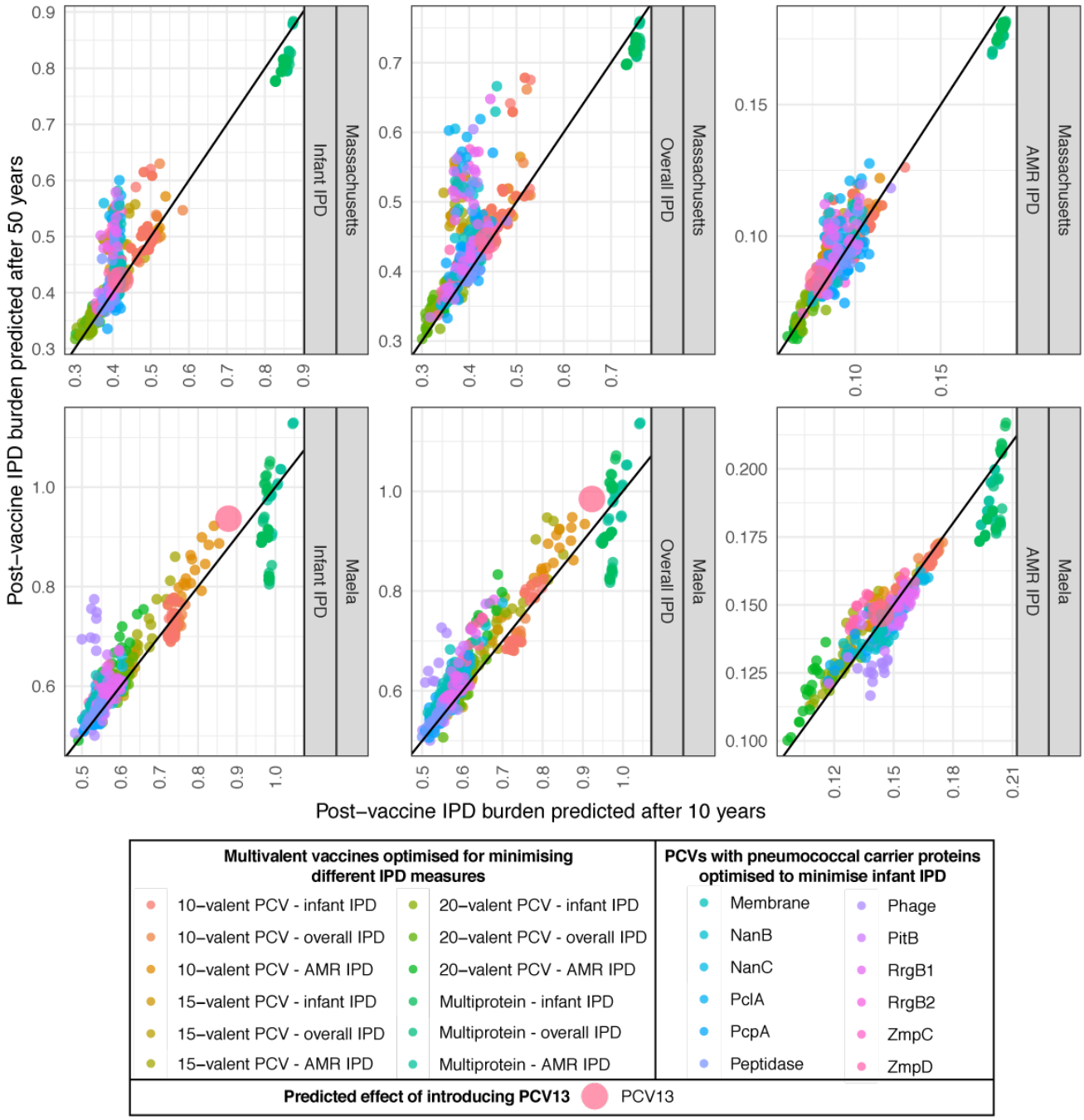
Designing ecologically-optimised pneumococcal vaccines using population genomics

