Frequency-dependent selection in vaccine-associated pneumococcal population dynamics

Frequency-dependent selection in vaccine-associated pneumococcal population dynamics

Authors:

Jukka Corander[1,2,3], Christophe Fraser[4], Michael U. Gutmann[5], Brian Arnold[6], William P. Hanage[6], Stephen D. Bentley[3], Marc Lipsitch[6,7], Nicholas J. Croucher[8,*]

Affiliations:

[1] Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[2] Department of Biostatistics, University of Oslo, 0317 Oslo, Norway

[3] Infection Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[4] Oxford Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[5] School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK

[6] Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

[7] Departments of Epidemiology and Immunology & Infectious Diseases, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

[8] MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, St. Mary's Campus, Imperial College London, London, W2 1PG, UK

*Correspondence to: Nick Croucher, n.croucher@imperial.ac.uk.

Abstract:

Many bacterial species are composed of multiple lineages distinguished by extensive variation in gene content. These often co-circulate in the same habitat, but the evolutionary and ecological processes that shape these complex populations are poorly understood. Addressing these questions is particularly important for *Streptococcus pneumoniae*, a nasopharyngeal commensal and respiratory pathogen, as the changes in population structure associated with the recent introduction of partial-coverage vaccines have significantly reduced pneumococcal disease. Here we show pneumococcal lineages from multiple populations each have a distinct combination of intermediate frequency genes. Functional analysis suggested these loci were likely subject to negative frequency-dependent selection (NFDS) through interactions with other bacteria, hosts, or mobile elements. Correspondingly, these genes had similar frequencies in four populations with dissimilar lineage compositions. These frequencies were maintained following substantial alterations in lineage prevalences once vaccination programmes began. Fitting a multilocus NFDS model of post-vaccine population dynamics to three genomic datasets using Approximate Bayesian Computation generated reproducible estimates of the influence of NFDS on pneumococcal evolution, the strength of which varied between loci. Simulations replicated the stable frequency of lineages unperturbed by vaccination, patterns of serotype switching, and clonal replacement. This framework highlights how bacterial ecology affects the impact of clinical interventions.

Population genomics has revealed many bacterial species exhibit extensive variation in their 'accessory' genomes. While neutral evolutionary models can account for such diversity[1–3], allowing for heterogeneity in the evolutionary rate between genes significantly improves their fit to genomic data[4–6], consistent with selection causing differences in gene content[7]. If recombination rates are sufficiently high, selection can alter the distribution of individual genes[8]. However, lower levels of recombination are associated with chromosome-wide sweeps, such that niche specialization at one or more loci can result in largely clonal 'ecotypes'[9,10]. Similarly, a recent model suggested selection acting on a high proportion of the genome could partition even freely-recombining bacteria into highly-diverged 'metabolic types'[11]. As well as adaptation to particular niches, this latter model[11] considered antigenic loci to be under negative frequency-dependent selection (NFDS), the situation in which alleles are most beneficial to genotypes when they are rare. This is based on the assumption antigens become more costly when common, because they are more frequently recognised by acquired immune responses.

Such NFDS has been proposed to explain the extensive antigenic diversity of the nasopharyngeal coloniser and respiratory pathogen *Streptococcus pneumoniae* (the pneumococcus)[11–13]. This variation makes anti-pneumococcal vaccine development challenging. The first licensed conjugate vaccine (PCV7) targeted seven of over ninety serotypes[14], and consequently was associated with 'serotype replacement' as vaccine types (VTs) were replaced by non-vaccine types (NVTs), with no overall change in carriage rates[15]. This was driven by both serotype switching, the replacement of VTs by NVTs that differed at few loci other than that which determined the serotype, and clonal replacement of VTs by distantly-related NVTs. These population dynamics are now amenable to detailed study, having been tracked by genomic surveillance of isolates carried by children in both Massachusetts (USA) [14] and Southampton (UK) [16,17], and isolates from invasive

pneumococcal disease in adults in Nijmegen (the Netherlands) [18]. Here we use the distribution of the accessory genome across isolates to develop a gene frequency-based model of bacterial population structure based on multiple NFDS mechanisms[19,20].

Results

Enrichment of loci under frequency-dependent selection in the accessory genome

Previous analyses of 5,442 clusters of orthologous genes (COGs) in the Massachusetts pneumococcal population suggested those present at intermediate frequencies were important in distinguishing sequence clusters[21]. To identify functions that were enriched in this set of genes, the 1,112 COGs present in 5% to 95% of isolates and 1,194 core COGs[14] were annotated by integrating multiple analyses (Fig 1a & Supplementary Datasets 1 and 2). The most substantive difference was in mobile genetic elements (MGEs; Fisher's exact test; odds ratio, OR = 336; two-sided $p < 2.2 \times 10^{-16}$). However, few of these genes were 'cargo' beneficial to the host bacterium, and were instead likely to be parasitic, consistent with the distribution of prophage between pneumococci[21,22]. Correspondingly, restriction modification systems (RMSs) that protect against MGE infection accounted for 2.4% of the intermediate-frequency genes, but were absent from the core COGs. These are most often advantageous when rare, such that the donor of an infecting MGE is unlikely to have the same system[19], but typically futile when ubiquitous. Hence the co-existence of lineages likely involves competition between bacteria and MGEs through 'kill-the-winner' dynamics, a form of NFDS in which an increase in a genotype's frequency would be associated with a counterbalancing rise in the prevalence of MGE genotypes able to infect such cells[23].

Annotation also suggested direct interference competition between bacteria was likely to be important in maintaining a diversity of lineages[24]. Bacteriocins, which mediate interstrain

killing[25], were significantly enriched in the accessory genome relative to the core (Fisher's exact test; OR = 24.0; two-sided $p < 2.2x10^{-16}$). Although regulatory components of the bacteriocin-like peptide (*blp*) locus were conserved across the population, most of the gene cluster was composed of various combinations of bacteriocin and immunity protein genes, many of which were found in multiple loci[26]. Despite this diversity, each of the previously-described fifteen monophyletic sequence clusters[14] was typically associated with one distinctive *blp* allele (Supplementary Fig 1), with an exception being sequence clusters (SCs) 3 and 14, which did not co-exist for long owing to vaccine-induced population dynamics[14].

Sequence clusters also varied in their complement of rarer bacteriocin biosynthesis gene clusters, including pneumocyclicin[27], pneumolancidin[28], two loci likely regulated by the TprA/PhrA quorum-sensing system[29], and other putative loci (Supplementary Fig 1). No individual gene cluster replicated the diversity of the *blp* locus, with sequence variation instead often corresponding to disruptive mutations in bacteriocin structural or biosynthetic genes. Assuming relevant phenotypes can be reliably inferred from the gene clusters, such mutations result in bacteria immune to the bacteriocin, but unable to kill competitors. These immune non-producers co-circulate with producer cells carrying the putatively fully-functional allele, and susceptible cells completely lacking the gene cluster. Analogous variation with respect to individual bacteriocins is likely present between the *blp* loci, given their diverse complements of production and immunity genes. If both biosynthesis and immunity functions are costly, these phenotypes can co-exist through rock-paper-scissors NFDS dynamics as producers kill susceptible cells, immune non-producers outcompete producers, and susceptible cells outcompete immune non-producers[30]. Hence the distinctive overall bacteriocin production profile of strains may be shaped by NFDS acting on multiple loci.

NFDS can also result from competition for resources[20,31]. A particular nutrient import strategy, either optimized for different nutrients[31] or different concentrations of the same nutrient[32], will become less advantageous as it becomes more common, as a consequence of more intense competition for the same resource[24]. While nutrient importers account for 11.5% of the core COGs, because many are universally necessary, they also make up 9.35% of the intermediate-frequency COGs. Hence they are significantly enriched relative to general metabolic genes in the latter category (Fisher's exact test, OR = 2.48, two-sided p = 2.61×10^{-8}). This suggests NFDS may sustain multiple nutrient acquisition strategies in the population as a consequence of interstrain competition for resources.

Antibiotic resistance, also variable between isolates, could be affected by similar competition[33]. If resistant bacteria are considered adapted to hosts consuming antibiotics, but suffering a cost in untreated hosts, then resistance will be most effective as a resource acquisition strategy where rare owing to the lessened competition with other strains. This could directly result in NFDS, although there are alternative explanations for the co-existence of sensitive and resistant pneumococci that instead imply NFDS through other mechanisms[34].

A further functional category to be enriched in the intermediate-frequency COGs relative to the core genome were genes encoding for the biosynthesis of immunogenic structures, such as surface proteins[35] or the capsule[36] (Fisher's exact test, OR = 2.56, two-sided p = 9.23×10^{-10}). These can be under NFDS as long as alleles are immunologically distinguishable, a criterion met by the serotype-defining capsule[36], as well as accessory antigens that are typically either present as large surface structures, or completely absent, such as the pili[37].

Therefore multiple disparate functions enriched in the intermediate-frequency genes relative to the core genome can each be understood as being subject to NFDS, albeit through different processes. While no NFDS mechanism could be identified for 32.1% of the intermediate frequency COGs, this category is likely to include both metabolic enzymes and signal transduction proteins linked to loci under NFDS on genomic islands, and loci under NFDS that cannot be identified as such owing to incomplete functional information. To test whether these inferences applied to other pneumococcal populations in a similar manner, further genomic datasets were compared to those from Massachusetts.

Population similarities in frequencies of genes, but not genotypes

Overall, 4,127 isolates were combined from available reference sequences, Massachusetts, Southampton, Nijmegen, and the Maela refugee camp in Thailand where the population is unvaccinated[38] (Supplementary Dataset 3). A new analysis identified 11,049 'global' COGs (gCOGs), from which a 'relaxed' core of 1,447 gCOGs was extracted to generate a maximum likelihood phylogeny (Fig 1b; Supplementary Fig 2). Strikingly, there was little evidence of genetic isolation-by-distance, as both vaccine-type status and country of isolation had a polyphyletic distribution, indicating a history of recombination and frequent international migration.

