Title:
Predicting evolution using frequency-dependent selection in bacterial populations

Authors:
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[9], Marc Lipsitch[2,10†], William P Hanage[2†]

[†]Co-senior authors

Affiliations:
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 Sanger Institute, 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** World Health Organization, Geneva Switzerland; **10** Department of Immunology and Infectious Diseases, T.H. Chan School of Public Health, Harvard University, Boston MA.

Pamela P Martinez pmartinez@hsph.harvard.edu
Brian J Arnold brianjohnarnold@gmail.com
Lindsay R Grant lgrant10@jhu.edu
Jukka Corander 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 obrienk@who.int
Marc Lipsitch mlipsitc@hsph.harvard.edu
William P Hanage whanage@hsph.harvard.edu

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

One Sentence Summary:

We develop estimates of pneumococcal strain fitness based on the frequencies of accessory genes in a population, and test them using our ability to predict the impact of vaccination.

Abstract:

Predicting how pathogen populations will change over time is challenging. Such has been the case with *Streptococcus pneumoniae*, an important human pathogen, and the pneumococcal conjugate vaccines (PCVs), which target only a fraction of the strains in the population. Here, we use the frequencies of accessory genes to accurately predict changes in the pneumococcal population after vaccination, hypothesizing that these frequencies reflect negative frequency-dependent selection (NFDS) on the gene products. We find that the standardized fitness of a strain estimated by an NFDS-based model at the time the vaccine is introduced accurately predicts the direction of the strain's prevalence change observed after vaccine introduction. Further, we are able to accurately predict the equilibrium post-vaccine population composition and assess the migration and invasion capacity of emerging lineages. 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 driving force.

Main Text:

Human interventions perturb microbial populations in many ways. Most obviously, the use of antibiotics or vaccines that target some strains and not others provide opportunities for new strains to emerge and become established. Examples include vaccines for antigenically diverse human pathogens like influenza, *Neisseria meningitidis*, *Haemophilus influenzae*, *Streptococcus pneumoniae*, and human papillomavirus (*1–3*). Predicting these changes is a central goal of population genomic and evolutionary studies (*4*). However, detailed predictions of how a population will respond to a selective pressure are challenging. Models that specify how mutations with a given fitness change in frequency over time 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) (*5*).

Ongoing efforts to control invasive disease caused by *Streptococcus pneumoniae*, a colonizer of the human nasopharynx and a cause of pneumonia, bacteremia, meningitis, and otitis media, underscore the difficulties of predicting changes after introduction of a vaccine (*6*). Pneumococcal conjugate vaccines (PCVs) target only a fraction of the antigenically diverse *S. pneumoniae* species, which contains over 90 distinct serotypes (*7*). Following widespread introduction of PCVs, non-vaccine serotypes (NVT) benefitted from the removal of their vaccine-serotype (VT) competitors and became more common in carriage and disease, with the gains from reducing VT disease partly offset by increases in NVT disease (*8–10*). These changes in the pathogen population varied by location and were not fully appreciated until retrospective analysis (*11–13*).

Our recent study of pneumococcal carriage isolates collected before and after vaccine introduction in the US Southwest illustrates the complexity in post-vaccine population dynamics, echoing findings from other studies (*13*). Pre-vaccine, the population consisted of multiple 'sequence clusters' which are closely related lineages, defined on the basis of sequence variation in loci present among all isolates (i.e., the core genome) (*14*), and which we henceforth call strains (Figure 1., See supplementary methods for exact details of how we have defined strains and Supplemental Figure 1.). Variation in genome content due to horizontal gene transfer is a

hallmark of prokaryotes; therefore, in addition to the core genome, we can define the accessory genome, as those genes not found in all isolates in the sample. Consistent with their close relatedness, each strain we identify comprises isolates that are fairly homogeneous – but not completely so - in the presence/absence of accessory genes and in phenotypic properties such as serotype and antibiotic resistance (*15*). As VT strains were removed following vaccination, they were replaced by NVT strains, including two NVT strains that had not been found before vaccination, but became common thereafter (*13*). Here, we subsequently find that there was considerable deviation from the null expectation that NVT strains would increase in prevalence proportionately to their pre-vaccine frequency; the most common NVT strains before vaccination were not necessarily the most prevalent 10 years afterwards (Figure 1, see Supplementary Information for details). In particular we find 14 of 35 strains identified among 937 pneumococcal isolates deviated significantly from the prevalence expected under proportional increase; nine increased more than expected and five increased less, annotated with plus and minus signs, respectively, in Figure 1B. The impact of vaccination on individual NVT strains was hence not easily predictable. Consequently, public health authorities and vaccine manufacturers have had to rely on post-vaccine surveillance to estimate the next epidemiologically important lineage and determine subsequent vaccine formulations. At best, this reduces the population impact of vaccination; at worst, it could unintentionally increase the prevalence of virulent or antibiotic resistant lineages (*16*).

