Prediction of post-vaccine population structure of *Streptococcus pneumoniae* using accessory gene frequencies

Taj Azarian[1,2], Pamela P Martinez[2], Brian J Arnold[2], Lindsay R Grant[3], Jukka Corander[4,5,6], Christophe Fraser[7], Nicholas J Croucher[8], Laura L Hammitt[3], Raymond Reid[3], Mathuram Santosham[3], Robert C Weatherholtz[3], Stephen D Bentley[6], Katherine L O'Brien[3], Marc Lipsitch[2,9,*], William P Hanage[2*]

1 Burnett School of Biomedical Sciences, University of Central Florida, Orlando, FL;
2 Center for Communicable Disease Dynamics, Department of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston MA;
3 Center for American Indian Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland;
4 Helsinki Institute for Information Technology, Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland.
5 Department of Biostatistics, University of Oslo, 0317 Oslo, Norway;
6 Infection Genomics, The Wellcome Trust, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK;
7 Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK;
8 MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK;
9 Department of Immunology and Infectious Diseases, T.H. Chan School of Public Health, Harvard University, Boston MA.

Word Count: Abstract - 272, Main text - 2569
References: 27

Pamela P Martinez pmartinez@hsph.harvard.edu
Brian J Arnold brianjohnarnold@gmail.com
Lindsay R Grant lgrant10@jhu.edu
Jukka Corrander jukka.corander@medisin.uio.no
Christophe Fraser christophe.fraser@bdi.ox.ac.uk
Nicholas J Croucher n.croucher@imperial.ac.uk
Laura Hammitt lhammitt@jhu.edu
Raymond Reid rreid2@jhu.edu
Mathuram Santosham msantosham@jhu.edu
Robert R Weatherholtz rweathe1@jhu.edu
Stephen D Bentley sdb@sanger.ac.uk
Katherine L O'Brien klobrien@jhu.edu
Marc Lipsitch mlipsitc@hsph.harvard.edu
William P Hanage whanage@hsph.harvard.edu

*Co-senior authors

Corresponding Author:
Taj Azarian, PhD MPH
Burnett School of Biomedical Science, College of Medicine
University of Central Florida
taj.azarian@ucf.edu

Abstract

Predictions of how a population will respond to a selective pressure are valuable, especially in the case of infectious diseases, which often adapt to the interventions we use to control them. Yet attempts to predict how pathogen populations will change, for example in response to vaccines, are challenging. Such has been the case with *Streptococcus pneumoniae*, an important human colonizer and pathogen, and the pneumococcal conjugate vaccines (PCVs), which target only a fraction of the strains in the population. Here, we use recent advances in knowledge of negative-frequency dependent selection (NFDS) acting on frequencies of accessory genes (i.e., flexible genome) to predict the changes in the pneumococcal population after intervention. Implementing a deterministic NFDS model using the replicator equation, we can accurately predict which pneumococcal lineages will increase after intervention. Analyzing a population genomic sample of pneumococci collected before and after vaccination, we find that the predicted fitness of a lineage post-vaccine is significantly and positively correlated with the observed change in its prevalence. Then, using quadratic programming to numerically solve the frequencies of non-vaccine type lineages that best restored the pre-vaccine accessory gene frequencies, we accurately predict the post-vaccine population composition. Additionally, we also test the predictive ability of frequencies of core genome loci, a subset of metabolic loci, and naïve estimates of prevalence change based on pre-vaccine lineages frequencies. Finally, we show how this approach can assess the migration and invasion capacity of emerging lineages, on the basis of their accessory genome. In general, we provide a method for predicting the impact of an intervention on pneumococcal populations and other bacterial pathogens for which NFDS is a main driving force.

Detailed predictions of how a population will respond to a selective pressure are challenging. While evolutionary models specify how mutations with a given fitness vary in frequency over time, these are often hard to apply in practice, as we typically do not know in advance important parameters such as the fitness value of particular alleles or how this is affected by their frequency (frequency-dependent selection) or genetic background (epistasis) (*1*). Prediction is especially valuable in the case of infectious disease, as pathogens adapt to the interventions we use to control them (*2*). For example, pneumococcal conjugate vaccines (PCVs) target a fraction of the strains of *Streptococcus pneumoniae*, a colonizer of the human nasopharynx and common cause of bacterial pneumonia, bacteremia, meningitis, and otitis media (*3*). Before PCV use, there was concern that non-vaccine serotypes (NVT) could benefit from the removal of their vaccine-serotype (VT) competitors and thereby become more common in carriage and disease. Serotype replacement has indeed occurred following introduction of PCVs, with the gains from reducing VT disease partly offset by increases in NVT disease (*4–6*).