The core alignment was also used to define 74 sequence clusters. Plotting the pairwise core genome divergence of isolates, represented by their cophenetic separation in the tree, against their accessory genome divergence, calculated as the Jaccard distance between the isolates' gCOG content, demonstrated members of the same sequence cluster were substantially more similar in their accessory, as well as core, genomes (Fig 1c). These differences between lineages were likely biologically meaningful, as they represented a significant proportion of the accessory genome and were preserved despite international dissemination of some genotypes and ongoing horizontal DNA transfer. Although some of

the previously-identified atypical unencapsulated lineages were associated with extensive private gene content[21], sequence clusters of encapsulated pneumococci each contained few unique accessory loci. The mean numbers of gCOGs present in ≥95% of the isolates in a given sequence cluster, but not meeting this criterion in any other sequence cluster in the same population, were just 16.75 in Massachusetts, 19.94 in Southampton, 19.46 in Nijmegen, and 15.02 Maela (Supplementary Fig 2). Sequence clusters' distinctiveness instead resulted from the polyclonal distribution of the 1,731 intermediate frequency gCOGs, present in between 5% and 95% of the pre-vaccination isolates in at least one population (Supplementary Fig 2). Hence a long history of recombination was reflected in intermediate-frequency loci being associated with multiple lineages, with each lineage in turn defined by a unique combination of intermediate-frequency loci.

Despite the lineages representing discrete and distinct sets of genotypes, their prevalences were highly heterogeneous between the four populations, with a significant correlation only between those in Massachusetts and Southampton (Fig 2a). In marked contrast, the frequencies of accessory gCOGs were strongly correlated between Massachusetts and every other population (Fig 2b; Pearson correlation, two-sided $p < 10^{-15}$ in all comparisons). This suggests pneumococcal populations are configured by genomic islands being maintained at equilibrium frequencies that are conserved between populations, consistent with their prevalence being influenced by NFDS[19]. A significant deviation between populations was the elevated frequency of Tn*916* in Maela; this transposon underlies tetracycline resistance[21], and therefore the difference is likely to represent a location-specific selection pressure rather than drift[39]. Hence selection appears to shape pneumococcal populations to be similar in frequencies of genes, rather than genotypes.

Vaccination as a test of negative frequency-dependent selection

The partial-coverage vaccines introduced to limit pneumococcal disease can be used as a natural experiment, to test whether loci expected to change in frequency due to association with VTs were actually maintained at equilibrium frequencies by NFDS. Although a significant correlation existed between pre- and post-PCV7 sequence cluster frequencies in the three vaccinated populations (Fig 2c), divergence in population composition was driven by the replacement of some VT sequence clusters with distantly-related NVT lineages. Across all comparisons of pre- and post-PCV7 populations, gCOG frequencies showed a stronger positive correlation. This stability in gene frequencies reflected the significant correlation between the post-PCV7 decrease in a gCOG's absolute frequency in VT isolates, and the contemporaneous increase in its absolute frequency in NVT isolates (Supplementary Fig 3), consistent with the NFDS hypothesis. The greatest deviation in the Massachusetts population was *wciN*, directly involved in the synthesis of the vaccine-targeted 6A and 6B capsules, reflecting differences in selection pressures between timepoints[14]. This suggested the equilibrium frequencies of the intermediate frequency gCOGs were likely to govern the post-vaccine restructuring of the population.

To quantify whether NFDS of intermediate frequency gCOGs could explain changes in pneumococcal populations better than a neutral model, a discrete time Wright-Fisher multilocus NFDS model was constructed in which the number of offspring produced by a genotype i at generation t, $X_{i,t}$, was distributed as:

$$X_{i,t} \sim Pois\left(\left(\frac{\kappa}{N_t}\right)(1-m)(1-v_i)\left(1+\sigma_f\right)^{\pi_{i,t}}\right)$$

General density-dependent competition was parameterised by the total number of pneumococci in the simulated population at time t, N_t, and the environment's carrying

capacity, κ. This was constant across *t*, reflecting the stable levels of pneumococcal carriage post-PCV7[15,16]. The other demographic process was migration, at rate *m* (per month-long generation), by which isolates in the resident simulated population were replaced by genotypes randomly selected from the genomic data from the same location. VT genotypes were subject to a fitness cost, *v*, representing vaccine efficacy at preventing transmission. The final term parameterised NFDS, the strength of which was determined by σ_f and the exponent $\pi_{i,t}$:

$$\pi_{i,t} = \sum_{l=1}^{L} g_{i,l}(e_l - f_{l,t})$$

where *l* is an intermediate frequency locus (gCOG or antibiotic resistance phenotype), and $g_{i,l}$ is a binary variable indicating whether *l* is present in genotype *i*. Each *l* has an equilibrium frequency e_l, its prevalence in the pre-vaccination sample, and an instantaneous frequency at generation *t*, $f_{l,t}$. Therefore $f_{l,t}$ determines whether *l* benefits its host, when it is rare relative to e_l, or has a net cost, when it is common relative to e_l. Model details are described in Supplementary Fig 4 and the Methods.

The σ_f, *v* and *m* parameters were estimated for the Massachusetts population using Approximate Bayesian Computation, an inference technique for intractable simulator-based models[40,41]. The simulated population was compared to the sequence cluster distribution across three time points (Fig 3a) using the Jensen-Shannon divergence (JSD) to determine similarity. Convergence of the parameter estimates found strong evidence for NFDS (σ_f significantly greater than its lower bound; Table 1 & Supplementary Table 1, Supplementary Fig 5). The precedent of other models[4,6] suggested the fit could be improved by allowing the strength of selection to be heterogeneous across loci. Hence an expanded

model featured a proportion, p_f, of the intermediate frequency loci experiencing NFDS at strength σ_f, while (1-p_f) experienced NFDS at strength σ_w (see Methods). Convergence of parameter estimates again found strong evidence for NFDS (σ_f and p_f significantly greater than their lower bounds; Table 1 & Supplementary Table 1, Supplementary Fig 5), with a substantial improvement over the homogeneous selection model, as quantified by the significantly smaller JSD values from appropriately parameterised simulations (Wilcoxon test on 100 simulation pairs, W = 9902, two-sided p = 4.73×10^{-33}; Supplementary Fig 6).

At the locus level, those genes subject to stronger NFDS stabilised close to their equilibrium frequencies, whereas the frequencies of those subject to weaker NFDS drifted near-neutrally (Supplementary Fig 6). At the lineage level, these simulations replicated three important facets of the post-vaccination population dynamics (Fig 3a & Supplementary Fig 7). The first was the stable post-vaccine prevalence of some NVT sequence clusters, such as SC4 and SC8. The second was serotype switching, the replacement of VT by NVT within sequence clusters that remained at stable overall frequencies, as observed in SC1, SC5, SC9 and SC15. The third was clonal replacement of VT by unrelated NVT, such as the contemporaneous disappearance of SC13, SC14, SC22 and SC24, and expansion of SC3, SC6, SC7 and SC11. These trends were not trivial to replicate. The same framework was used to fit a neutral model (NFDS eliminated, with $\sigma_f = 0$); a serotype-focused single locus NFDS model (e_l applied to serotype, rather than locus, frequencies), and an ecotype model (e_l applied to sequence cluster, rather than locus, frequencies). Both the neutral and serotype models poorly reproduced the stability of SC8's frequency, serotype switching within SC9 and SC15, or any patterns of clonal expansion. The ecotype models better reproduced NVT sequence cluster stability and serotype switching, but did not replicate the observed patterns of clonal replacement. All of these models resulted in significantly worse fits to the data than the heterogeneous multilocus NFDS model (Supplementary Figs 6 & 7).

The estimated vaccine selection strength, v, of 0.081 per month from the heterogenous rate multilocus NFDS is consistent with PCV7's halving of the rate at which VT are acquired[42], if pneumococci transmit at least once every six months, an interval similar to the carriage duration of VT serotypes[43]. Similarly, the estimated migration rate, m, of 0.0044 per month suggests half the resident Massachusetts pneumococcal population is replaced by immigrant strains over approximately 13 years, which is realistic given the 50% probability that a pneumococcal lineage was detectable in different localities within Massachusetts after 3-4 years[14].

Consistent evidence of NFDS in other populations

The homogeneous and heterogeneous multilocus NFDS models were also fitted to similar surveillance data from Southampton (Supplementary Fig 5 & 8). The JSD values for the heterogenous rate model were again reproducible and significantly smaller than for the homogeneous rate version (Wilcoxon test on 100 simulation pairs, $W = 9954$, two-sided $p = 1.01 \times 10^{-33}$). The point estimates of parameter values were again robust and, in the case of the three parameters determining the strength of NFDS, very similar to those for Massachusetts (Table 1 & Supplementary Table 1).

However, the vaccine selection strength was estimated to be 2.54-fold higher in Southampton than in Massachusetts. This difference is likely attributable to the substantially higher PCV7 coverage in children under 24 months of age in the years immediately after the vaccine's introduction in the UK relative to the USA[44,45], combined with the lower age range included in the Southampton study, excluding older children who are less likely to have been immunized, or in whom natural acquisition of immunity blunted the selective pressure of the vaccine[16,46]. Simulations using these point estimates again

replicated the strain dynamics observed in the genomic sample (Fig 3b). VT SC5 and SC18 were eliminated at realistic rates; NVT SC3, SC19 and SC35 remained at stable frequencies; serotype switching occurred within SC1 and SC9, while NVT SC2 rose in prevalence at a much faster rate than same lineage did in Massachusetts.