A clue to the post-vaccine success of pneumococcal strains may lie in the accessory genome (*17*, *18*). In many bacteria, this can be a large fraction of the total number of genes found in a species (i.e., the 'pan genome') (*19*, *20*). A population genomic study of pneumococci in Massachusetts children found that vaccination had remarkably little effect, after six years, on the overall frequency of the individual accessory genes (defined as clusters of orthologous genes or COGs) (*18*). Despite the fact that nearly half the pre-vaccine population had serotypes targeted by the vaccine, only two of >3000 loci in the accessory genome significantly decreased in frequency, and none increased (*18*). More recently, a geographically diverse sample of pneumococcal genomes showed that while the distribution of strains varied widely across the globe, the proportion of isolates in each sample containing each individual accessory gene was highly

consistent across locations (*17*). Where vaccine was introduced, accessory gene frequencies were restored even after the introduction of vaccine (*13*, *17*).

Negative frequency-dependent selection (NFDS) has been proposed as the mechanism by which the frequencies of loci were restored and maintained after vaccine introduction (*17*). NFDS is a type of balancing selection, which maintains diversity by favoring variants when rare, but exacting a cost when they become common, such that the frequency of the variant stabilizes at intermediate values, or in some instances result in frequency oscillations. Examples include host immunity and bacteriophage predation, and as such, balancing selection is increasingly recognized as a key contributor to population composition (*21*, *22*). Among pneumococci, similar processes, driven by host immunity, have been proposed to explain the co-existence of multiple serotypes (*23*) and vaccine-induced metabolic shifts (*24*). NFDS acting through multiple mechanisms has been used to retrospectively explain changes in strain prevalence following vaccination (*17*). Here, we present flexible, easily computable statistics that estimate the fitness of any strain using the contents of its accessory genome as a proxy for how it will be affected by NFDS, dependent on the frequencies of other strains in the population, and specifically of the accessory genes they carry. Even though we do not know the specific loci under selection or the mechanism involved, through representative population sampling, we are able to make predictions about the composition of a population as well as predict the fitness of any strain in any population, whether or not it has yet appeared in that population. This predictive model offers a way to study population processes and the response to interventions. Further, since our approach does not require a full mechanistic accounting of the underlying evolutionary drivers, it may offer wide application to investigate the phenomenon among other organisms where sufficient sampling of the population is available.

We hypothesized that evolutionary dynamics could be predicted on the premise that after perturbation, strains characterized by accessory genomes that could best restore the pre-perturbation accessory-gene-frequency equilibrium would have the highest fitness and therefore increase in prevalence disproportionately to their competitors. To this end, we implemented a deterministic model using the replicator equation to calculate the fitness of a strain based on its accessory genome, using vaccination as an example of perturbation (*25*, *26*) (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 strain i strain (sequence cluster $SC_i, i = \{1, \ldots, n\}$), n is the total number of strains, and ω_i denotes the predicted fitness of strain i (adapted from Ref. (*17*)), and φ is the average population predicted fitness. In this model, we define ω_i as the scalar product of two vectors whose elements correspond to the accessory genes: the vector $k_{i,l}$ ($l = \{1, \ldots, n_{loci}\}$) with values between 0 and 1 for the frequency of gene l in strain i, and the vector $(e_l - f_l)$ containing the difference between the pre-vaccine frequency e_l of each gene l and f_l, which is the gene's expected frequency post-vaccination, based on removing the VTs from the pre-vaccine population (eq. 2). Intuitively, the vector $(e_l - f_l)$ represents the vacancy that vaccination produces in the population in terms of the accessory loci it removes, and ω_i quantifies the ability of strain i to fill that gap.

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

In contrast with previous more complex models (*17*), we do not define carrying capacity, migration, mutation, or recombination rate, requiring only knowledge of the accessory gene frequencies at equilibrium and which strains they are associated with; these quantities can be estimated from a population survey prior to the perturbation of interest. We assume that the impact of recombination on the accessory genome is negligible over the relatively short time period we study here. Using simulated data, we first assessed the ability of a strain's standardized predicted fitness ($\omega_i - \varphi$) to predict the direction of its change in frequency (Figure 2A), based on its ability to resolve the vaccine-induced perturbation. Note that this predicted fitness uses only data available before vaccine rollout. Using this model, we show that the predicted fitness accurately estimates the direction of a simulated strain's adjusted frequency change (positive predictive value (PPV) = 99.9%, negative predictive value (NPV) = 83.9%, 1000 simulations), independent of the initial pre-vaccine frequency (Figure 2B).