While serotype replacement has become evident, the scale of that replacement, NVT serotypes involved, and overall changes in the pathogen population structure were not appreciated until retrospective analysis (*7–9*). The apparent unpredictability of replacement is illustrated by our recent analysis of genomic data from 937 pneumococcal carriage isolates collected before and after vaccine introduction among Native American communities in the southwest United States (*9*). Before vaccine introduction, the population consisted of multiple lineages (often referred to as sequence clusters, SCs) including SCs comprising VT only, mixed VT and NVT, or NVT only (Figures 1 and 2). After vaccine introduction, there was non-uniform expansion of NVT SCs as well as the appearance of two previously unobserved SCs (*9*). There was considerable deviation from the null expectation that SCs including NVT would change in prevalence in proportion to their pre-vaccine frequency i.e., SC prevalence rank was not maintained (Figure 2, see Supplementary Information for details). Particularly, among 35 SCs, we find nine that increased significantly more than expected and five that increased significantly less. This illustrates the difficulty of prediction; even if we could be reasonably sure serotype replacement would occur, we would not have been able to say exactly which lineages would increase the most. Consequently, researchers are left playing a game of evolutionary whack-a-mole where post-vaccine pathogen surveillance is used to estimate the next epidemiologically important lineage

and determine subsequent vaccine formulations; then the cycle repeats. At best, this reduces the population impact of vaccination; at worst, it could unintentionally increase the prevalence of virulent or antibiotic resistant lineages (*10*).

One clue into the post-vaccine success of pneumococcal lineages may lie in the frequencies of the loci that make up the accessory genome (i.e., those genes not found in all strains within the population) (*11*, *12*). Corander and colleagues recently demonstrated that while the distribution of pneumococcal SCs were not correlated across diverse geographies, the frequency of accessory clusters of orthologous genes (COGs) were (*11*). Further, these frequencies were restored even after significant lineage perturbation induced by the introduction of PCV7 (*9*, *11*). Corander *et al*. 2017 went on to propose negative frequency-dependent selection (NFDS) as a mechanism for maintaining intermediate frequency loci. Similar processes, driven by host immunity, have been proposed to explain the co-existence of multiple serotypes (*13*) and vaccine-induced metabolic shifts among pneumococci (*14*).

Based on the observations of Corander *et al*, we hypothesized that predictions of accurate post-vaccination evolutionary dynamics could be made on the premise that after vaccination, SCs with accessory genomes that could best restore COG frequencies perturbed by removal of vaccine serotypes, would increase disproportionately to other SCs. To this end, we implemented a deterministic NFDS model using the replicator equation (*15*, *16*) to predict pneumococcal evolution after a perturbation from near equilibrium COG frequencies (eq. 1).

$$(1)\quad \frac{dx_i}{dt} = x_i(\omega_i - \varphi), \qquad \varphi = \sum_{j=1}^{n} x_j\,\omega_j$$

Under this formulation, x_i denotes the frequency of each i^{th} sequence cluster ($SC_i, i = \{1, \ldots, n\}$), n is the total number of SCs, and ω_i denotes the predicted fitness of SC_i (adapted from Ref. (*11*)), and φ is the average population predicted fitness. In this model, we define ω_i for each SC_i as the dot product of two vectors whose elements correspond to the COGs: a vector $k_{i,l}$ ($l = \{1, \ldots, n_{loci}\}$) with elements {0,1} for the absence or presence of the COG_l in SC_i, respectively, and a vector containing difference between the pre-vaccine frequency e_l of each

COG_l and f_l, which is the COG's expected frequency post-vaccination based on depleting the VT from the pre-vaccine population (eq. 2).

$$(2)\quad \omega_i = \sum_{l=1}^{L} k_{i,l}\,(e_l - f_l)$$

Intuitively, the vector $(e_l - f_l)$ represents the shape of the "hole" left in the population by vaccination, and $k_{i,l}$ quantifies the ability of SC_i to fill that hole. We make no explicit assumptions about carrying capacity, migration, mutation, or recombination rate, requiring only knowledge of the population structure (SCs) and COG frequencies before the intervention (e.g., vaccine); these quantities can be estimated from a pre-vaccine population survey with genome sequencing. However, there is an implicit assumption that over the study period recombination negligibly affects the frequency and distribution of COGs. Using simulated data, we first assessed the ability of a SC's standardized fitness ($\omega_i - \varphi$) immediately after intervention to predict the direction of change in SC frequencies from pre-vaccine to the post-vaccine equilibrium (Figure 3A). The predicted fitness represents the SC's ability to resolve the perturbation resulting from the vaccine-induced population bottleneck. Using this model, we show that the predicted fitness accurately estimates the direction of a simulated SC's adjusted frequency change (positive predictive value (PPV) = 99.9%, negative predictive value (NPV) = 83.9%, 1000 simulations), independent of the initial pre-vaccine SC frequency (Figure 3B). The accuracy of prediction was also robust to varying the proportion of the population removed between 5-30% of SCs to mimic the affect of PCV7 on the pneumococcal population.