Caroline Colijn, Jukka Corander and Nicholas J. Croucher

Supplementary Materials

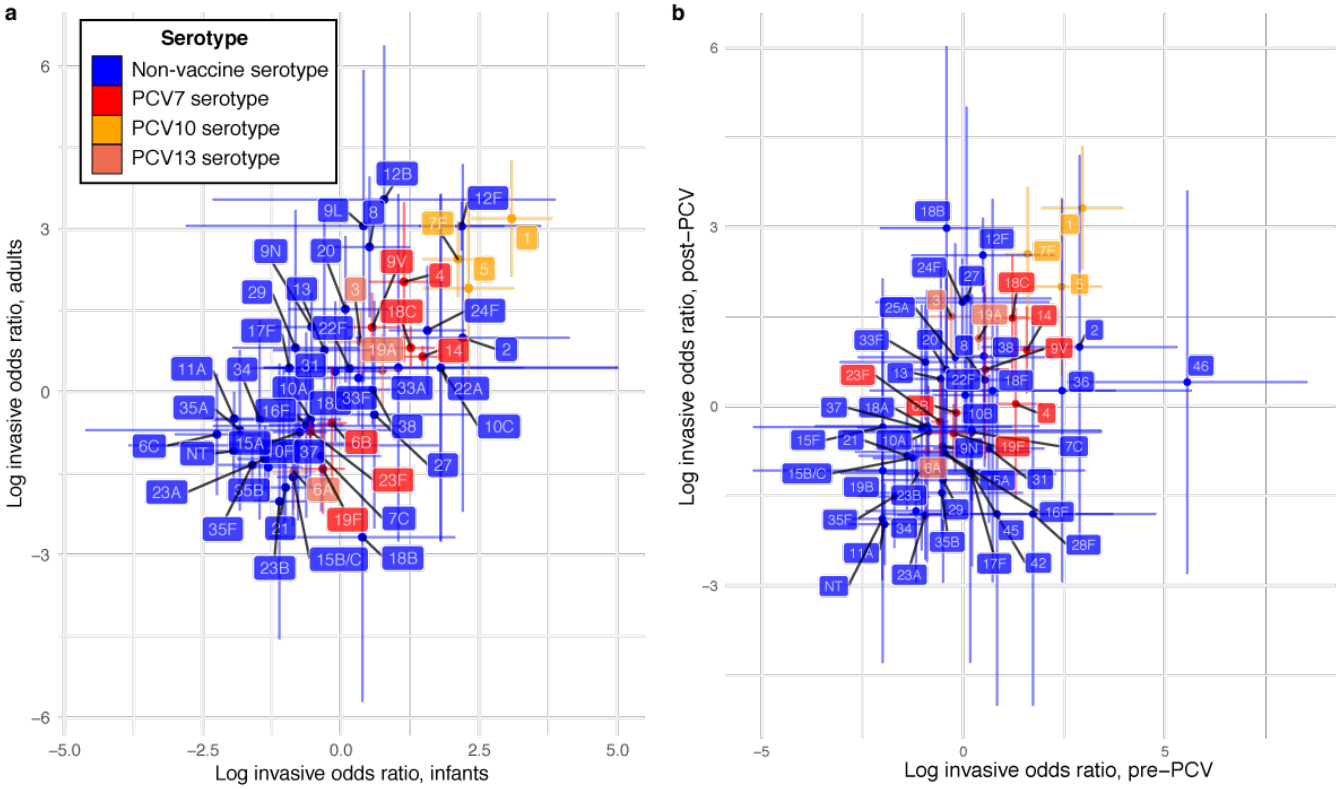
Supplementary Figures 1-11

Supplementary Tables 4-6



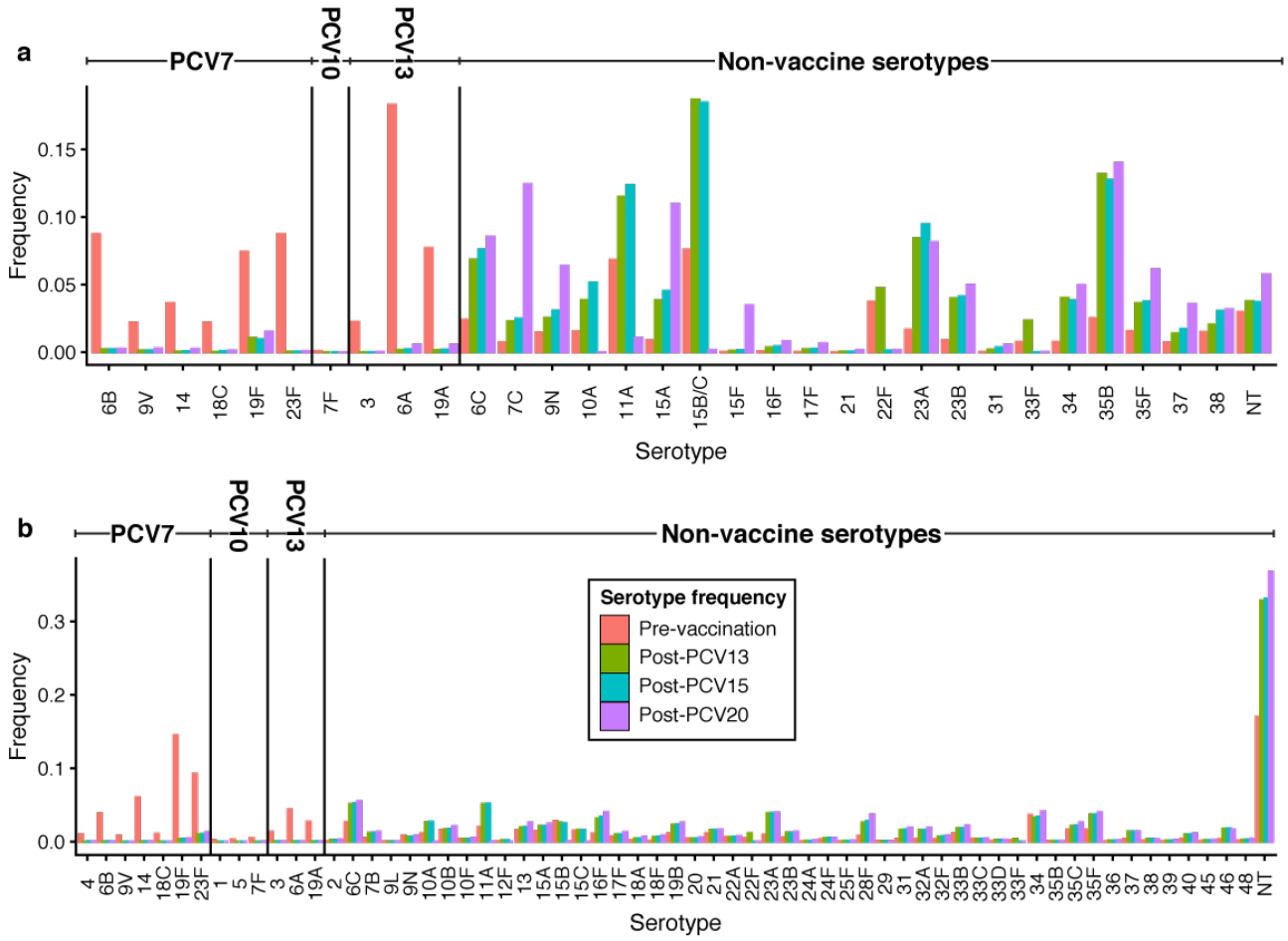
Supplementary Figure 1

These scatterplots compare the IPD burden measures, used for vaccine optimisation, at 10 and 50 years post-vaccination in Massachusetts (n = 480 optimised formulations in each plot) and Maela (n = 440 optimised formulations in each plot). Each plot also displays the simulated effect of introducing PCV13 into a vaccine-naïve population. Plots are separated by population and IPD burden measure. Points are coloured to indicate the constraint on the formulation and the criterion used for optimisation. The line of identity is marked in black. The IPD measures are generally similar at the two timepoints, albeit with some predicted deviations in infant and overall IPD in the Massachusetts population. Hence evaluating the vaccine formulations 10 years after their introduction likely provides a reasonable estimate of their long-term efficacy.