Next, we asked whether this approach could predict the post-vaccine composition of pneumococcal populations, and specifically the relative contribution of each strain to serotype replacement. To test this, we evaluated a pneumococcal sample from the southwest US,

comprised of 937 strains collected before and after the introduction of vaccine. For each strain present before vaccine introduction, we calculated a predicted fitness based on its accessory genome. In a few cases the strains we identified, as described above, contained both VT and NVT serotypes. This is not unexpected, because homologous recombination is known to occasionally transfer the genes that confer serotype into different genomic backgrounds. Where this was the case we removed the VTs and considered the remainder in isolation as an NVT strain. We identified accessory genes as detailed in supplementary materials, finding 2371 loci that were present in between 5% and 95% of isolates. 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 strains increased by the same proportion from their pre-vaccine prevalence (Adjusted R^2=0.44, $p<<0.001$, Figure 3A). The predicted trajectory following vaccination, whether increasing or decreasing in frequency, was accurately predicted for 28 of the 31 tested strains identified in the sample (Figure 3A-B). Strains with a positive adjusted prevalence change had substantially higher predicted fitness than those with a negative one (mean fitness increased vs. decreased 6.4, -2.4; 95% CI of the difference: 5.0-12.5, $p<0.001$, Figure 3B).

While the predicted fitness estimates how successful each strain will be following vaccination, the new equilibrium proportion of the population represented by each strain is of more direct interest for evolution and public health. Thus, we used an optimization technique, quadratic programming, to calculate the NVT strain composition that produced accessory gene frequencies closest to those observed in the previous population, specifically focusing on the 27 strains that were observed pre-vaccine (see Supplementary Information for more details). This method predicted the strain composition of the population following vaccination well i.e., the 95% confidence interval of the observed vs. predicted post-vaccine strain 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 3C). As with our method estimating strain fitness, this approach also accurately predicted which strains would increase in prevalence (PPV=71.4%, NPV=92.3%, Fisher's exact test score = 25.4, $p=0.001$, Figure 3D). In comparison, a naïve estimate based solely on pre-vaccine prevalence performed poorly (Supplemental Figure 5A), as

expected given the discordance in the pre- to post-vaccine rank changes illustrated in Figure 1. 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, though not as accurately as the accessory genome (Supplemental Figure 5). This finding must be considered in the context of recombination, selection, and the evolutionary timescale impacting the pneumococcal genome, which may impact the varying magnitude of NFDS signal across sets of loci. Despite moderate levels of bacterial recombination among pneumococci, there remains appreciable linkage disequilibrium between loci nearby as well as genome-wide (*5*), which makes it difficult to discern the relative selective importance of any particular locus. Exactly which genomic elements are responsible for the predictive ability we document here is unknown but is obviously of interest and should be a focus for future work.

We can retrospectively calculate the predicted fitness of the two strains (shown as SC-10 and SC-24 in Supplemental Figure 1) 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 (*18*, *27*), we found that these two strains had higher predicted fitness than any of the other potential migrant strains that were not present in our southwest US sample before vaccination. Indeed, only two of the strains present before vaccination in the southwest US (SCs 23 and 9) had a higher predicted fitness. This suggests we can use this approach to quantify which strains are most likely to successfully invade a population.

There are two distinct ways in which NVT strains can fill the gap left by vaccination, depending on how closely related they are to the removed VT strains. Where they are very closely related, to the extent that they are considered to be the same strain by analysis of their core genome (which does not necessarily include the genes that determine serotype), they are expected to be almost identical in their accessory genome. There are two examples of this in our dataset (see SC-09 and SC-23 in Supplemental Figure 4) and as we would expect, both were more successful than average following vaccination. We hence expect that for any strain that contains both VT and NVT representatives, the NVT fraction will increase post vaccination. Where such close relatives are not available, the perturbation of the accessory genome may be addressed by NVT strains with varying degrees of relatedness in terms of core genome distance. As shown by the

varied association between core and accessory loci and respective predicted fitness values (Supplemental Figures 3 and 4), strains may be divergent in core genome distance but share similar accessory genomes and consequently comparable predicted fitness. A clear example is the 'invading' strain SC-24 that appeared post-vaccine, which is similar to strain SC-09 in its accessory genome, but sufficiently distinct in its core genome to constitute a different strain (Supplemental Figures 1 and 4).