The homogeneous and heterogeneous rate multilocus NFDS models were also fitted to a genomic dataset from cases of invasive pneumococcal disease in Nijmegen. The heterogeneous model was again a significantly closer fit to the genomic data, as assessed by the JSDs (Wilcoxon test, $W = 3988$, two-sided $p = 0.0135$; Table 1 and Supplementary Fig 5). Precisely replicating the observed population dynamics was difficult (Fig 3c and Supplementary Fig 9), owing to the sparser sampling, particularly post-PCV7, and inevitable bias towards more invasive genotypes in this dataset. While the estimated strength of NFDS was similar to both Massachusetts and Southampton, the estimated vaccine selection strength was lower than in these infant carriage surveillance projects, consistent with the Nijmegen collection being isolated in an adult population primarily protected by herd immunity[18]. Correspondingly, fitting the heterogenous rate model to the Maela dataset, isolated from an entirely unvaccinated community, estimated v close to zero (Table 1).

NFDS acting on genomic islands can also affect variation in the core genome. Comparisons between pre- and post-vaccination populations, and between different locations, revealed allele frequencies of core genome single nucleotide polymorphisms (SNPs) typically showed very similar correlations to those of accessory loci frequencies (Supplementary Fig 10). This was not a consequence of tight linkage between SNPs in the regions flanking genomic islands (Supplementary Fig 10). Nevertheless, simulations in which NFDS acted on only accessory loci precisely replicated the post-vaccination changes in the core SNP allele frequencies, and similar correlations to those between collections were observed in

simulations where the Massachusetts population was gradually replaced with isolates from other datasets (Supplementary Fig 10). Therefore while it is possible core genome loci may also be under NFDS, the observed correlations can be attributed to NFDS acting only on accessory loci.

Consequences of NFDS for the impact of vaccination

Simulations were used to investigate counterfactual scenarios. In the absence of vaccination ($v = 0$), the pre-PCV7 populations were stable in Massachusetts (Supplementary Fig 7), Southampton (Supplementary Fig 8) and Nijmegen (Supplementary Fig 9). Eliminating migration ($m = 0$) significantly increased the proportion of VTs observed in simulations in all three populations (Wilcoxon tests; Massachusetts, $W = 0$, two-sided $p = 2.56\text{x}10^{-34}$; Southampton, $W = 0$, two-sided $p = 2.56\text{x}10^{-34}$; Nijmegen, $W = 1453$, two-sided $p = 4.50\text{x}10^{-18}$), highlighting the importance of imported or previously rare NVTs in driving out VTs. However, removing NFDS significantly decreased the proportion of VTs observed in all three populations (Wilcoxon tests; $W = 10000$, two-sided $p = 2.56\text{x}10^{-34}$ in Massachusetts and Southampton; $W = 9979$, two sided $p = 4.81\text{x}10^{-34}$ in Nijmegen). This is because following vaccination, those loci enriched in VT genotypes become increasingly advantageous to their bacterial hosts as they become rarer, resulting in NFDS slowing the rate at which VT genotypes are eliminated until such loci rise in frequency in NVT genotypes.

Discussion

These combined analyses of multiple population genomic datasets suggest that NFDS plays an important role both in the stable structuring of pneumococcal populations, and their dynamics following disruption by vaccine-induced immunity. According to the best-fitting model, relatively strong NFDS acts on a few hundred accessory genes, corresponding to

5.0% of the Massachusetts pangenome, and 8.3% of that in Southampton. This cumulative effect across multiple loci in complex populations is predicted to maintain stable lineage compositions in the absence of disruption by vaccination, without the oscillatory dynamics associated with some single locus NFDS processes[19,30,47,48]. Hence multiple lineages can persistently coexist within this framework despite their confinement to a niche, the human nasopharynx, that is physiochemically homogeneous compared with the varied environments inhabited by species often considered as split into ecotypes, such as *Escherichia coli*. Furthermore, although intraspecific recombinations are slow over the timescales simulated in this study[14], horizontal DNA transfer has comprehensively reassorted genomic islands between genotypes over the species' history. Their consequent polyclonal distribution means accessory locus frequencies can be preserved by multiple lineage combinations, thereby accounting for the diverse population structures observed globally, and the panoply of strains they contain[49]. While the NFDS processes represented in the multilocus model were also sufficient to explain the major post-vaccination population changes, further work is required to determine whether core loci are also involved. Continued development of such quantitative models with large genomic datasets promises to improve our understanding of how diverse selective pressures affecting bacterial populations shape their response to public health interventions, and how best to design novel pathogen control strategies.

Methods

Annotation of the accessory genome

The previously analysed Massachusetts population[14,50] contained 1,112 COGs present in between 5% and 95% of the 616 isolates and 1,194 COGs present in a single copy in every isolate. Information on whether these were associated with capsule polysaccharide synthesis, antibiotic resistance, RMSs, Pneumococcal Pathogenicity Island 1 or MGEs was

extracted from previously described analyses[14,21,50]. Coding sequences (CDSs) associated with proteinaceous immunogenic structures were identified through the results of protein antigen array data[35]. Candidate bacteriocins were identified using the BAGEL3 algorithm[51]. The variation at the *blp* locus, and the other putative bacteriocin production loci, was manually identified within *de novo* assemblies of the Massachusetts isolates using Artemis and ACT[52]. The heatmap showing the distribution of the *blp* alleles in Supplementary Fig 1 was generated by mapping Illumina reads for each of the Massachusetts isolates against the concatenated set of loci using BWA with default settings[53]. Further information on COG functional domains[14] and previous automated annotations[50] was additionally used to manually curate all available information into the annotation and classification in Supplementary Datasets 1 and 2.

Bioinformatic analysis of genomic data

The isolate collections analysed each came from systematic sampling of defined host populations. The Massachusetts pneumococcal dataset, isolated from the nasopharynxes of children up to five years of age during routine primary care physician visits, consisted of the 616 *de novo* assemblies generated with Velvet[54] as described previously[14,50]. VelvetOptimiser[55] was used to assemble data from the Maela collection[38] (3,085 genomes), isolated from the nasopharynxes of infants up to two years of age, and their mothers, in a Thai refugee camp; the Southampton collection[16] (516 genomes), isolated from the nasopharynxes of children up to four years of age during outpatient visits; and the Nijmegen[18] collection (337 genomes), isolated from adults hospitalised with bacteraemic pneumonia. These were supplemented with 20 complete, publically available reference genomes (Supplementary Dataset 3). To standardise these genome collections relative to the Massachusetts dataset, assemblies were discarded if they were less than 1.98 Mb, or greater than 2.19 Mb, in length; or had an N50 less than 15 kb[14,50]; or necessary information

was absent from the public databases. Of the 4,586 genomes, 4,462 met these criteria and were included in a preliminary analysis that identified non-pneumococcal streptococci, which were then excluded from the final analysis. Consequently, the final dataset of 4,127 genomes contained 20 reference sequences, 616 Massachusetts sequences, 491 Southampton sequences, 337 Nijmegen sequences, and 2,663 Maela sequences.

Each genome was processed with RNAmmer v1.2, to annotate rRNA[56]; tRNAscan-SE v1.3.1, to annotate tRNA[57]; Rfam scan, to annotate other non-coding RNA[58]; scanned for BOX, RUP and SPRITE repeats using HMM profiles[59,60]; and Prodigal v2.6[61] , to annotate CDSs using a model trained on the genome of *S. pneumoniae* ATCC 700669[62]. CDSs that overlapped with the non-coding RNA or short interspersed repeat sequences were then removed from the annotation, and the remaining set translated to allow a non-redundant set of proteins to be identified. A version without low complexity regions was generated by filtering with segmasker[63] and masking choline binding domains. All-against-all comparisons of these protein databases were then generated using BLAT v0.34[64]. Global COGs (gCOGs) were then generated using COGtriangles and COGcognitor[65], and through linking pairs of highly similar sequences, as described previously[14]. The gCOG nomenclature was then applied to the full, redundant set of protein sequences.

To correct for misassemblies, particularly those reflecting differences between the methods used to assemble the Massachusetts isolates' genomes and those from other populations, false positive CDSs were eliminated from the intermediate frequency gCOGs. A database generated from the annotation of *S. pneumoniae* ATCC 700669[62] was used to search intermediate frequency gCOG DNA sequences using BLASTALL v2.2.25. This identified 39 gCOGs corresponding to fragments of tRNA, oligomers of choline binding domains, or antisense fragments of insertion sequences. This left a final set of 11,049 gCOGs, of which

1,731 were present at a frequency between 5% and 95% in the pre- or peri-vaccination samples (grouped as "pre-vaccination" samples in the Results section) of at least one of the four study populations.

To transfer the functional annotation onto the gCOG sequences, the annotated protein sequences from Massachusetts in Supplementary Table 1 were used to identify identical proteins in the gCOG dataset. When COGs could not be matched to gCOGs through this approach, links were instead made through searching gCOGs for proteins with identity to the middle 50% of annotated protein sequences from Massachusetts. These links were then manually curated to categorise the 1,731 intermediate frequency gCOG sequences where possible, as shown in Fig 2.

Analysis of population structure

To analyse the overall population structure, a 'relaxed' core set of 1,447 gCOGs were identified that met two criteria: first, that they were present in at least 95% of the isolates; and second, that the total number of gCOG representatives was less than 105% of the number of isolates containing the gCOG, to exclude gCOGs that are present in high copy number in some, or all, genomes. A codon alignment was then generated for each gCOG using mafft v7.221[66], excluding any sequences from isolates containing more than one representative of the gCOG. These were concatenated, with gap sites used to pad regions where data were missing for a particular isolate, and a 293,508 bp alignment of polymorphic sites extracted using SNP-sites[67]. A phylogeny was generated from this alignment using FastTree2 with the 'fastest' option[68].

Population structure was analysed with hierarchical BAPS clustering[69] using five independent runs of the estimation algorithm starting from the upper bound of 200-500

clusters, which all converged to the same posterior mode. Two polyphyletic primary BAPS clusters were split into their secondary level clusters, yielding 73 sequence clusters that were almost entirely congruent with the phylogeny, and SC0, which remained polyphyletic. The monophyletic sequence clusters most similar to those in Massachusetts[14] were numbered accordingly. The plot in Fig 1c combined cophenetic distances from the core genome phylogeny, extracted with Bioperl[70], and Jaccard distance calculated from the presence and absence matrix of gCOGs using the R package vegan[71]. For each isolate, 100 comparator isolates were selected at random, and this sample of pairwise comparisons used to generate the plot.