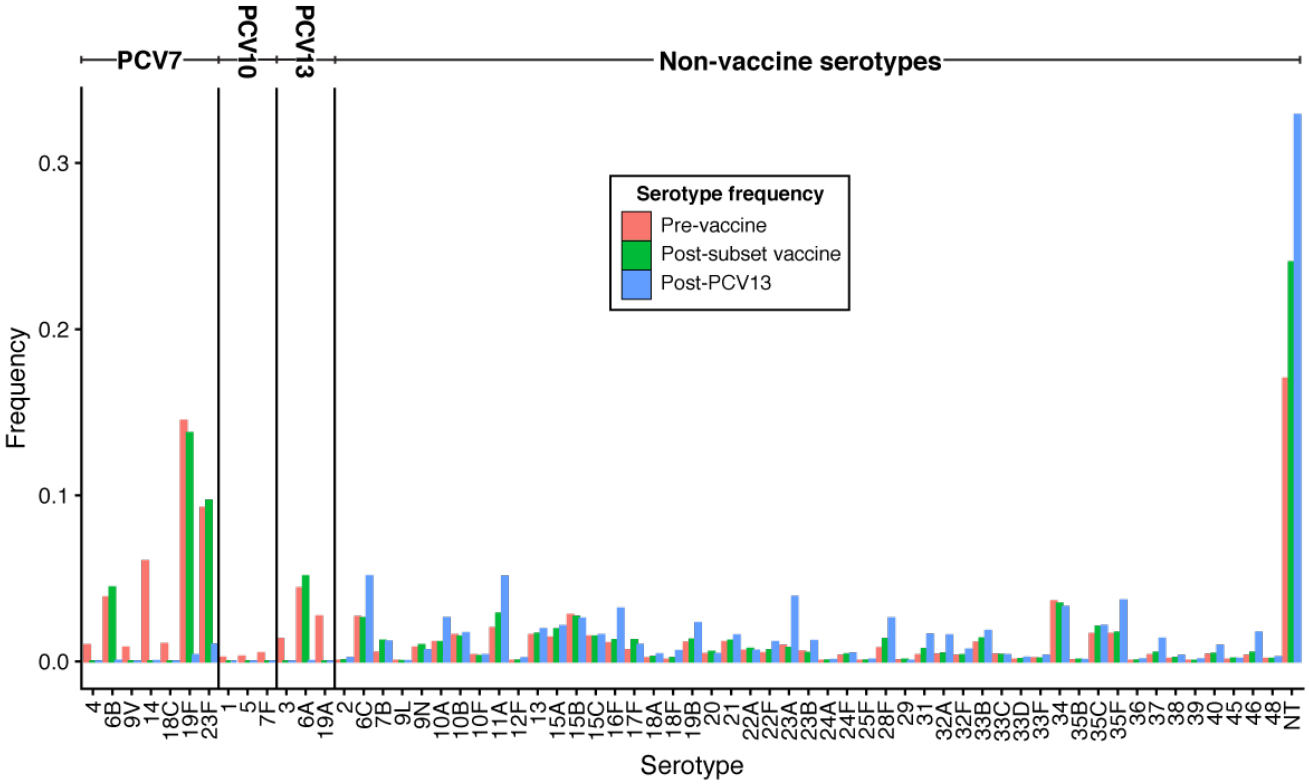
Supplementary Figure 2

Relationships between serotype invasiveness estimates. Each point represents the invasiveness logarithmic odds ratios for a particular serotype, calculated by analysing different datasets with a random effects model. The error bars represent the 95% confidence intervals calculated from the same analyses. **a**, A scatterplot comparing serotype invasiveness in infants and adults (n = 51 serotypes). This shows the same data as in Figure 1, but with all included serotypes labelled. **b**, A scatterplot comparing serotype invasiveness pre- and post-PCV introduction (n = 53 serotypes). This plot shows the estimates of the logarithmic odds ratios of invasiveness from the meta-analysis, split by whether they were estimated using data collected pre- or post-PCV introduction. Considerable variation is evident between the two periods, but the vaccine serotypes' invasiveness did not differ to a notably greater extent than that of the non-vaccine serotypes. This suggests PCVs do not have a substantial effect on the invasiveness of serotypes they target. Therefore the simulations appear justified in associating the same invasiveness with a serotype, regardless of whether it is in the selected PCV formulation or not.