The ability of this type of balancing selection to determine the strain composition of a population is consistent with findings from environmental microbiology research on multiple bacterial species (*28*). Among pneumococci, changes in population dynamics after the introduction of vaccine have been explained by selection on many different aspects of the organism, including metabolic types, antibiotic resistance, carriage duration, recombination rates, and serotype competition, all of which are likely to be relevant contributors alongside, or components of, the accessory genome (*16*, *24*, *29*, *30*). We provide a simple and effective approach for estimating the fitness of any strain in a population evolving under NFDS acting on accessory loci. All that is required is knowledge of the strain composition of the population and the accessory loci associated with each strain, as this approach does not depend on NFDS acting on particular known biological functions to predict the consequences of vaccination. It is quite conceivable that a minority of loci are involved, including even SNPs in the core genome, which also show a correlation (see Supplemental Figure 5 and (*17*)). We do not wish to imply that the sorts of selection discussed here act alone. Our previous work suggests the interplay between host immunity and polymorphic protein antigens may play a significant role (*31*), and other work suggests an important role for metabolic loci in the core genome (*24*). Phage predation and defense as well as antibiotic resistance all likely contribute to the observed signal (*15*). Certainly, as shown by two outliers to predictions (SC-18 and SC-25 in Figure 2C), we acknowledge that the model does not currently capture all population dynamics. Heterogeneity among loci in strength of NFDS strength could account for some of these discrepancies, as indicated by retrospective model fitting. Nevertheless, given the many potential pressures, mostly not directly observable, that we might expect to structure the pneumococcal population it is notable how effectively this approach can predict the impact of this perturbation. Overall, we find a significant relationship between predicted fitness and the adjusted prevalence change of a

strain. By optimizing the prevalence of each strain conditional on the gene frequencies before vaccination we can estimate the equilibrium population after vaccination, using both the predicted fitness and numerical approximations of the post-vaccine equilibrium. Repeating this analysis on the equilibrated post-vaccine pneumococcal population using the recently introduced vaccine with higher valency (13 compared to seven targeted serotypes), we now predict that strain SC-26, which contains serotype 15B/C, will increase in carriage prevalence (see Supplemental material).

This work suggests numerous potential directions for future work, among them identifying the specific accessory loci or other genomic elements that are responsible for what we observe. Expanding the model to include immigration of other strains 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 gene frequencies on the timescale of population-level shifts in lineage composition. The emergence of new strains, characterized by novel combinations of accessory loci, is expected to be limited by the other strains present in the population in ways that are not presently well understood.

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 perturbing the pneumococcal population that may be useful to predict the outcomes of future interventions including vaccines. 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 those strains whose removal would result in a predicted re-equilibration that favors the least virulent or most drug-susceptible lineages. The pervasive finding of accessory genomes in most bacterial species is usually explained by specialization of lineages to specific niches; however, it could also reflect widespread NFDS, and so future work should seek for evidence of similar signal in the core and accessory genome of other bacteria.

References and Notes:

1. P. Gray *et al.*, Evaluation of HPV type-replacement in unvaccinated and vaccinated adolescent females-Post-hoc analysis of a community-randomized clinical trial (II). *Int. J. cancer*. **142**, 2491–2500 (2018).
2. R. I. Menzies, P. Markey, R. Boyd, A. P. Koehler, P. B. McIntyre, No evidence of increasing *Haemophilus influenzae* non-b infection in Australian Aboriginal children. *Int. J. Circumpolar Health*. **72**, 20992 (2013).
3. C. Hogea, T. Van Effelterre, A. Vyse, Exploring the population-level impact of MenB vaccination via modeling: Potential for serogroup replacement. *Hum. Vaccin. Immunother.* **12**, 451–66 (2016).
4. B. R. Levin, M. Lipsitch, S. Bonhoeffer, *Science (80-.).*, in press (available at http://science.sciencemag.org/content/283/5403/806.abstract).
5. B. J. Arnold *et al.*, *Genetics*, in press, doi:10.1534/genetics.117.300662.
6. 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).
7. S. D. Bentley *et al.*, Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. *PLoS Genet.* **2**, 0262–0269 (2006).
8. D. M. Weinberger, R. Malley, M. Lipsitch, Serotype replacement in disease after pneumococcal vaccination. *Lancet*. **378**, 1962–73 (2011).
9. 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).
10. 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).
11. 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).
12. W. P. Hanage *et al.*, Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics*. **2**, 80–84 (2010).
13. 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).
14. L. Cheng, T. R. Connor, J. Siren, D. M. Aanensen, J. Corander, Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Mol. Biol. Evol.* **30**, 1224–1228 (2013).
15. N. J. Croucher *et al.*, Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* **5**, 1–12 (2014).
16. 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.
17. J. Corander *et al.*, Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* **1**, 1950–1960 (2017).
18. N. J. Croucher *et al.*, Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–63 (2013).
19. H. Tettelin, D. Riley, C. Cattuto, D. Medini, Comparative genomics: the bacterial pan-