Of the polymorphic sites in the core genome, 282,043 corresponded to a base in the *S. pneumoniae* ATCC 700669 reference genome. For each population, the set of sites that were both biallelic and had a non-reference allele frequency between 5% and 95% in that population were extracted with VCFtools v0.1.14[72]; there were 27,616 of these in the Massachusetts dataset, 26,954 in the Southampton dataset, 28,396 in the Nijmegen dataset, and 30,579 in the Maela dataset. The r^2 statistics between these polymorphic sites, and between the binary presence and absence information of accessory gCOGs with a representative in the *S. pneumoniae* ATCC 700669 genome, were then calculated with VCFtools by treating each isolate as a phased haplotype. These were used to generate the linkage analysis plots in Supplementary Fig 10.

Inference of antibiotic resistance profiles

Individual isolates' genotypes were used to predict their antimicrobial resistance profiles. The presence of *aph*3' (the gCOG CLS350021) was inferred to cause resistance to aminoglycosides; the presence of *tetM* (CLS03712) was inferred to cause resistance to tetracycline; the presence of *cat* (CLS01043) was inferred to cause resistance to

chloramphenicol; and the presence of *ermB* (CLS01283), *mef* (CLS02227), or both was inferred to cause macrolide resistance[62,73]. These gCOGs themselves were removed from the set of loci used in the simulations, and the inferred antibiotic resistance phenotype used instead.

Non-susceptibility to other antibiotics is determined by core genome loci; to incorporate these into the model, resistant alleles of relevant loci were treated analogously to the presence of an accessory resistance gene. The presence of the I100L substitution in the dihydrofolate reductase protein (CLS03211) was inferred to result in resistance to trimethoprim[74,75], and the presence of an insertion shortly after S61 in the dihydropteroate synthase protein (CLS01442) was inferred to result in resistance to sulphamethoxazole[76]. Three penicillin-binding proteins substantially contribute to β-lactam resistance. Using a similar approach to that of Li *et al*[77], the population-wide protein sequences of Pbp1A (CLS01776), Pbp2X (CLS01031) and Pbp2B (CLS01093) were aligned with mafft v7.221[66], and the transpeptidase domain regions extracted. Following validation using the isolates from Massachusetts[14], sequences exhibiting less than 97% amino acid identity with the susceptible alleles defined by Li *et al* in the multiple sequence alignment were considered resistance-associated. These antibiotic resistance phenotypes were included as intermediate frequency loci if they met the criteria for a given population.

Multilocus negative frequency dependent selection model

The multilocus negative frequency dependent selection model was generated within a discrete-time Wright-Fisher framework[78,79]. Although such models were designed with a number of strong assumptions, the results of simulations have been found robust to violations of these conditions[80]. Each individual i had a genotype g_i defined by a binary string representing the presence and absence of each gCOG or antibiotic resistance

phenotype present at an intermediate frequency in the starting population. The number of offspring arising from i at time t is a Poisson-distributed random variable $X_{i,t}$. This Poisson approximation is justifiable if only a small proportion of descendants survive to the next generation[78], as is likely to be the case for a nasopharyngeal coloniser with a small within-host effective population size[81] that experiences a strong bottleneck at transmission. To allow for differential reproductive success between genotypes in a manner that depended on the composition of the overall population, $X_{i,t}$ was parameterised using the function (Supplementary Fig 4):

$$X_{i,t} \sim Pois\left(\left(\frac{\kappa}{N_t}\right)(1-m)(1-v_i)\left(1+\sigma_f\right)^{\pi_{i,t}}\right)$$

The four components of the function each correspond to a different biological process. General density-dependent selection depends on κ, the carrying capacity of the environment, and N_t, the total number of individuals at time t. This maintained an approximately stable population size throughout simulations. This is justifiable, as *S. pneumoniae* colonization levels did not substantially change in the years immediately after PCV7's introduction.

Migration into the population occurred at rate m, subject to the limits $0 \leq m \leq 1$, and therefore the reproductive fitness of resident individuals was reduced by a factor of (1-m) accordingly to maintain an approximately constant population size of κ. The number of immigrating individuals at time t, $N_{m,t}$, was a random variable calculated as:

$$N_{m,t} \sim Bin(m, \kappa)$$

Migrant individuals were selected, with replacement, from all isolates observed at any time point in the geographically-specified dataset being studied. Therefore it was the only mechanism by which genotypes not present in the pre-vaccine genome samples could enter the simulated population. To prevent artefactually improving the fit of the model at high values of m through sampling all isolates in proportion to their observed frequency, the selection of an immigrating isolate was biased such that it was equally likely to come from any sequence cluster with at least one representative in the studied population, although these were present at very different frequencies within each population. Hence the probability of an immigrating individual being of genotype i and sequence cluster s, $p_{m,s,i}$, was:

$$p_{m,s,i} = \frac{n_{s,i}}{Sn_s}$$

Where S is the number of sequence clusters in the population, $n_{s,i}$ is the number of isolates in sequence cluster s of genotype i in the genome dataset, and n_s was the number of isolates in the sequence cluster s in the genomic dataset.

The vaccine selection pressure to which individual i was subject, v_i, depended on whether the individuals were of a vaccine serotype or not; for PCV7, the vaccine serotypes were 4, 6B, 9V, 14, 18C, 19F and 23F, as well as 6A, a vaccine-related type to which PCV7 elicited strong cross-immunity[14]. Consequently, v_i was determined as:

$$v_i \begin{cases} v & if\ isolate\ has\ a\ vaccine\ serotype \\ 0 & otherwise \end{cases}$$

Where v was subject to the constraint $0 \leq v \leq 1$.

In the homogeneous rate multilocus model of NFDS, the magnitude of this pressure was determined by the term $\left(1+\sigma_f\right)^{\pi_{i,t}}$, where $\sigma_f \geq 0$. The selection pressure depended on the genotype g_i and distribution of intermediate frequency loci at time t, as summarised by the exponent $\pi_{i,t}$. The calculation of $\pi_{i,t}$ necessitated determining the frequency $f_{l,t}$ of each locus l at time t in the simulation, using the binary variables $g_{i,l}$ that represent presence or absence of l in i:

$$f_{l,t} = \frac{\sum_{i=1}^{N_t} g_{i,l}}{N_t}$$

These were compared to the equilibrium frequencies, e_l, of the same loci, which were assumed to correspond to their frequencies in the sample of G_0 genomes from isolates sampled pre- or peri-vaccination:

$$e_l = \frac{\sum_{i=1}^{G_0} g_{i,l}}{G_0}$$

The overall deviation of the L accessory genome loci included in the simulations, for individual i at time t, $\pi_{i,t}$, was calculated as:

$$\pi_{i,t} = \sum_{l=1}^{L} g_{i,l}\left(e_l - f_{l,t}\right)$$

Therefore if all accessory genes are at their equilibrium frequencies, then $\left(1+\sigma_f\right)^{\pi_{i,t}} = 1$, and NFDS has no effect on an individual's reproductive fitness. When a genotype contains

many genes rarer than their equilibrium frequencies, $\left(1+\sigma_f\right)^{\pi_{i,t}} > 1$, and NFDS increases an individual's reproductive fitness. Lastly, when a genotype contains many genes more common than their equilibrium frequencies, $\left(1+\sigma_f\right)^{\pi_{i,t}} < 1$, and therefore NFDS reduces an individual's reproductive fitness. In the absence of *l* from an individual's genotype, $f_{l,t}$ has no direct effect on its fitness.

Extension to heterogeneous frequency-dependent selection

Two further parameters were introduced when accessory genes were split into two categories, each subject to a different level of frequency dependent selection. The σ_w parameter represented the strength of weaker NFDS acting on a fraction, $(1-p_f)$, of the accessory genes included in the model. To facilitate inference of these two parameters, it was assumed that loci under weaker negative frequency dependent selection would vary in frequency to a greater extent between the initial and final genomic samples; therefore the accessory loci were ordered by the statistic Δ_l:

$$\Delta_l = \frac{(f_{l,t>0} - e_l)^2}{(1 - e_l(1 - e_l))}$$

Where e_l is the frequency of the gCOG or antibiotic resistance phenotype across all pre- or peri-vaccination samples, as defined previously, and $f_{l,t>0}$, is its frequency across all post-vaccination samples. The denominator is intended to emphasise the effects of gCOGs at frequencies of approximately 50%, which are likely to have a large effect on the overall population structure. The proportion p_f of genes for which Δ_l was smallest were considered subject to NFDS with strength $(1+\sigma_f)$, whereas the rest where subject to NFDS of strength $(1+\sigma_w)$. If the *L* loci were ordered by ascending values of Δ_l, then l_f was the highest ranking

meeting the criterion, $\frac{l_f}{L} \leq p_f$. This resulted in two distinct measures of the deviation of $f_{l,t}$ from e_l:

$$\pi_{i,t} = \sum_{l=1}^{l_f} g_{i,l}(e_l - f_{l,t})$$

And:

$$\omega_{i,t} = \sum_{l=l_f+1}^{L} g_{i,l}(e_l - f_{l,t})$$

Hence the modified offspring distribution was:

$$X_{i,t} \sim Pois\left(\left(\frac{\kappa}{N_t}\right)(1-m)(1-v_i)\left[(1+\sigma_f)^{\pi_{i,t}} + (1+\sigma_w)^{\omega_{i,t}}\right]\right)$$

Simulations and parameter estimation

The model was implemented in C++ using the GNU scientific library, and is available for download from https://github.com/nickjcroucher/multiLocusNFDS. In each simulation, genotypes were represented by the gCOGs and antibiotic resistance phenotypes present in between 5% and 95% of the pre- or peri-vaccination population. Hence L was 1,090 for Massachusetts, 1,175 for Southampton, 1,090 for Nijmegen and 1,254 for Maela. For simplicity, κ was assumed to represent the number of pneumococci likely to transmit between individuals in the sampled population. This was estimated to correspond to 25% colonisation of children under ten years of age in the USA and European samples. In Massachusetts[15], an under ten population of 828,129 in 2000 [82] implied a bacterial population size of 2×10^5 (10^5 was actually used for model fitting for computational

efficiency; comparing simulations demonstrated this had no detectable effect on the results); in Southampton (including Hampshire and Portsmouth), an under ten population of 202,404 in 2011[83] implied a bacterial population size of $5x10^4$; and in Nijmegen (including Arnhem), an under ten population of 77,753 in 2011[84] implied a bacterial population of $2x10^4$. An elevated colonisation rate of 50%[85] was used for Maela, where estimating that 15% of the 40,000 residents being under 10 implied a bacterial population size of $3x10^3$.