Next, we asked whether this approach could predict the post-vaccine composition of real pneumococcal populations, and more specifically, the relative contribution of lineages to serotype replacement, without the need for full forward simulation. To test this, we evaluated a pneumococcal sample from the southwest US, comprised of 937 strains collected before and after the introduction of PCV7. For each NVT SC present pre-vaccine (considering NVT taxa only, in the case of SCs that contained both) we calculated their predicted fitness based on their accessory genome. We identified COGs as detailed in supplementary materials, using the 2371 COGs found in between 5% and 95% of the population and calculated the frequency of each COG among NVT taxa in each SC before vaccination. We found the predicted fitness value was significantly and positively correlated with the adjusted prevalence change – its change in

prevalence minus what would be expected if all NVT SCs increased by the same proportion from their pre-vaccine prevalence (Adjusted R^2=0.44, p<<0.001, Figure 4A). More than 90% of the SCs were accurately assigned based on whether they increased or decreased after vaccine (Figure 4A-B). SCs with a positive adjusted prevalence change had substantially higher predicted fitness than those with a negative one (p=0.012, Figure 4B).

While the predicted fitness accurately determines the direction of prevalence change, it does not provide the SC prevalence once the population has achieved post-vaccine equilibrium. To address this limitation we used quadratic programming (QP) to numerically identify the set of NVT SC prevalence that produced COG frequencies closest to those observed pre-vaccine (see Supplementary Information for more details). In short, assuming the pre-vaccine COG frequencies represent an equilibrium, we removed the VT population and then asked which combination (estimated as a proportion) of the NVT SCs present pre-vaccine best restored the equilibrium COG frequencies. QP accurately predicted SC frequencies following vaccination i.e., the 95% confidence interval of the observed vs. predicted post-vaccine SC frequencies included the line of equality (1:1 line), which denotes a perfect prediction, and the intercept and slope did not differ significantly from zero and one, respectively (p=0.26; intercept 95% CI: -0.005, 0.030; slope 95% CI: 0.257, 1.075, Figure 4C). In addition, QP also accurately predicted which SCs would have a positive prevalence change (PPV=71.4%, NPV=92.3%, Fisher's exact test score = 25.4, p=0.001, Figure 4D). These results were also robust to restricting the post-vaccine population to only those isolates collected in 2010 (n=119), prior to the introduction of PCV13 (Supplementary Information). In comparison, a naïve estimate based solely on pre-vaccine prevalence performed poorly (Figure 5A), as expected given the discordance in the pre- to post-vaccine rank changes illustrated in Figure 2. We further tested the predictive value of different genomic elements, finding that core genome loci (n_{loci}= 17,101) and metabolic loci (n_{loci}=5,853) were also capable of predicting the impact of vaccine (Figure 5). This finding must be considered in the context of recombination, selection, and the evolutionary timescale impacting the pneumococcal genome. Despite moderate levels of bacterial recombination among pneumococci, there remains appreciable linkage disequilibrium between loci nearby as well as genome-wide (*1*), which makes it difficult to discern the relative selective importance of

any particular locus. Operationally, COGs provide accurate prediction using significantly fewer loci and are easy to obtain using widely available genomic tools.

Given that the predicted fitness estimation requires the pre-vaccine SC prevalence, we can only retrospectively calculate the predicted fitness of the two SCs (SC10 and SC24) that emerged over the study period and compare them with samples collected elsewhere. Comparing with a carriage dataset of 1,354 pneumococci collected in Massachusetts children (*12*, *17*), we found that SCs 10 and 24, observed in the present sample only post-PCV, had higher predicted fitness than any of the other potential migrant SCs found only in Massachusetts and not our southwest US sample before vaccination, and higher predicted fitness than any SC seen in both carriage collections, except for SCs 23 and 9. As such, we can use this approach to ask which lineages are most likely to successfully invade following vaccination, and given that SCs 10 and 24 were present in USA carriage samples around the time of PCV7's introduction, they appear to have been primed for emergence.

Considering population structure and accessory genome content, post-vaccine COG frequencies may be restored by: 1) replacement by NVT SCs with varying degrees of relatedness in terms of core genome distance, or 2) clonal replacement by NVT strains belonging to SCs containing both VT and NVT taxa. In the southwest US sample, we observe both. Regarding the former, we find that the similarity of SCs in terms of COG presence-absence is only weakly associated with the phylogenetic distance between them, calculated from the core genome (see Supplemental Information; Supplemental Figures 2A and 3). Therefore, SCs may be divergent in core genome distance but share similar accessory genomes and comparable predicted fitness, as shown by the varied association between core and accessory loci and respective predicted fitness values (Supplemental Figure 2). A clear example was the post-PCV7 success of SC24, which possessed a high predicted fitness due to COG similarity with SC9. Regarding the latter, we find the NVT component of SCs containing both VT and NVT taxa (e.g., SC09 and SC23) possessed high predicted fitness, as expected under our model since these taxa are very similar to those that have been removed in both their core and accessory loci. Hence we should strongly expect the NVT part of any mixed SCs to increase post vaccination, especially since these NVT taxa are sometimes similar to their VT counterparts in terms of serotype properties such as capsule

thickness and charge, which are independently correlated with prevalence (*18*, *19*). A good example of this is the serotype 15B/C component of SC26, which we now predict to be successful following the more recent introduction of a vaccine incorporating six additional serotypes (PCV13) and which has indeed been noted to be increasing in recent samples (*20*–*22*). This information may be relevant for current vaccine considerations.

