S.R.H., N.J.C. and J.C. designed the analyses. C.C., S.R.H., P.M., L.C., A.P., D.M.A., A.E.M., A.J.P., S.J.S., D.H. and J.C. performed the analyses. C.C., S.D.B., S.R.H., A.E.M. and N.J.C. wrote the manuscript. All authors read and approved the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

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

1. O’Brien, K.L. & Nohynek, H. Report from a WHO Working Group: standard method for detecting upper respiratory carriage of Streptococcus pneumoniae. Pediatr. Infect. Dis. J. 22, e1–e11 (2003).
2. Adelfa, I.M. et al. Pre-vaccination nasopharyngeal pneumococcal carriage in a Nigerian population: epidemiology and population biology. PLoS ONE 7, e30548 (2012).
3. Hanage, W.P. et al. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. Epidemics 2, 80–84 (2010).
4. Croucher, N.J. et al. Rapid pneumococcal evolution in response to clinical interventions. Science 331, 430–434 (2011).
5. Croucher, N.J. et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat. Genet. 45, 656–663 (2013).
6. Steinmoen, H., Knutsen, E. & Havarstein, L.S. Induction of natural competence in Streptococcus pneumoniae triggers lysis and DNA release from a subtraction of the cell population. Proc. Natl. Acad. Sci. USA 99, 7681–7686 (2002).
7. Hanage, W.P., Fraser, C., Tang, J., Connor, T.R. & Corander, J. Hyper-recombination, diversity, and antibiotic resistance in pneumococci. Science 324, 1454–1457 (2009).
8. Hiller, N.L. et al. Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog. 6, e1001108 (2010).
9. Eldholm, V., Johnsrud, O., Haugen, K., Ohnstad, H.S. & Havarstein, L.S. Fratricide in Streptococcus pneumoniae: contributions and role of the cell wall hydrolases CbpD, LyoA and LytC. Microbiology 155, 2223–2234 (2009).
10. Wei, H. & Havarstein, L.S. Fratricide is essential for efficient gene transfer between pneumococci in biofilms. Appl. Environ. Microbiol. 78, 5897–5905 (2012).
11. Donkor, E.S. et al. High levels of recombination among Streptococcus pneumoniae isolates from the Gambia. mBio 2, e00040–11 (2011).
12. Turner, P. et al. A longitudinal study of Streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. PLoS ONE 7, e38271 (2012).
13. Turner, C. et al. High rates of pneumonia in children under two years of age in a South East Asian refugee population. PLoS ONE 8, e54026 (2013).
14. Cheng, L., Connor, T.R., Siren, J., Aarnes, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS Software. Mol. Biol. Evol. 30, 1224–1228 (2013).
15. Corander, J., Marttinen, P., Siren, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9, 539 (2008).
16. Salter, S.J. et al. Variation at the capsule locus, cps, of mistyped and non-typable Streptococcus pneumoniae isolates. Microbiology 158, 1560–1569 (2012).
17. Hsieh, Y.C. et al. Serotype competence and penicillin resistance in Streptococcus pneumoniae. Emerg. Infect. Dis. 12, 1709–1714 (2006).
18. Zapun, A., Contreras-Martel, C. & Vernet, T. Penicillin-binding proteins and β-lactam resistance. FEMS Microbiol. Rev. 32, 361–385 (2008).
19. Adrian, P.V. & Klugman, K.P. Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of Streptococcus pneumoniae. Antimicrob. Agents Chemother. 41, 2406–2413 (1997).
20. Padayachee, T. & Klugman, K.P. Novel expansions of the gene encoding dihydropteroate synthase in trimethoprim-sulfamethoxazole-resistant Streptococcus pneumoniae. Antimicrob. Agents Chemother. 43, 2225–2230 (1999).
21. Silver, L.L. Multi-targeting by monotherapeutic antibacterials. Nat. Rev. Drug Discov. 6, 41–55 (2007).
22. Hage, C.W., Gambel, J.M., Sirjan, A., Pitarangsi, C. & Echeverria, P. Trends in antibiotic resistance among diarrheal pathogens isolated in Thailand over 15 years. Clin. Infect. Dis. 26, 341–345 (1998).
ONLINE METHODS

Sample collection. Nasopharyngeal swabs in this study were collected between November 2007 and November 2010 from 528 infants over the first 24 months of life, as well as from 242 mothers. The swabs, also part of the study described in refs. 12,13, were collected and processed according to World Health Organization (WHO) pneumococcal carriage detection protocols23. All isolates were serotyped and then tested for antibiotic susceptibilities (Supplementary Note). A smaller collection of 3,085 single-colony isolates was randomly selected for whole-genome sequencing in such a way that about 100 isolates were recovered from each consecutive month of the study period, except the start and end of the sampling period when the number of isolates were lower. DNA extraction for each isolate was performed using the RBC Bioscience MagCore HF16 platform. Our collection is tabulated in Supplementary Table 1.