Each simulation was run for a number of timesteps corresponding to the number of months spanned by the genomic collection, excluding early or late years in which sampling was sparse. The well-sampled periods were the 72 months between spring 2001 and spring 2007 for Massachusetts[50]; the 48 months between spring 2007 and spring 2011 for Southampton[16]; the 120 months between 2001 and 2011 for Nijmegen[18]; and the 24 months between 2007 and 2009 for Maela[85]. All isolates from a single winter were assigned to the year in which the season ended. In simulations of the Nijmegen population, where a substantial proportion of samples pre-dated the vaccine's introduction, $v = 0$ for years up to 2007. In each case, the starting population for the simulation, of size κ, was generated by randomly resampling with replacement from the genotypes present in the pre- and peri-vaccination samples in each study; hence the 'pre-vaccination' population consisted of isolates sampled up to spring 2001 in Massachusetts, up to spring 2007 in Southampton, and up to 2007 in Nijmegen. These were the genomic samples used to calculate e_l for all intermediate frequency loci; all later samples were used to calculate $f_{l,t>0}$ in the definition of Δ_l.

At each time t at which a genomic sample was available, the equivalent number of genotypes was randomly sampled from the simulated population. The similarity between

the simulated and genomic samples at t was then calculated as the Jensen-Shannon divergence[86] (JSD_t) between the real and simulated samples:

$$JSD_t = \sum_{\forall s} \sum_{v=0}^{v=1} \left[\frac{1}{2} \left(f_{t,s,v} ln \left(\frac{f_{t,s,v}}{f_{t,s,v} + a_{t,s,v}} \right) \right) + \frac{1}{2} \left(a_{t,s,v} ln \left(\frac{a_{t,s,v}}{f_{t,s,v} + a_{t,s,v}} \right) \right) \right]$$

Where $f_{t,s,v}$ is the simulated frequency of genotypes of sequence cluster s and vaccine type status v at time t, and $a_{t,s,v}$ is the equivalent value from the genomic sample. This value was summed over all vaccine type statuses and sequence clusters for each timepoint sampled in the genomic dataset to calculate the overall divergence of the simulation from the sampled data.

Each set of simulations was run with variation in the parameters v (range 0-0.5); m (range 0-0.2); σ_f (range 10^{-6}-0.22); σ_w (range 10^{-6}-0.15; only in the heterogeneous rates model), and p_f (range 0-1; only in the heterogeneous rates model). Model fitting was achieved through Approximate Bayesian Computation with the BOLFI algorithm[40], run for 2,000 iterations of Bayesian optimisation to identify best-fitting parameter sets through minimizing the JSD (Table 1 & Supplementary Fig 5). Point estimates of parameter values were generated based on the Gaussian process minimisers, with the distribution of the projected JSD values shown for each fit in Supplementary Fig 5. Exploration of parameter space was performed with logarithmically transformed values to avoid discontinuity of the approximate likelihood function near the natural boundary and to enable better fit of the Gaussian process regression. The 95% posterior credible intervals for the parameters were obtained using three generations of sequential Monte Carlo sampling with the same default settings as used in Gutmann and Corander[40] for the pneumococcal day care transmission model.

Alternative model formulations

To test whether equivalently good fits to the genomic data could be achieved using different approaches within the same framework, alternative model formulations were tested. The neutral model was fitted in the same way as the multilocus NFDS models, except that σ_f was fixed at zero. The serotype NFDS model assumed all serotypes were present at equilibrium frequencies in the pre-vaccine samples, and therefore $\pi_{i,t}$ was calculated as the deviation of an isolate's serotype from its initial frequency. This was fitted using both the homogeneous and heterogeneous selection rate models. In the latter case, Δ_l was calculated by comparing the serotype e_l values with their post-vaccination frequencies, as for the intermediate frequency loci. The same parameter ranges were used as for the multilocus NFDS model, except σ_f and σ_w were allowed to take values within the range 10^{-6}-25 to compensate for the single locus contributing to $\pi_{i,t}$ and $\omega_{i,t}$. Additionally, to avoid many lower frequency sequence clusters evolving neutrally, serotypes were considered to be at intermediate frequencies if they were between 1% and 99% prevalence in the pre- or peri-vaccination population.

The ecotype NFDS model assumed each sequence cluster was adapted to a specific ecological niche, and therefore was present at an equilibrium frequency in the pre-vaccine samples. Therefore $\pi_{i,t}$ and $\omega_{i,t}$ were calculated as the deviation of an isolate's sequence cluster from its initial frequency. This was fitted using both the homogeneous and heterogeneous selection rate models using the same parameter ranges and intermediate frequency range as for the serotype NFDS model, as well as the same approach to the calculation of Δ_l. For both the serotype and ecotype models, fitting was conducted with BOLFI[40] as for the multilocus NFDS model, using JSDs to quantify the differences between the simulated and sampled populations. Results are shown in Table 1. These implementations are not intended to represent the optimal versions of each model, but

instead demonstrate that the fits of the multilocus NFDS models cannot be trivially replicated through changing the genetic basis of NFDS.

Simulations in which isolates from two populations were combined used the pre-vaccination population from Massachusetts and post-vaccine isolates from one of the alternative populations. The initial population was drawn only from Massachusetts; both these isolates, and those from the alternative dataset, could enter the simulated population through migration. When the alternative population corresponded to Southampton or Nijmegen, the population size, number of generations, parameter point estimates, Δ_l and e_l values were those of the alternative population. When the alternative population was Maela, the e_l and Δ_l values were those of the alternative population, but the simulations were otherwise parameterised for the Massachusetts population, due to the difficulty of obtaining robust point estimates for parameters from the Maela population as a consequence of the lack of vaccine introduction in this location.

Statistical analyses

Statistical analyses, including calculation of Pearson's R^2, Wilcoxon tests, interquartile ranges and Fisher's exact tests, were performed in R[87]. Estimation of parameter values and credibility intervals through model fitting were performed with BOLFI[40]. All reported p values are two-sided.

Code availability

The model code used in this analysis is freely available from the GitHub repository, https://github.com/nickjcroucher/multilocusNFDS.

Data availability

The sequence datasets analysed in the current study are available in the public sequence databases with the accession codes listed in Supplementary Dataset 3. The epidemiological and phylogenetic data analysed in the current study are available from https://microreact.org/project/multilocusNFDS.

References:

1. Haegeman, B. & Weitz, J. S. A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13,** 196 (2012).
2. Baumdicker, F., Hess, W. R. & Pfaffelhuber, P. The infinitely many genes model for the distributed genome of bacteria. in *Genome Biology and Evolution* **4,** 443–456 (2012).
3. Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J. & Hanage, W. P. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microb. Genomics* **1,** 10.1099/mgen.0.000038 (2015).
4. Hogg, J. S. *et al.* Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8,** R103 (2007).
5. Collins, R. E. & Higgs, P. G. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.* **29,** 3413–3425 (2012).
6. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* **5,** 233–242 (2013).
7. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2,** 17040 (2017).
8. Shapiro, B. J. *et al.* Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science* **336,** 48–51 (2012).
9. Cohan, F. Bacterial species and speciation. *Syst Biol* **50,** 513–524 (2001).
10. Cohan, F. M. What are Bacterial Species? *Annu. Rev. Microbiol.* **56,** 457–487 (2002).

11. Watkins, E. R. *et al.* Vaccination Drives Changes in Metabolic and Virulence Profiles of *Streptococcus pneumoniae*. *PLoS Pathog.* **11,** e1005034 (2015).
12. Regev-Yochay, G. *et al.* Re-emergence of the type 1 pilus among *Streptococcus pneumoniae* isolates in Massachusetts, USA. *Vaccine* **28,** 4842–4846 (2010).
13. Cobey, S. & Lipsitch, M. Niche and neutral effects of acquired immunity permit coexistence of pneumococcal serotypes. *Science* **335,** 1376–1380 (2012).
14. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45,** 656–663 (2013).
15. Huang, S. S. *et al.* Continued impact of pneumococcal conjugate vaccine on carriage in young children. *Pediatrics* **124,** e1-11 (2009).
16. Gladstone, R. A. *et al.* Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine* **33,** 2015–2021 (2015).
17. Gladstone, R. A. *et al.* Pre-vaccine serotype composition within a lineage signposts its serotype replacement – a carriage study over 7 years following pneumococcal conjugate vaccine use in the UK. *Microb. Genomics* **3,** 119 (2017).
18. Cremers, A. J. H. *et al.* The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. *Sci. Rep.* **5,** 14952 (2015).
19. Levin, B. R. Frequency-dependent selection in bacterial populations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (1988). doi:10.1098/rstb.1988.0059
20. Maynard Smith, J. *Evolutionary Genetics*. *New York* **2,** (1998).
21. Croucher, N. J. *et al.* Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* **5,** 5471 (2014).
22. Croucher, N. J. *et al.* Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLOS Biol.* **14,** e1002394 (2016).
23. Takeuchi, N., Cordero, O. X., Koonin, E. V & Kaneko, K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. *BMC Biol.*

13, 20 (2015).

24. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12,** 263–273 (2014).

25. Dawid, S., Roche, A. M. & Weiser, J. N. The blp bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both in vitro and in vivo. *Infect Immun* **75,** 443–451 (2007).