The potential of NFDS to structure a pathogen population is consistent with findings from environmental microbiology research on multiple bacterial species (*23*). Among pneumococci, changes in population dynamics after the introduction of vaccine have been explained by metabolic types, antibiotic resistance, carriage duration, recombination rates, and serotype competition, which may involve NFDS as well as other types of selection (*10*, *14*, *24*, *25*). In the case of three distinct pneumococcal carriage samples, COG frequencies consistently rebounded after being perturbed by PCV (*11*). Our approach, which does not require forward simulation, is predicated on the relatively simple hypothesis that SCs whose COG frequencies best resolve the PCV perturbation will be more successful in the post-vaccine evolutionary landscape. Indeed, we find a significant linear relationship between predicted fitness and the adjusted prevalence change of a SC. By optimizing NVT SC prevalence conditional on the pre-PCV7 COG frequencies equilibrium, we are able to recover observed post-PCV7 SC prevalence. Both the predicted fitness and numerical approximations of the post-vaccine equilibrium by QP robustly predicted SC trajectory after PCV7, and the same rationale leads us to predict that serotype 15BC from SC26 will now increase in carriage prevalence following the use of PCV13.

It should be noted that we do not have a full mechanistic account of how selection produces the equilibrium of COG frequencies, which we have used to predict the consequences of vaccination. It is quite conceivable that a minority of COGs or other loci are involved (e.g., polymorphic protein antigens (*26*)), or SNPs in the core genome, which also show a correlation (see Supplementary Information and (*11*)), and the correlations that we have leveraged to predict the impact of vaccination are due to the amount of genetic linkage that persists in the pneumococcus despite appreciable transformation and recombination. Further, as evidenced by two outliers to QP predictions (SC18 and SC25 in Figure 3C), we acknowledge that the model does not currently capture how the ordering of strain invasions may affect the emergence of

genotypes in the years post-PCV or other types of selection acting on pneumococci. Stochasticity resulting from patterns of strain migration will inevitably have an effect. This model is also restricted to the period in which recombination likely plays a limited role in transferring COGs and affecting their frequencies, although previous modeling suggests this period lasts several years (*27*). Ultimately, expanding the model to include immigration of other SCs and disentangling the relative contribution of selection on various loci is likely to be a fruitful area for future research. One area worth exploring is the degree to which recombination acts to maintain COG frequencies on the timescale of population level shifts in lineage composition.

Predicting evolution is a central goal of population genomics especially when related to human health. While evolutionary theory provides an understanding of bacterial population processes including the relative success of lineages, distribution of phenotypes, and ecological niche adaption, these analyses are often conducted retrospectively. Here, we demonstrate a method for predicting the impact of vaccination on the pneumococcal population and make future predictions based on the PCV13-era data. By incorporating information on invasive capacity, these predictions could be extended to inform changes in invasive disease rates. These dynamics may suggest novel vaccine strategies in which one could target not only prevalent serotypes but also those serotypes whose removal would result in a predicted re-equilibration that favors the least virulent or most drug-susceptible lineages. As NFDS appears to be pervasive among bacterial populations, future studies should assess extant pathogen genomic samples for this signal in both the core and accessory genomes.