Whole-genome sequencing. All processing and sequencing of genomic DNA was performed by the core sequencing teams at the Wellcome Trust Sanger Institute. All samples were sequenced as multiplexed libraries using Illumina HiSeq 2000 analyzers on 75-bp paired-end runs as described in ref. 4, giving mean coverage of 276.67 reads per nucleotide. MLST and serotype, for both capsulated and non-capsulated isolates, were determined directly from the sequence and compared to previously experimentally determined typing information to verify sample identities (Supplementary Table 2 and Supplementary Note).

Sequence assembly. Genomes for 3,085 strains were de novo assembled using an iterative process involving Velvet24, SSPACE25, GapFiller26, BWA27 and Bowtie28 (Supplementary Note). A final quality control step was performed on each assembly, with the reads mapped back to the assembly using SMALT 0.5.7. The assembly pipeline gave on average a total length of 2,161,240 bp from 111–279 contigs with average contig length of 33,191 bp and average N50 of 65,656 bp.

Mapping. To estimate the whole population structure, reads from all 3,085 samples were mapped onto a single reference genome, S. pneumoniae ATCC700669 (European Molecular Biology Laboratory (EMBL) accession FM211187)29, using SMALT 0.5.7 to generate a coarse but sufficient alignment for determining the population structure. The 2,221,315-bp reference gave on average 82.33% mapping coverage. Bases were called using the method described in ref. 30. To improve resolution for a deeper analysis, closely related references were employed for finer mapping in seven dominant clusters. Public references as well as draft references generated from these data (Supplementary Table 7 and Supplementary Note) were used for mapping as described previously. The final alignment for deeper analysis included indels using the method described in ref. 30.

Estimating population structure by clustering and approximate phylogeny. On the basis of coarse mapping against the core genome of S. pneumoniae ATCC 700669, BAPS software v6.0 (refs. 15,31–33) was used to estimate the population structure (for application7,26–28). As described in refs. 14,37, we used BAPS in a hierarchical manner to resolve the population structure at a fine level of detail (Supplementary Note).

Also, an approximately maximum-likelihood phylogenetic tree was estimated by FastTree30 using a GTR + CAT (General Time Reversible with per-site rate CATegories) model of approximation for site rate variation. With 1,000 resamples, 80.6 and 32.6% of the branches had over 0.700 and 1.000 bootstrap supports, respectively.

Estimating evolutionary parameters. Rates of single-nucleotide substitution as well as recombination were determined for each of the seven dominant clusters (Supplementary Fig. 4a–g and Supplementary Note). Recombination SNPs were separated from mutation SNPs using methods described previously

For mutation rates, there was difficulty in correlating the overall accumulation of mutation SNPs through time from the whole cluster owing to a narrow sampling timeframe (Supplementary Fig. 5). Therefore, correlations were performed within subclades of dominant clusters instead of using each whole cluster to capture signals (Supplementary Fig. 6 and Supplementary Note). Mutation rates were calculated with BEAST39 using the skyline population size prior and a relaxed lognormal clock model (tabulated in Supplementary Table 3). Comparison of the nucleotide substitution rate in different clusters was conducted using the Kruskal–Wallis test.

Rates of homologous recombination (r/m) were calculated given numbers of recombination events and the number of single polymorphic sites produced by the algorithm described in ref. 4, excluding any signals located in the regions of mobile genetic elements. Note that this is different from r/m calculation as originally used in ref. 40 to avoid a bias introduced by the genetic distance between the recombination donors and recipients (Supplementary Note). We calculated r/m using two different approaches. The first approach involved modeling the relationship between recombination events and mutations as a linear regression (Supplementary Fig. 7), using ranked recombination events as the outcome and the ranked number of SNPs as the predictor variable. The slope represents r/m (Supplementary Table 4). We used the ANCOVA test to determine the significant difference in recombination rates between clusters when its statistical assumptions were met. The second approach involved using the arithmetic mean of r/m for a cluster, averaged from the r/m of each branch within a cluster. The mean of the distribution of r/m for the cluster was reported in Supplementary Table 4. The Kruskal–Wallis test was used to test for differences in r/m between clusters calculated by arithmetic mean.