26. Miller, E. L., Abrudan, M. I., Roberts, I. S. & Rozen, D. E. Diverse Ecological Strategies Are Encoded by *Streptococcus pneumoniae* Bacteriocin-Like Peptides. *Genome Biol. Evol.* **8,** 1072–90 (2016).

27. Bogaardt, C., van Tonder, A. J. & Brueggemann, A. B. Genomic analyses of pneumococci reveal a wide diversity of bacteriocins - including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics* **16,** 554 (2015).

28. Maricic, N., Anderson, E. S., Opipari, A. M. E., Yu, E. A. & Dawid, S. Characterization of a multipeptide lantibiotic locus in *Streptococcus pneumoniae*. *MBio* **7,** (2016).

29. Hoover, S. E. *et al.* A new quorum-sensing system (TprA/PhrA) for *Streptococcus pneumoniae* D39 that regulates a lantibiotic biosynthesis gene cluster. *Mol. Microbiol.* **97,** 229–243 (2015).

30. Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. M. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* **418,** 171–174 (2002).

31. Stewart, F. M. & Levin, B. R. Partitioning of Resources and the Outcome of Interspecific Competition: A Model and Some General Considerations. *Am. Nat.* **107,** 171 (1973).

32. Levin, B. R. Coexistence of two asexual strains on a single resource. *Science* **175,** 1272–1274 (1972).

33. Colijn, C. & Cohen, T. How competition governs whether moderate or aggressive treatment minimizes antibiotic resistance. *Elife* **4,** (2015).

34. Lehtinen, S. *et al.* Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 1075–1080 (2017).
35. Croucher, N. J. *et al.* Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E357–E366 (2017).
36. Croucher, N. J. *et al.* Selective and Genetic Constraints on Pneumococcal Serotype Switching. *PLoS Genet* **11,** e1005095 (2015).
37. Bagnoli, F. *et al.* A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. *J Bacteriol* **190,** 5480–5492 (2008).
38. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46,** 305–309 (2014).
39. Goossens, H. *et al.* Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* **365,** 579–587 (2005).
40. Gutmann, M. U. & Corander, J. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *J Mach Learn Res* **16,** (2016).
41. Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S. & Corander, J. Fundamentals and Recent Developments in Approximate Bayesian Computation. *Syst. Biol.* **66,** e66–e82 (2017).
42. Rinta-Kokko, H., Dagan, R., Givon-Lavi, N. & Auranen, K. Estimation of vaccine efficacy against acquisition of pneumococcal carriage. *Vaccine* **27,** 3831–3837 (2009).
43. Lipsitch, M. *et al.* Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model. *Epidemiology* **23,** 510–9 (2012).
44. Health Protection Agency COVER programme, October to December 2008: Quarterly

vaccination coverage statistics for children aged up to five years in the United Kingdom. *Heal. Prot. Rep.* **3,** 8–15 (2009).

45. Nuorti, J. P., Martin, S. W., Smith, P. J., Moran, J. S. & Schwartz, B. Uptake of pneumococcal conjugate vaccine among children in the 1998-2002 United States birth cohorts. *Am J Prev Med* **34,** 46–53 (2008).

46. Huang, S. S., Finkelstein, J. A., Rifas-Shiman, S. L., Kleinman, K. & Platt, R. Community-level predictors of pneumococcal carriage and resistance in young children. *Am J Epidemiol* **159,** 645–654 (2004).

47. Durrett, R. & Levin, S. Allelopathy in Spatially Distributed Populations. *J. Theor. Biol.* **185,** 165–171 (1997).

48. Gupta, S., Ferguson, N. & Anderson, R. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* **280,** 912–915 (1998).

49. Henriques-Normark, B., Blomberg, C., Dagerhamn, J., Bättig, P. & Normark, S. The rise and fall of bacterial clones: *Streptococcus pneumoniae*. *Nat. Rev. Microbiol.* **6,** 827–837 (2008).

50. Croucher, N. J. *et al.* Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci. Data* **2,** 150058 (2015).

51. van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J. & Kuipers, O. P. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res.* **41,** (2013).

52. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672–2676 (2008).

53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

54. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18,** 821–829 (2008).

55. Gladman, S. VelvetOptimiser. (2010). at <http://www.vicbioinformatics.com/software.velvetoptimiser.shtml>
56. Lagesen, K. *et al.* RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35,** 3100–3108 (2007).
57. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
58. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41,** (2013).
59. Croucher, N. J., Vernikos, G. S., Parkhill, J. & Bentley, S. D. Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics* **12,** 120 (2011).
60. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7,** (2011).
61. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
62. 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).
63. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
64. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12,** 656–664 (2002).
65. 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).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
67. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2,** (2016).
68. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood

trees for large alignments. *PLoS One* **5,** e9490 (2010).

69. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30,** 1224–1228 (2013).

70. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12,** 1611–1618 (2002).

71. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14,** 927–930 (2003).

72. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

73. Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331,** 430–434 (2011).

74. Pikis, A., Donkersloot, J. A., Rodriguez, W. J. & Keith, J. M. A conservative amino acid mutation in the chromosome-encoded dihydrofolate reductase confers trimethoprim resistance in *Streptococcus pneumoniae*. *J Infect Dis* **178,** 700–706 (1998).

75. Maskell, J. P., Sefton, A. M. & Hall, L. M. C. Multiple mutations modulate the function of dihydrofolate reductase in trimethoprim-resistant *Streptococcus pneumoniae*. *Antimicrob. Agents Chemother.* **45,** 1104–1108 (2001).

76. Haasum, Y. *et al.* Amino acid repetitions in the dihydropteroate synthase of *Streptococcus pneumoniae* lead to sulfonamide resistance with limited effects on substrate Km. *Antimicrob. Agents Chemother.* **45,** 805–809 (2001).

77. Li, Y. *et al.* Penicillin-Binding Protein Transpeptidase Signatures for Tracking and Predicting β-Lactam Resistance Levels in *Streptococcus pneumoniae*. *mBio* **7,** (2016).

78. Fisher, R. A. The Genetical Theory of Natural Selection. *Genetics* **154,** 272 (1930).

79. Wright, S. Evolution in Mendelian populations. *Genetics* **16,** 97–159 (1931).

80. Der, R., Epstein, C. & Plotkin, J. B. Dynamics of neutral and selected alleles when the

offspring distribution is skewed. *Genetics* **191,** 1331–1344 (2012).

81. Li, Y., Thompson, C. M., Trzciński, K. & Lipsitch, M. Within-host selection is limited by an effective population of *Streptococcus pneumoniae* during nasopharyngeal colonization. *Infect. Immun.* **81,** 4534–4543 (2013).

82. US Census Bureau. *Census 2000*. *US Census Bureau* (2000). at <http://quickfacts.census.gov/qfd/states/13/13135.html>

83. Office for National Statistics. *Census 2011*. *Census* (2011). at <http://www.ons.gov.uk/ons/guide-method/census/2011/index.html>

84. Statistics Netherlands *Dutch Census 2011*. (2011). at <https://ec.europa.eu/CensusHub2/>

85. 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).

86. Wong, A. K. C. & You, M. Entropy and Distance of Random Graphs with Application to Structural Pattern Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-7,** 599–609 (1985).

87. R Core Development Team *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2011).

Acknowledgments

We thank Dr Rebecca Gladstone, Dr Johanna Jefferies, Dr Saul Faust and Dr Stuart Clarke for sharing epidemiological data on the Southampton isolates. NJC was funded by a Sir Henry Dale fellowship, jointly funded by the Wellcome Trust and Royal Society (Grant Number 104169/Z/14/Z). JC was funded by the COIN Centre of Excellence. ML was funded by NIH grant R01 AI048935 and WPH by NIH grant R01 AI106786.

Author contributions

JC, CF, BA, WPH, ML and NJC designed the model; JC, MUG and NJC fitted the model; WPH, SDB and NJC analysed the genomic data; JC and NJC initially drafted the manuscript, with all authors contributing to the final version.

Competing financial interests

ML has consulted for Pfizer, Affinivax and Merck and grant support not related to this paper from Pfizer and PATH Vaccine Solutions. WPH, ML and NJC have consulted for Antigen Discovery Inc.