References

1. B. J. Arnold *et al.*, *Genetics*, in press, doi:10.1534/genetics.117.300662.
2. B. R. Levin, M. Lipsitch, S. Bonhoeffer, *Science (80-.).*, in press (available at http://science.sciencemag.org/content/283/5403/806.abstract).
3. B. Wahl *et al.*, Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000-15. *Lancet. Glob. Heal.* **6**, e744–e757 (2018).
4. D. M. Weinberger, R. Malley, M. Lipsitch, Serotype replacement in disease after pneumococcal vaccination. *Lancet*. **378**, 1962–73 (2011).
5. S. Flasche *et al.*, Effect of pneumococcal conjugate vaccination on serotype-specific carriage and invasive disease in England: a cross-sectional study. *PLoS Med.* **8**, e1001017 (2011).
6. W. P. Hausdorff, W. P. Hanage, Interim results of an ecological experiment—Conjugate vaccination against the pneumococcus and serotype replacement. *Hum. Vaccin. Immunother.* **12**, 358–374 (2016).
7. W. P. Hanage *et al.*, Carried pneumococci in Massachusetts children: the contribution of clonal expansion and serotype switching. *Pediatr. Infect. Dis. J.* **30**, 302–8 (2011).
8. W. P. Hanage *et al.*, Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics*. **2**, 80–84 (2010).
9. T. Azarian *et al.*, The impact of serotype-specific vaccination on phylodynamic parameters of Streptococcus pneumoniae and the pneumococcal pan-genome. *PLOS Pathog.* **14**, e1006966 (2018).
10. U. Obolski, J. Lourenço, S. Gupta, Vaccination can drive an increase in frequencies of antibiotic resistance among non-vaccine serotypes of Streptococcus pneumoniae. *Proc. Natl. Acad. Sci. U. S. A.* (2017), pp. 1–12.
11. J. Corander *et al.*, Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* **1**, 1950–1960 (2017).
12. N. J. Croucher *et al.*, Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–63 (2013).
13. S. Cobey, M. Lipsitch, Niche and Neutral Effects of Acquired Immunity Permit Coexistence of Pneumococcal Serotypes. *Science (80-.).* **335**, 1376–1380 (2012).
14. E. R. Watkins *et al.*, Vaccination Drives Changes in Metabolic and Virulence Profiles of Streptococcus pneumoniae. *PLoS Pathog.* **11**, e1005034 (2015).
15. J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge university press, 1998; http://ebooks.cambridge.org/ref/id/CBO9781139173179).
16. P. D. Taylor, L. B. Jonker, Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978).
17. P. Mitchell *et al.*, Population genomics of pneumococcal carriage in Massachusetts children following PCV-13 introduction. *bioRxiv* (2017) (available at http://biorxiv.org/content/early/2017/12/16/235192.abstract).
18. Y. Li, D. M. Weinberger, C. M. Thompson, K. Trzciński, M. Lipsitch, Surface charge of *Streptococcus pneumoniae* predicts serotype distribution. *Infect. Immun.* **81**, 4519–24 (2013).
19. C. Hyams *et al.*, Effects of Streptococcus pneumoniae strain background on complement resistance. *PLoS One*. **6**, e24581 (2011).

20. P. L. Ho *et al.*, Increase in the nasopharyngeal carriage of non-vaccine serogroup 15 Streptococcus pneumoniae after introduction of children pneumococcal conjugate vaccination in Hong Kong. *Diagn. Microbiol. Infect. Dis.* **81**, 145–148 (2015).
21. S. S. Richter *et al.*, Changes in pneumococcal serotypes and antimicrobial resistance after introduction of the 13-Valent conjugate vaccine in the United States. *Antimicrob. Agents Chemother.* **58**, 6484–6489 (2014).
22. R. Kaur, J. R. Casey, M. E. Pichichero, Emerging streptococcus pneumoniae strains colonizing the nasopharynx in children after 13-valent pneumococcal conjugate vaccination in comparison to the 7-valent era, 2006-2015. *Pediatr. Infect. Dis. J.* **35**, 901–906 (2016).
23. O. X. Cordero, M. F. Polz, Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).
24. S. Cobey, M. Lipsitch, Pathogen Diversity and Hidden Regimes of Apparent Competition. *Am. Nat.* **181**, 12–24 (2013).
25. J. Lourenço *et al.*, Lineage structure of Streptococcus pneumoniae may be driven by immune selection on the groEL heat-shock protein. *Sci. Rep.* **7**, 9023 (2017).
26. T. Azarian *et al.*, Association of pneumococcal protein antigen serology with age and antigenic profile of colonizing isolates. *J. Infect. Dis.* **215** (2017), doi:10.1093/infdis/jiw628.
27. R. Mostowy *et al.*, Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution. *PLoS Genet.* **10**, e1004300 (2014).