genome. *Curr. Opin. Microbiol.* **11**, 472–7 (2008).
20. M. F. Polz, E. J. Alm, W. P. Hanage, Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29**, 170–5 (2013).
21. S. Lehtinen *et al.*, Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc. Natl. Acad. Sci.* **114**, 1075–1080 (2017).
22. O. X. Cordero, M. F. Polz, Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–73 (2014).
23. S. Cobey, M. Lipsitch, Niche and Neutral Effects of Acquired Immunity Permit Coexistence of Pneumococcal Serotypes. *Science (80-.).* **335**, 1376–1380 (2012).
24. E. R. Watkins *et al.*, Vaccination Drives Changes in Metabolic and Virulence Profiles of Streptococcus pneumoniae. *PLoS Pathog.* **11**, e1005034 (2015).
25. J. Hofbauer, K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge university press, 1998; http://ebooks.cambridge.org/ref/id/CBO9781139173179).
26. P. D. Taylor, L. B. Jonker, Evolutionary stable strategies and game dynamics. *Math. Biosci.* **40**, 145–156 (1978).
27. P. K. Mitchell *et al.*, Population genomics of pneumococcal carriage in Massachusetts children following introduction of PCV-13. *Microb. Genomics*. **5** (2019), doi:10.1099/mgen.0.000252.
28. O. X. Cordero, M. F. Polz, Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12** (2014), pp. 263–273.
29. S. Cobey, M. Lipsitch, Pathogen Diversity and Hidden Regimes of Apparent Competition. *Am. Nat.* **181**, 12–24 (2013).
30. 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).
31. 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.

Funding: T.A. and W.P.H. were funded by NIH grant R01 AI106786. T.A., P.P.M. and M.L. were funded by NIH grant R01 AI048935. N.J.C. was funded by a Sir Henry Dale fellowship, and jointly funded by the Wellcome Trust and Royal Society (Grant Number 104169/Z/14/Z). J.C. was funded by European Research Council grant number 742158.

Author contributions: T.A., P.P.M., M.L., W.P.H. conceived the study and designed the model; P.P.M. and T.A. fitted the model; J.C., C.F., and N.J.C. contributed to model evaluation; T.A., P.P.M., and S.D.B. analyzed the genomic data; all authors contributed to the interpretation of the results; T.A. and P.P.M initially drafted the manuscript, with all authors contributing to the final version; L.L.H., R.R., M.S., R.C.W. and K.L.OB. produced data on which the model was based.

Competing interests: M.L. has consulted for Pfizer, Affinivax and Merck and has received grant support not related to this paper from Pfizer and PATH Vaccine Solutions. W.P.H., M.L., and N.J.C. have consulted for Antigen Discovery Inc. The authors have declared that no competing interests exist. K.L.OB. has received grant support for pneumococcal work not related to this paper from Pfizer, GSK, and Gavi. K.L.OB. has consulted for Merck and Sanofi Pasteur. L.R.G, L.L.H., and R.C.W. have received grant support not related to this paper from Pfizer, Merck and GSK.

Data and materials availability: Whole-genome sequencing data are available from NCBI under BioProject PRJEB8327: https://www.ncbi.nlm.nih. gov/bioproject/PRJEB8327. Accession numbers and accompanying metadata have previously been published.

List of Supplementary Materials:

Supplemental Methods

Supplemental Results

Figures S1-S5

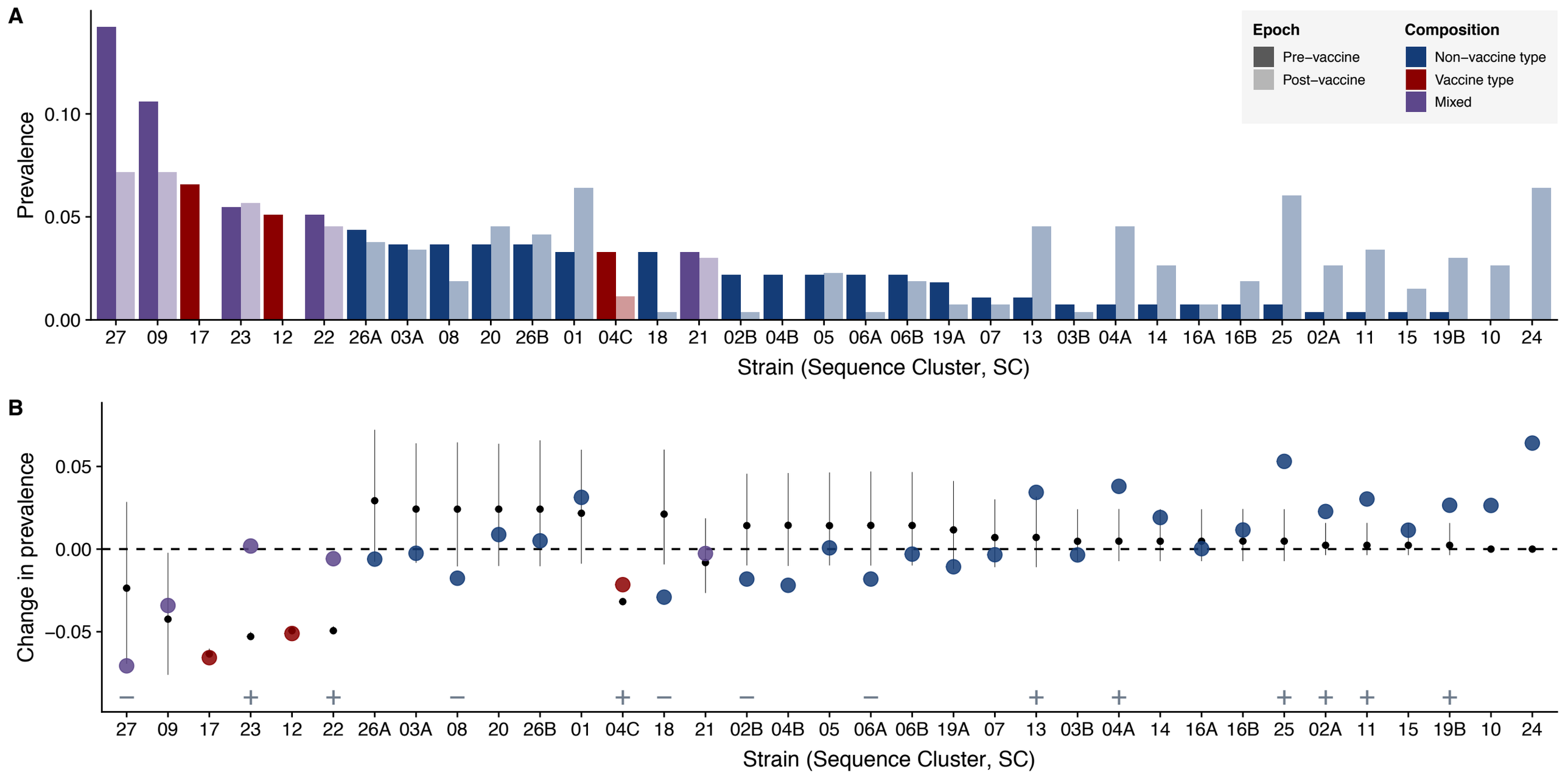
References 1-18

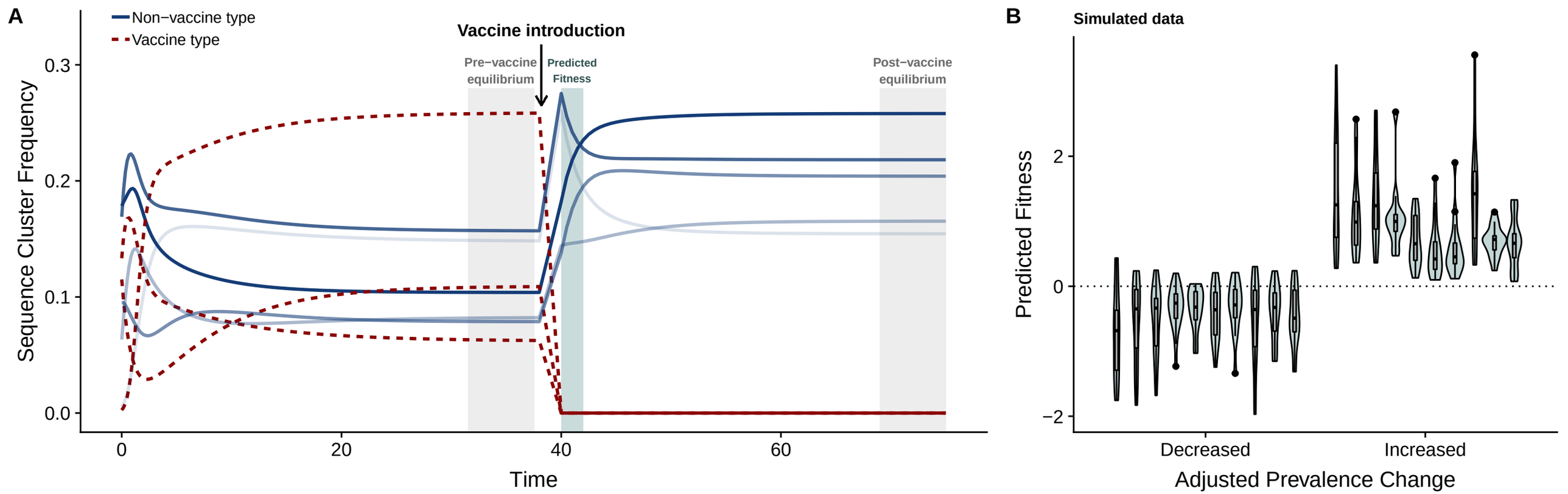
Figures

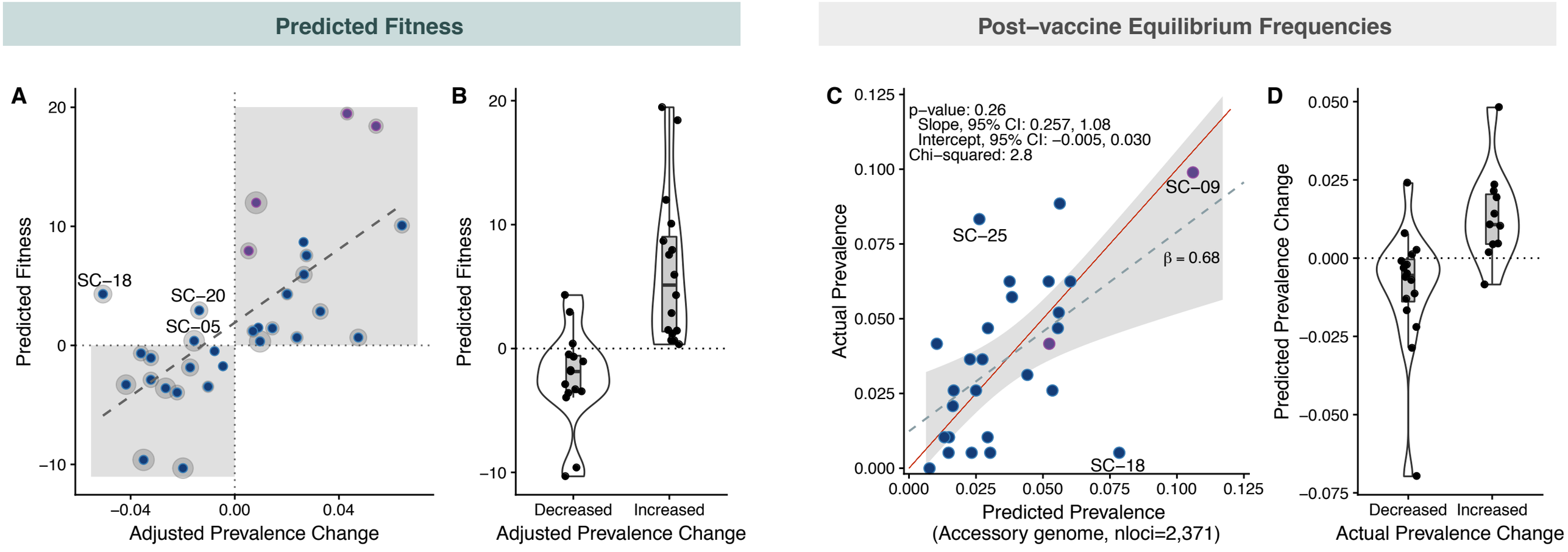
Figure 1. A.) Pre-vaccine to post-vaccine change in prevalence of strains (sequence clusters, SCs). Strains are ordered from highest to lowest pre-vaccine prevalence. B.) Observed change in prevalence from pre-vaccine to post-vaccine ordered by strain as in (A) compared to that expected under a null model (i.e., not using the predictive methods in this paper). Observed changes in prevalence are represented by points colored by serotype composition of the strain: non-vaccine serotype (NVT) only, vaccine-serotype (VT) only, and mixed VT and NVT (VT-NVT). The point and whiskers show 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 of proportional increase where only the VT/NVT status of a strain determined its fitness. The dot is the median, and the whiskers give the 2.5% and 97.5% quantiles of predicted changes under the null model using 10,000 bootstraps from pre-vaccine and post-vaccine samples. Significant differences are denoted with plus and minus signs specifying strains that were significantly more (n=9) or less (n=5) common, respectively, than expected under the null model. Among the most successful were strains that contained both 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. SC27 is polyphyletic, comprised of an aggregate of strains that are at low frequency in the overall population. Compared to strains comprised of solely NVT isolates, those with mixed NVT-VT 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 strains 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 2. A) Descriptive representation of the strain prevalence at different stages relative to vaccine introduction: pre-vaccine equilibrium, vaccine introduction, and post-vaccine introduction. We modeled a population of VT and NVT strains (represented as unique genotypes with alleles 1 or 0 at a locus denoting the presence or absence of a single accessory locus) and simulated the removal of VT genotypes, following the post-vaccine population to equilibrium (details in Supplemental Information). The x-axis represents iterations of the simulation depicting the passage of time. A total of 8 strains are present at time zero and the system is allowed to evolve until it reaches a steady state ('pre-vaccine equilibrium'). Three strains were 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 reaches a second steady state ('post-vaccine equilibrium'). Different shades of red represent the rank of the strain frequencies in the post-vaccine equilibrium. B) Predicted fitness from simulated data. Ten independent replicates were calculated for 2371 accessory loci, 35 randomly chosen strains, and 3 vaccine types (consistent with our empirical observations (*13*)). The predicted fitness accurately predicts the direction of the adjusted prevalence change in 90.1% of cases (mean of 1000 simulations).