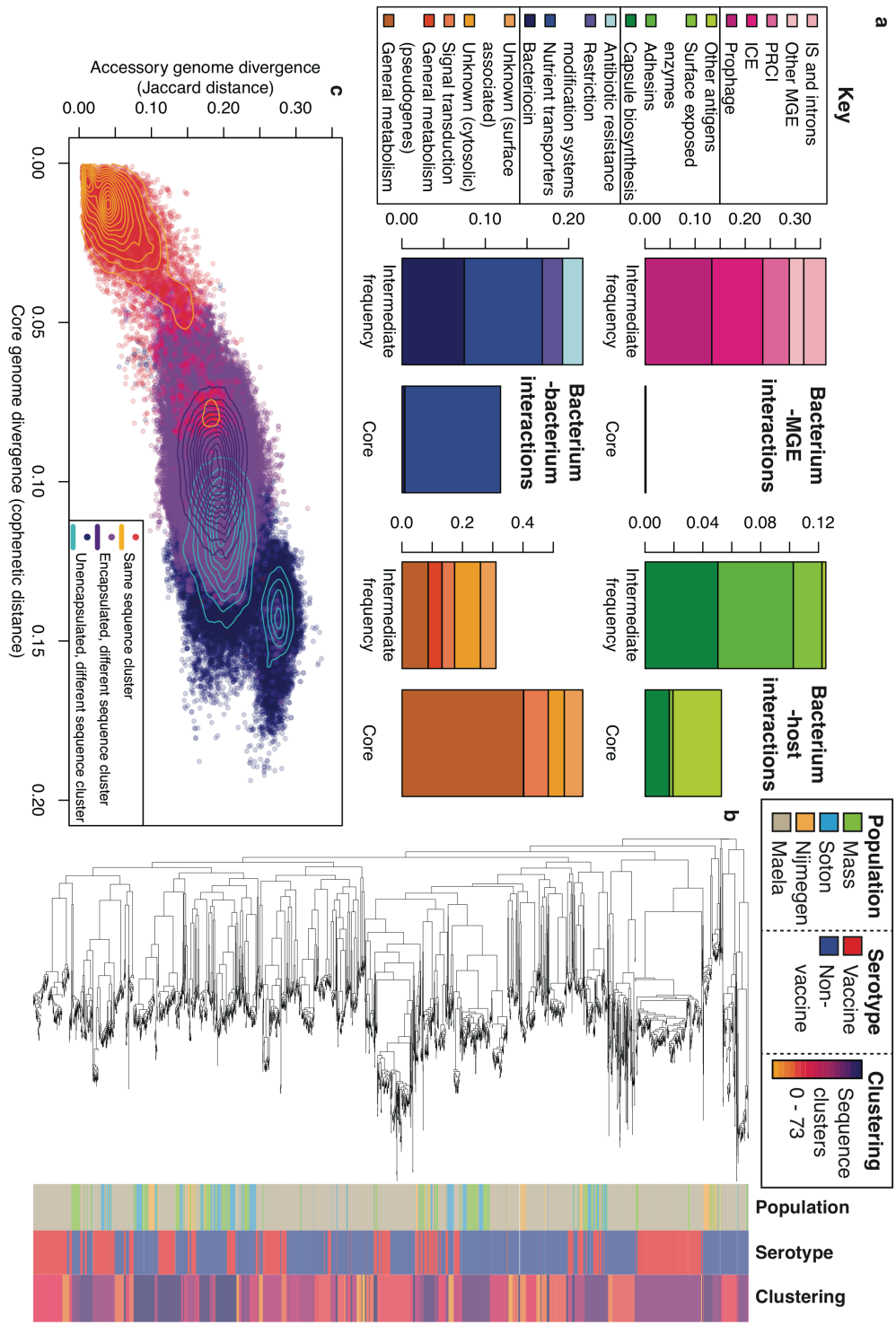
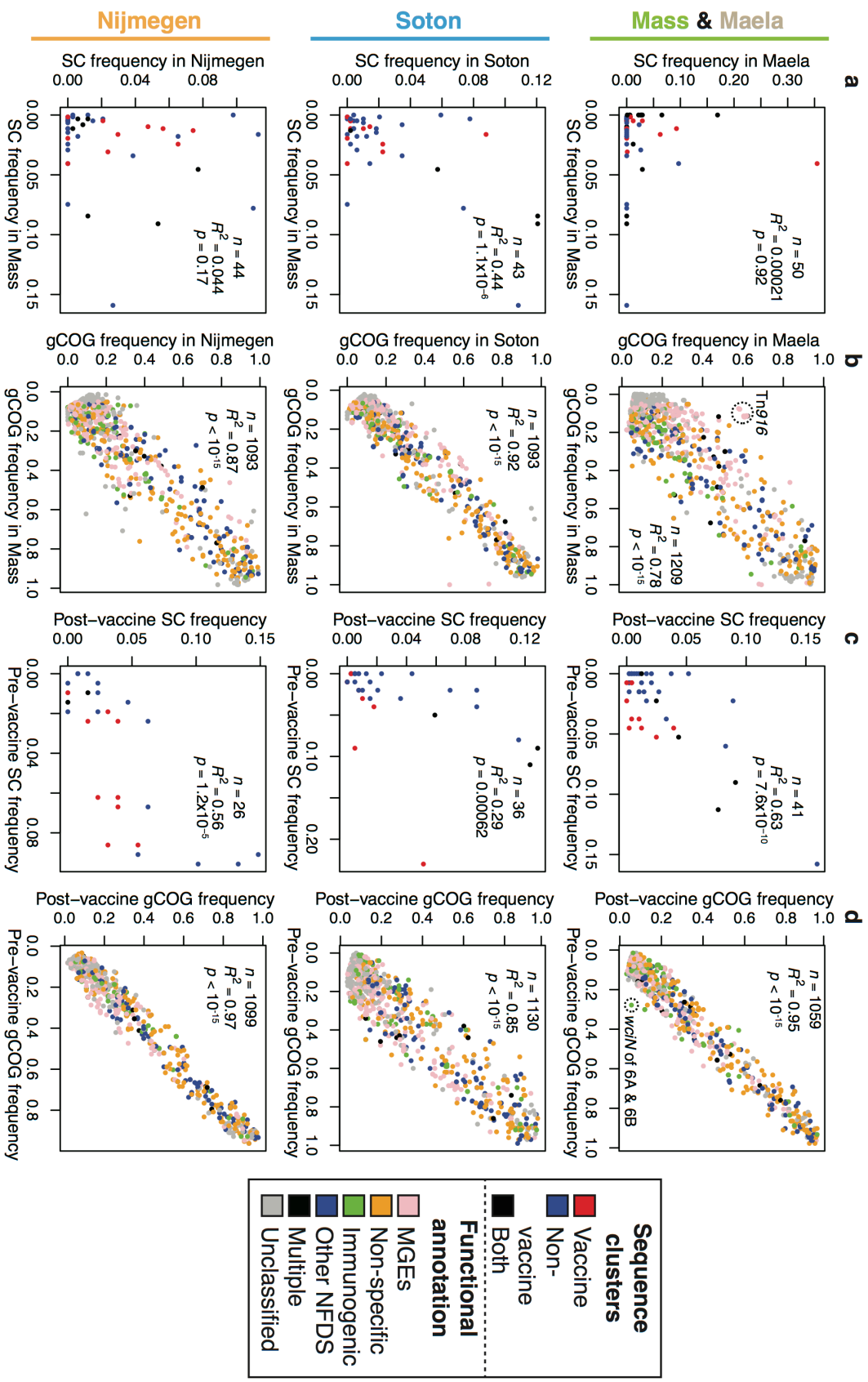

a
Key
IS and introns
Other MGE
PRCI
ICE
Prophage
Other antigens
Surface exposed enzymes
Adhesins
Capsule biosynthesis
Antibiotic resistance
Restriction modification systems
Nutrient transporters
Bacteriocin
Unknown (surface associated)
Unknown (cytosolic)
Signal transduction
General metabolism
(pseudogenes)
General metabolism
Bacterium -MGE interactions
Bacterium -host interactions
Bacterium -bacterium interactions
Intermediate frequency
Core
0.00
0.10
0.20
0.30
0.04
0.08
0.12
0.0
0.2
0.4
b
Population
Mass
Soton
Nijmegen
Maela
Serotype
Vaccine
Non-vaccine
Clustering
Sequence clusters
0 - 73
Population
Serotype
Clustering
c
Accessory genome divergence (Jaccard distance)
Core genome divergence (cophenetic distance)
Same sequence cluster
Encapsulated, different sequence cluster
Unencapsulated, different sequence cluster

Figure 1. Diversity and structure of the pneumococcal population. **a** Functional classification of the 1,112 intermediate-frequency and 1,194 core COGs in the Massachusetts pneumococcal population, as detailed in Supplementary Datasets 1 and 2. Each barchart compares the frequencies of functional categories in intermediate-frequency and core COGs. Categories are grouped as likely to be under NFDS resulting from bacterium-MGE interactions (pink segments), bacterium-bacterium interactions (blue segments), or bacterium-host interactions (green segments). The chart with orange segments shows the frequencies of loci with roles in general metabolism or signal transduction, or otherwise could not be classified. **b** Population structure of the 4,127 isolates from Massachusetts (Mass), Southampton (Soton), Nijmegen and Maela (Supplementary Dataset 3). The maximum likelihood phylogeny was generated from 1,447 core gCOGs. The adjacent columns contain a row for each genome, which represent the population in which the bacterium was isolated, its susceptibility to PCV7-induced immunity, and sequence cluster classification. **c** Comparison of core genome divergence, quantified as the cophenetic distance between isolates in the core genome phylogeny, and the accessory genome divergence, quantified as the Jaccard distance between the gCOG content of isolates. Each point is a pairwise comparison between randomly sampled isolates (excluding the polyphyletic SC0), which was coloured orange if the isolates belonged to the same sequence cluster; purple if they belonged to different sequence clusters but were both encapsulated; or otherwise dark blue, revealing the presence of some genetically divergent unencapsulated genotypes. Isocontour lines quantify the distribution of points in each category.


Mass & Maela
Soton
Nijmegen
a
b
c
d
SC frequency in Maela
SC frequency in Soton
SC frequency in Nijmegen
SC frequency in Mass
n = 50
R2 = 0.00021
p = 0.92
n = 43
R2 = 0.44
p = 1.1x10-6
n = 44
R2 = 0.044
p = 0.17
gCOG frequency in Maela
gCOG frequency in Soton
gCOG frequency in Nijmegen
gCOG frequency in Mass
Tn916
n = 1209
R2 = 0.78
p < 10-15
n = 1093
R2 = 0.92
p < 10-15
n = 1093
R2 = 0.87
p < 10-15
Post−vaccine SC frequency
Pre−vaccine SC frequency
n = 41
R2 = 0.63
p = 7.6x10-10
n = 36
R2 = 0.29
p = 0.00062
n = 26
R2 = 0.56
p = 1.2x10-5
Post−vaccine gCOG frequency
Pre−vaccine gCOG frequency
wciN of 6A & 6B
n = 1059
R2 = 0.95
p < 10-15
n = 1130
R2 = 0.85
p < 10-15
n = 1099
R2 = 0.97
p < 10-15
Sequence clusters
Vaccine
Non-vaccine
Both
Functional annotation
MGEs
Non-specific
Immunogenic
Other NFDS
Multiple
Unclassified

1

Figure 2. Distribution of genetic diversity between populations. Column (**a**) compares the distribution of sequence clusters between populations; the frequency of each sequence cluster in Massachusetts is shown on the horizontal axis, and the corresponding frequencies in Maela, Southampton and Nijmegen are shown on the vertical axes in the plots from top to bottom. Red points correspond to predominantly VT (≥75%) sequence clusters; blue points to predominantly NVT (≥75%) sequence clusters, and black points to mixed sequence clusters. Column (**b**) compares the distribution of gCOGs between populations. The frequency of each in Massachusetts is shown on the horizontal axis, and the corresponding frequencies in Maela, Southampton and Nijmegen are shown on the vertical axes. Only gCOGs present at a mean frequency between 5% and 95% across the two compared populations were included, and the corresponding points are coloured according to the functional annotation of COGs in Fig 1a. The elevated frequencies of gCOGs encoded by Tn*916*, including the *tetM* tetracycline resistance gene, in Maela are annotated. Column (**c**) compares the pre- and post-vaccination frequencies of sequence clusters in Massachusetts, Southampton and Nijmegen. Points are coloured as in (a), showing the general decline in the frequency of VT sequence clusters. Column (**d**) compares the pre- and post-vaccination frequency of gCOGs in Massachusetts, Southampton and Nijmegen. Only gCOGs with an overall frequency between 5% and 95% in the relevant population were included in the panels. Points are coloured as in (b). The reduced frequency of the *wciN* allele involved in synthesis of the VT 6A and 6B capsules is annotated. As the relationships between gCOG frequencies were linear, each panel displays Pearson's correlation statistics, including two-sided *p* values.