Figure Legend

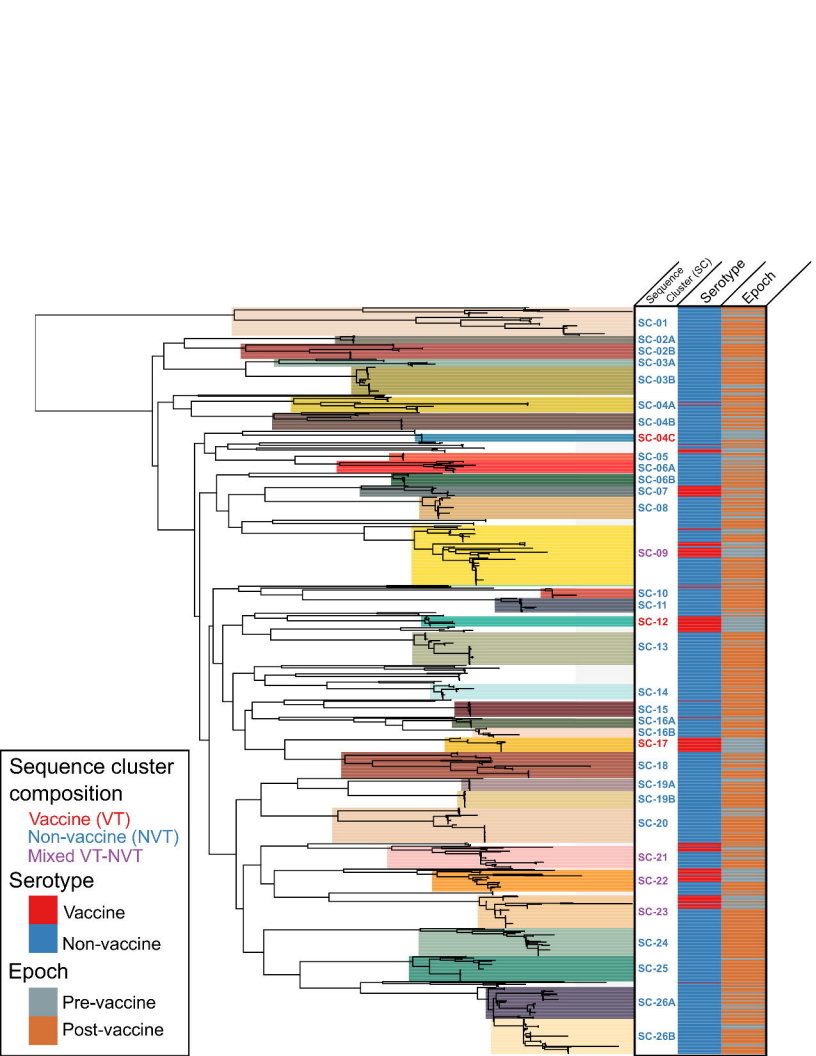
Figure 1. Maximum likelihood phylogeny inferred from a core genome alignment of 937 isolates obtained from a subset of three observational studies of pneumococcal carriage conducted among Native American communities in the southwest US from 1998 to 2012. Thirty-five sequence clusters (SCs), which were identified by subdividing the pneumococcal population on core genome diversity, are shaded by color on the phylogeny and labeled to the right. SCs were assigned based on core genome nucleotide diversity and are numbered such that closer numbers are more genomically similar. The text color of the SC label indicates the serotype composition. The heatmap indicates PCV7 vaccine and non-vaccine serotypes and collection period, illustrating the removal of vaccine serotypes by the introduction of PCV7.

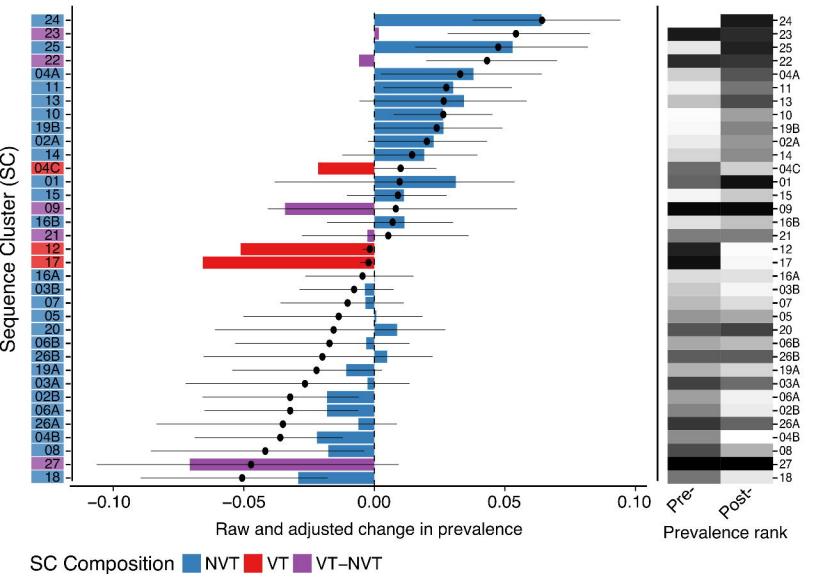
Figure 2. Raw and adjusted change in sequence cluster (SC) prevalence from pre-vaccine (PCV7) to post-vaccine (PCV7). The bars represent the raw change in SC prevalence and are colored based on serotype composition: non-vaccine serotype (NVT) only, vaccine-serotype (VT) only, and mixed VT and NVT (VT-NVT). The point and whisker show the adjusted changes in prevalence and 95% confidence intervals. Adjusted estimates are the raw prevalence change minus the prevalence change expected if all VT had declined by 96.4% (the overall population frequency change) and all NVT had increased by 68.7% (the overall population frequency change) – i.e. in a null model where only the VT/NVT status of a strain determined its fitness. Confidence intervals were obtained from sampling 10,000 bootstraps from pre- and post-vaccine samples. The heatmap to the right of the plot illustrates the prevalence rank before and after the introduction of vaccine. Darker colors represent higher prevalence values. Among the most successful were SCs that contained both PCV7 VT and NVT isolates (SC22 and SC23) whose NVT component included serotypes 6C, 15C, and 35B, as well as SC24 and SC25, which were dominated by the NVT serotypes 23A and 15C, respectively. Compared to SCs comprised of solely NVT isolates, mixed NVT-VT lineages had marginally higher risk differences, indicating greater success than expected under the null model (β=0.03, SE=0.015, $F(1,29)$=3.67, p=0.06). Two SCs that emerged during the study period (SCs 10 and 24) were not included in this analysis as they were not present at the first time point