Figure 3. (A) Relationship between predicted fitness of 31 strains and their adjusted prevalence change from pre- to post-vaccine. Predicted fitness was calculated using data solely from the pre-vaccine sample, with the exceptions of strains for which there were no non-vaccine serotype (NVT) isolates present in the sample before the introduction of PCV7. For those strains, data were imputed from the time point during which they were first observed (see Supplemental Information). Four strains were excluded either because they were polyphyletic (SC27) or had no NVT isolates present pre- or post-vaccine, and therefore could not be imputed (SC-04C, SC-12, and SC-17). Gray circles around points are scaled to the standard errors of each adjusted estimate. The points are colored by serotype composition of strains: 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 strains are annotated. (B) Comparison of predicted fitness between strains that increased or decreased based on their adjusted prevalence change ($p=0.012$). (C) Scatterplot of actual versus predicted prevalence of 27 strains at post-vaccine equilibrium based on quadratic programming. Points are colored based on serotype composition as described in panel A. Accessory loci frequencies pre-vaccine were regressed on the accessory gene frequencies of NVT strains. 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. The predictions remained significant if SC09 was removed (slope, 95% CI: 0.021, 1.05; intercept, 95% CI: -0.003, 0.03; $p=0.19$, chi-squared=3.5). (D) Comparison of the predicted prevalence change from quadratic programming analysis between strains that increased or decreased based on their actual pre- to post-vaccine prevalence change ($p=0.001$).


A
Epoch
Pre−vaccine
Post−vaccine
Composition
Non−vaccine type
Vaccine type
Mixed
Prevalence
0.10
0.05
0.00
27 09 17 23 12 22 26A 03A 08 20 26B 01 04C 18 21 02B 04B 05 06A 06B 19A 07 13 03B 04A 14 16A 16B 25 02A 11 15 19B 10 24
Strain (Sequence Cluster, SC)
B
Change in prevalence
0.05
0.00
−0.05
27 09 17 23 12 22 26A 03A 08 20 26B 01 04C 18 21 02B 04B 05 06A 06B 19A 07 13 03B 04A 14 16A 16B 25 02A 11 15 19B 10 24
Strain (Sequence Cluster, SC)


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
Post−vaccine Equilibrium Frequencies
A
Predicted Fitness
Adjusted Prevalence Change
SC−18
SC−20
SC−05
B
Predicted Fitness
Decreased
Increased
Adjusted Prevalence Change
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
Predicted Prevalence
(Accessory genome, nloci=2,371)
SC−09
SC−25
SC−18
β = 0.68
D
Predicted Prevalence Change
Decreased
Increased
Actual Prevalence Change