Recombination hotspots and associations of some hotspots with temporal changes in antibiotic consumption. Hotspots are defined as the genome locations where recombination events have occurred at higher frequency. Here we used the 95th percentile of each cluster’s own population site frequencies as a cutoff. Recombination events observed in major lineages, the number of excluded events owing to mobile genetic elements and the number of events associated with hotspots are summarized in Supplementary Table 8. To investigate the level of admixture of each hotspot gene, pbl1a, pbl2b, pbl2x, folA and folP, phylogenies of individual gene trees (Supplementary Figs. 1a–c and 4b), as well as a concatenated phylogenetic tree (Fig. 4a), were constructed with RAxML v7.0.4 (ref. 41) using a GTR model with a gamma correction for site rate variation using 100 bootstraps.

The trend of recombination was estimated through the detected phenotypes observed in the presence and absence of recombination in the subpopulation including the seven most prevalent clusters. Note that five isolates with missing phenotypes (Supplementary Table 1) were not included in this analysis. On the basis of the prediction of recombination from seven dominant clusters, strains undergoing recombination at pbl1a, pbl2b, pbl2x, folA or folP and their phenotypic resistance to β-lactam and cotrimoxazole were compared against the strains with no recombination events observed at these sites (Supplementary Table 9). The statistical significance of differences between the recombining group and the non-recombining group was estimated by two-tailed Fisher’s exact test. Alternative murM and murN genes associated with high β-lactam resistance42 were also considered. However, only two candidates with partial matches were observed and are thus less likely to explain trends in β-lactam resistance.

Temporal trends in recombination were determined by comparing phenotype difference in strains showing evidence of recent recombination (recombination events predicted at the external branches) to strains whose ancestors had undergone recombination (recombination events predicted at internal nodes). Using a two-tailed Fisher’s exact test, the statistical difference between these groups was estimated.

Trends in antibiotic consumption obtained from recommended treatments are tabulated in Supplementary Table 5.

Searching for potential recombination donors given recipient blocks. On the basis of the sequence identity of the recombination fragments detected in recipient strains, potential donors from the rest of the population were assessed. The number of recipient blocks used for this analysis is summarized in Supplementary Table 8. As described further (Supplementary Figs. 8 and 9 and Supplementary Note), several criteria were applied to maximize detection specificity and reduce the detection of false positives. Probabilities of a single isolate, as well as of each BAPS cluster, acting as a donor for a recipient
were then calculated. Further, we evaluated relationships between cluster size, cluster diversity and probability of being a donor. On the basis of Spearman ranking correlation, the association between both features and the probability of their becoming donors were estimated.

**Visualization of phylogenetic trees.** Display and manipulation of phylogenetic trees was performed using the online tool Interactive Tree of Life43 and the software package Circos44.

**Statistical tests.** All statistical tests and associated diagrams were generated in R version 2.11.1. Statistical analyses are discussed in relevant sections in the text.

23. O’Brien, K.L. & Nohynek, H. Report from a WHO working group: standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*. Pediatr. Infect. Dis. J. 22, 133–140 (2003).
24. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008).
25. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578–579 (2011).
26. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. Genome Biol. 13, R56 (2012).
27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
28. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
29. Croucher, N.J. et al. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* spaType2F ST81. J. Bacteriol. 191, 1480–1489 (2009).
30. Harris, S.R. et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474 (2010).
31. Corander, J., Waldmann, P. & Sillanpaa, M.J. Bayesian analysis of genetic differentiation between populations. Genetics 163, 367–374 (2003).
32. Corander, J. & Tang, J. Bayesian analysis of population structure based on linked molecular information. Math. Biosci. 205, 19–31 (2007).
33. 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).
34. Corander, J., Connor, T.R., O’Dwyer, C.A., Kroll, J.S. & Hanage, W.P. Population structure in the Neisseria, and the biological significance of fuzzy species. J. R. Soc. Interface 9, 1208–1215 (2011).
35. Muthreja, A. et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477, 462–465 (2011).
36. Cheng, L., Connor, T.R., Siren, J., Ananensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol. Biol. Evol. 30, 1224–1228 (2013).
37. Willems, R.J. et al. Restricted gene flow among hospital subpopulations of *Enterococcus faecium*. mBio 3, e00151–12 (2012).
38. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490 (2010).
39. Drummond, A.J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214 (2007).
40. Feil, E.J., Maiden, M.C., Achtman, M. & Spratt, B.G. The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. 16, 1496–1502 (1999).
41. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690 (2006).
42. Smith, A.M. & Klugman, K.P. Alterations in MurM, a cell wall muropeptide branching enzyme, increase high-level penicillin and cephalosporin resistance in *Streptococcus pneumoniae*. Antimicrob. Agents Chemother. 45, 2393–2396 (2001).
43. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39, W475–W478 (2011).
44. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645 (2009).