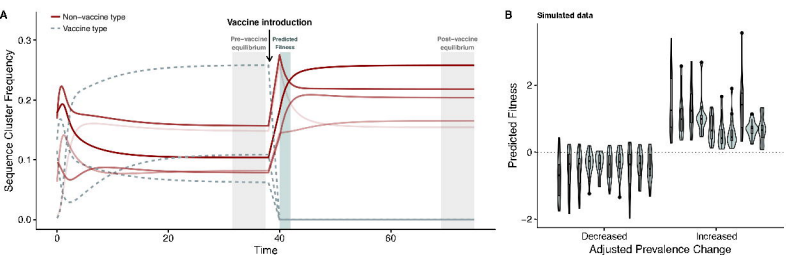
Figure 3. A) Descriptive representation of the Sequence Cluster (SC) prevalence at different stages. We modeled a population of VT and NVT SCs (represented as unique genotypes with alleles 1 or 0 at a locus denoting the presence or absence of a single COG) and simulated the removal of VT genotypes, following the post-vaccine population to equilibrium (details in Supplemental Information). A total of 8 SCs are present at time zero and the system is allowed to evolve until it reaches a steady state ('pre-vaccine equilibrium'). Three SCs where then targeted to mimic a vaccine introduction, which removes them from the system. The predicted fitness was then estimated from the period just after the vaccine introduction, when the population has been depleted of VT but relative prevalence of NVT have not changed – a quantity that can be calculated from pre-vaccine data alone. Finally, the system goes to a second steady state ('post-vaccine equilibrium'). Different shades of red represent the rank of the SC frequencies in the post-vaccine equilibrium. B) Predicted fitness from simulated data. Ten independent replicates were calculated for 2371 COGs, 35 randomly chosen SCs, and 3 vaccine types removed from 35 SCs (consistent with our empirical observations(*9*)). The predicted fitness accurately predicts the direction of the adjusted prevalence change 90.1% of the time (mean of 1000 simulations).

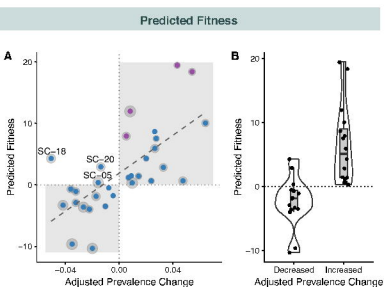
Figure 4. (A) Relationship between predicted fitness of a sequence cluster (SC) and its adjusted prevalence change from pre- to post-vaccine. Predicted fitness was calculated using data solely from the pre-vaccine sample, with the exceptions of SCs 04C, 10, and 22-24 for which there were no non-vaccine serotype (NVT) isolates present in the sample before the introduction of PCV7. For those SCs, data were imputed from the time point during which they were first observed (see supplemental methods). Gray circles around points are scaled to the standard errors of each adjusted prevalence change estimate. The points are colored by SC composition: NVT only (blue) and mixed vaccine serotype (VT) and NVT (purple). The grey shaded quadrants indicate regions of accurate prediction of the prevalence change direction (increased post-vaccine vs. decreased) given the predicted fitness value. Three outlier SCs are annotated. (B) Comparison of predicted fitness between SCs that increased or decreased based on their adjusted prevalence change (p=0.012). (C) Scatterplot of actual versus predicted prevalence of SCs at post-vaccine based on quadratic programming analysis. Points are colored based on SC serotype composition as described in panel A. Accessory loci frequencies pre-vaccine were regressed on the COG frequencies of NVT SCs. We tested the significance of slope $\beta = 1$ and intercept $\alpha = 0$ using Student's t test ($p < .05$). The line of equality (1:1 line), in red, shows the accuracy of the predicted to actual frequencies. Two outliers are annotated. (D) Comparison of the predicted prevalence change from quadratic programming analysis between SCs that increased or decreased based on their actual pre- to post-vaccine prevalence change (p=0.001).

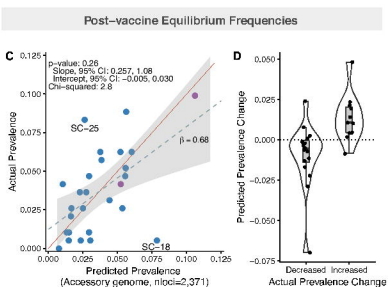
Figure 5. Comparison of naïve and quadratic programming predictions using varying genomic loci. A) Naïve estimate of SC prevalence based on NVT-VT composition and pre-PCV7 SC frequency, B) 2,371 COG predictions present among 5-95% of taxa, C) 17,101 biallelic polymorphic sites found in 1,111 genes in the core genome and present among 5-95% of taxa, D) 5,853 biallelic polymorphic sites found in 272 metabolic genes present in the core genome and present among 5-95% of taxa. For each model, the slope and 95% confidence interval of the actual vs. predicted prevalence values are compared to the 1:1 line using a linear test. P-values >0.05 indicate that the intercept and slope of observed versus predicted post-vaccine SC frequencies did not differ significantly from zero and one (joint model). The χ^2 values are presented in relation to the line of equality (in red). Figures E-H show comparisons of the predicted prevalence change from quadratic programming analysis between SCs that increased or decreased based on their actual pre- to post-vaccine prevalence change. The positive predictive value (PPV, i.e. the probability that a SC predicted to increase post-vaccine truly increased) and negative predictive value (NPV, i.e., the probability that a SC predicted to decrease post-vaccine truly decreased) are also presented for each model.