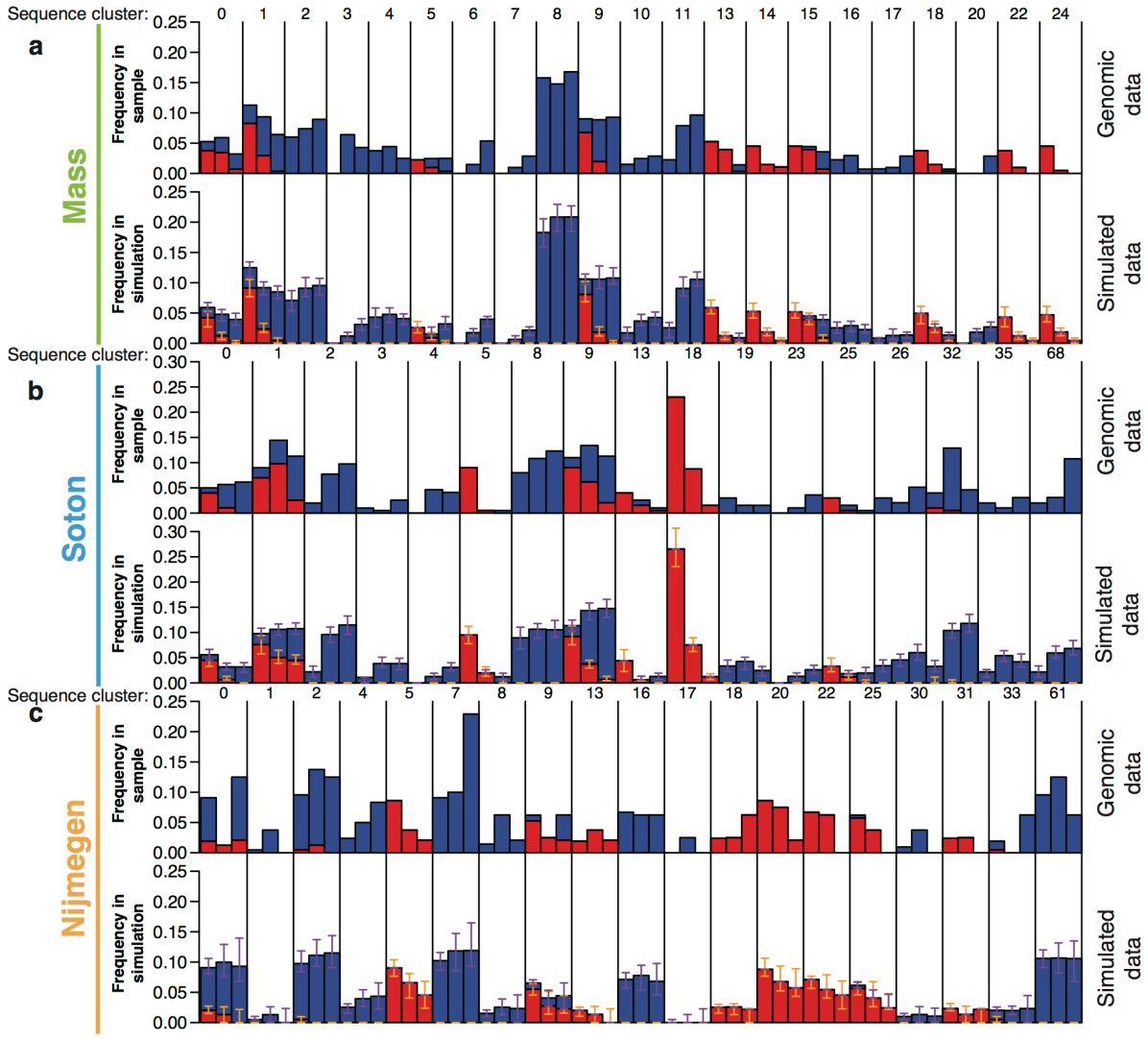


Figure 3. Comparing the sampled and simulated pneumococcal populations. In each barplot, the bacterial population is split into sequence clusters by vertical black lines, annotated at the top of the graph. Each sequence cluster is split into three timepoints: pre-vaccination, a midpoint sample and a late sample. Only sequence clusters present at greater than 2.5% frequency at one of these timepoints in the genomic sample are included in the graphs; full results are shown in the supplementary materials. The bars at each timepoint are split into red segments, for VT isolates, and blue segments, for NVT isolates. In each comparison, the top row is the genomic sample against which simulations were evaluated. The bottom row summarises the output of 100 simulations using the heterogeneous rate

multilocus NFDS model performed using the point estimate parameter values from Table 1. At the times at which samples were present in the respective genomic collections, the same numbers of isolates were randomly selected from the simulated populations. The bars represent the median result, and the error bars (orange for VT isolates, and purple for NVT isolates) represent the interquartile range observed across the simulations. (**a**) The results for Massachusetts split isolates into pre-vaccination (2001; 133 isolates), midpoint (2004; 203 isolates) and late (2007; 280 isolates) samples. (**b**) The results for Southampton, splitting isolates into pre-vaccination (up to 2007; 100 isolates), midpoint (2008-2009; 194 isolates) and late (2010-2011; 195 isolates) samples. (**c**) The results for Nijmegen, splitting isolates into pre-vaccination (up to 2007; 209 isolates), midpoint (2008-2009; 80 isolates) and late (2010-2011; 48 isolates) samples.

Tables

Population	Model	Maximal NFDS strength, σ_f	Vaccine selection strength, v	Migration rate, m	Proportion of loci under strong NFDS, p_f	Weaker NFDS strength, σ_w
Mass	Neutral	-	0.0375	0.0073	-	-
Mass	Homogeneous rate multilocus NFDS	0.0075 (0.0017 - 0.0234)	0.0733 (0.0430 - 0.1207)	0.0057 (0.0020 - 0.0131)	-	-
Mass	Heterogeneous rate multilocus NFDS	0.1363 (0.0213- 0.2113)	0.0812 (0.0491- 0.1254)	0.0044 (0.0015- 0.0165)	0.2483 (0.1197- 0.5448)	0.0023 (0.0010- 0.0514)
Mass	Homogeneous rate serotype NFDS	0.0333	0.0415	0.0071	-	-
Mass	Heterogeneous rate serotype NFDS	3.2613	0.0394	0.0053	0.1862	0.0127
Mass	Homogeneous rate ecotype	3.4514	0.0525	0.0090	-	-
Mass	Heterogeneous rate ecotype	1.0101	0.0541	0.0071	0.99	0.0009
Soton	Homogeneous rate multilocus NFDS	0.0028 (0.0010 - 0.0117)	0.1175 (0.0667 - 0.2262)	0.0032 (0.0011 - 0.0132)	-	-
Soton	Heterogeneous rate multilocus NFDS	0.1393 (0.0121 - 0.2148)	0.2063 (0.0832 - 0.3150)	0.0124 (0.0012 - 0.0394)	0.4035 (0.1005 - 0.5951)	0.0023 (0.0010 - 0.0238)
Nijmegen	Homogeneous rate multilocus NFDS	0.0605 (0.0012 - 0.0966)	0.0318 (0.0011 - 0.2621)	0.0018 (0.0009 - 0.0184)	-	-
Nijmegen	Heterogeneous rate multilocus NFDS	0.1462 (0.0013 - 0.2012)	0.0381 (0.0016 - 0.3235)	0.0015 (0.0009 - 0.0060)	0.1988 (0.0013 - 0.8356)	0.0032 (0.0010 - 0.1247)
Maela	Heterogeneous rate multilocus NFDS	0.1115 (0.0020 - 0.2138)	0.0011 (0.0010 - 0.0354)	0.0227 (0.0012 - 0.0568)	0.4995 (0.0028 - 0.9468)	0.0129 (0.0010 - 0.1416)

Table 1. Parameter estimates from model fits achieved through Approximate Bayesian Computation with BOLFI, run for 2,000 iterations. The displayed values represent point estimates of parameters generated based on the Gaussian process minimisers, with 95% credibility intervals in parentheses where calculated. The simplest neutral model required fitting only v and m to the genomic data. Homogeneous rate (σ_f, v and m) and heterogeneous rate (σ_f, v, m, p_f and σ_w) fits are shown for the multilocus NFDS model, in which

intermediate frequency gCOGs and resistance phenotypes have equilibrium frequencies; for the serotype NFDS model, in which serotypes have equilibrium frequencies; and for the ecotype model, in which sequence clusters have equilibrium frequencies. Replicate fits of the heterogeneous rate multilocus NFDS models to the Massachusetts, Southampton and Nijmegen datasets are shown in Supplementary Table 1 to demonstrate the robustness of the fitting process to stochastic effects.