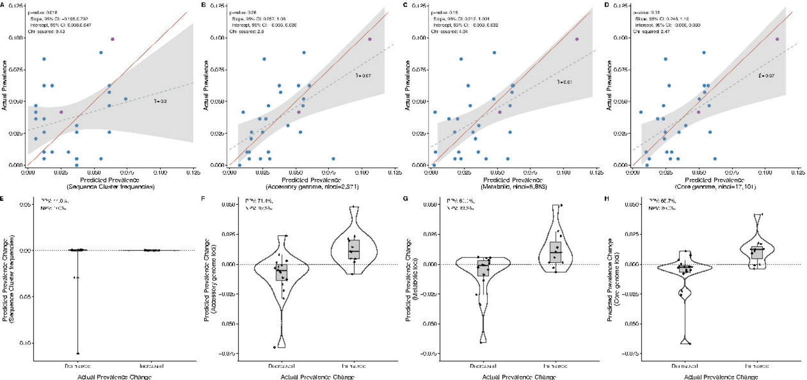

Sequence Cluster (SC)
Serotype
Epoch
SC-01
SC-02A
SC-02B
SC-03A
SC-03B
SC-04A
SC-04B
SC-04C
SC-05
SC-06A
SC-06B
SC-07
SC-08
SC-09
SC-10
SC-11
SC-12
SC-13
SC-14
SC-15
SC-16A
SC-16B
SC-17
SC-18
SC-19A
SC-19B
SC-20
SC-21
SC-22
SC-23
SC-24
SC-25
SC-26A
SC-26B
Sequence cluster composition
Vaccine (VT)
Non-vaccine (NVT)
Mixed VT-NVT
Serotype
Vaccine
Non-vaccine
Epoch
Pre-vaccine
Post-vaccine


Sequence Cluster (SC)
24
23
25
22
04A
11
13
10
19B
02A
14
04C
01
15
09
16B
21
12
17
16A
03B
07
05
20
06B
26B
19A
03A
02B
06A
26A
04B
08
27
18
−0.10
−0.05
0.00
0.05
0.10
Raw and adjusted change in prevalence
Pre-
Post-
Prevalence rank
SC Composition
NVT
VT
VT−NVT


A
Non−vaccine type
Vaccine type
Vaccine introduction
Pre−vaccine equilibrium
Predicted Fitness
Post−vaccine equilibrium
Sequence Cluster Frequency
0.0
0.1
0.2
0.3
0
20
40
60
Time
B
Simulated data
Predicted Fitness
−2
0
2
Decreased
Increased
Adjusted Prevalence Change


Predicted Fitness
A
Predicted Fitness
SC−18
SC−20
SC−05
Adjusted Prevalence Change
B
Predicted Fitness
Decreased
Increased
Adjusted Prevalence Change


Post−vaccine Equilibrium Frequencies
C
p−value: 0.26
Slope, 95% CI: 0.257, 1.08
Intercept, 95% CI: −0.005, 0.030
Chi−squared: 2.8
Actual Prevalence
SC−25
β = 0.68
SC−18
Predicted Prevalence
(Accessory genome, nloci=2,371)
D
Predicted Prevalence Change
Decreased
Increased
Actual Prevalence Change


A
B
C
D
E
F
G
H
Actual Prevalence
Predicted Prevalence
(Sequence Cluster frequencies)
(Accessory genome, nloci=2,371)
(Metabolic, nloci=5,853)
(Core genome, nloci=17,101)
p-value: 0.018
Slope, 95% CI: -0.195,0.792
Intercept, 95% CI: 0.008,0.047
Chi-squared: 9.43
p-value: 0.26
Slope, 95% CI: 0.257, 1.08
Chi-squared: 2.8
p-value: 0.15
Slope, 95% CI: 0.212, 1.001
Chi-squared: 4.04
p-value: 0.31
Slope, 95% CI: 0.248, 1.10
Intercept, 95% CI: -0.006, 0.030
Chi-squared: 2.47
β = 0.67
Predicted Prevalence Change
(Sequence Cluster frequencies)
(Accessory genome loci)
(Metabolic loci)
(Core genome loci)
Actual Prevalence Change
Decreased
Increased