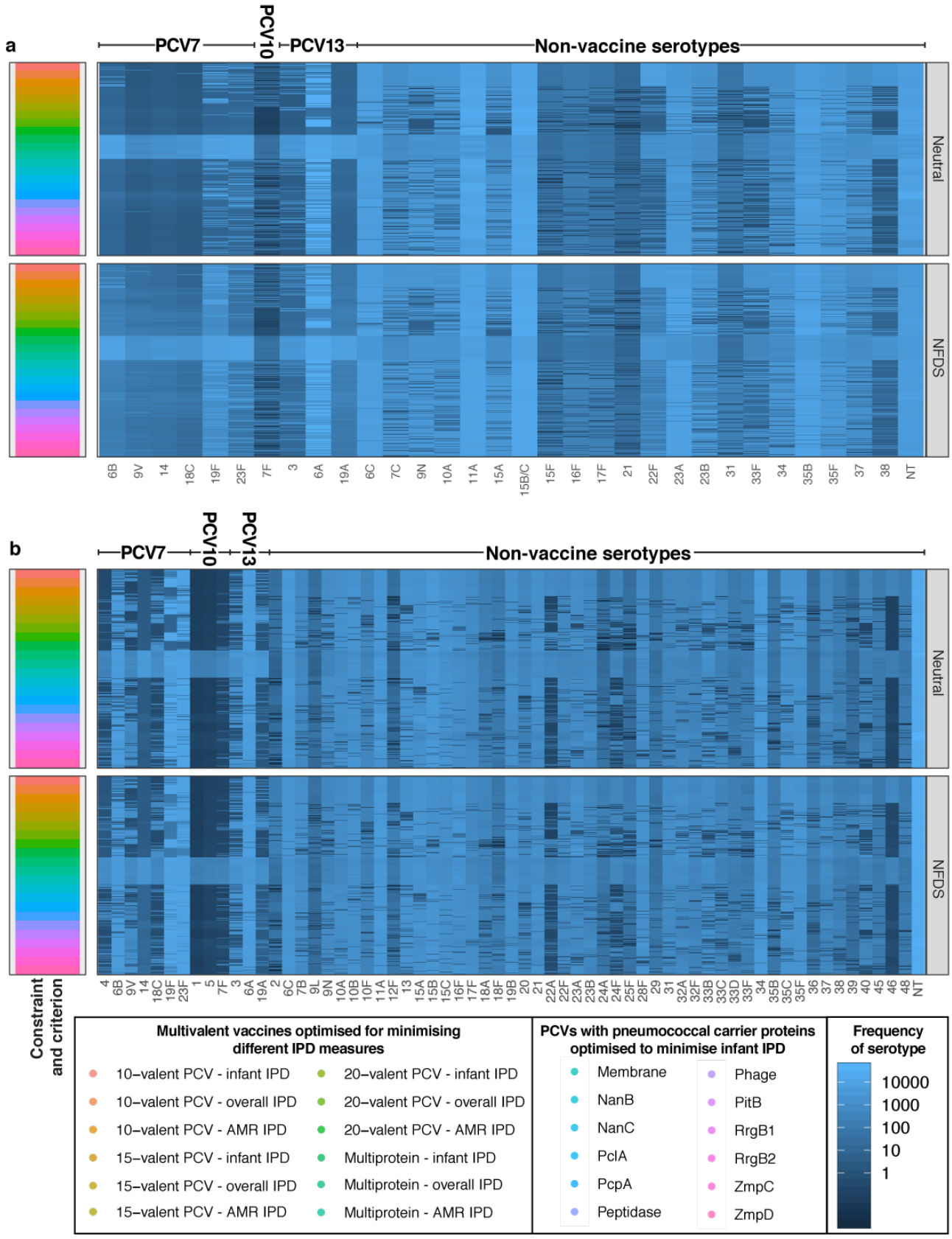
Supplementary Figure 3

Comparison between the simulated impact of the currently-licensed PCV13 formulation, and the forthcoming PCV15 and PCV20 formulations, in **a**, Massachusetts and **b**, Maela. The predicted frequencies of different serotypes 10 years post-vaccine introduction (assuming no prior PCV introduction in each case) are shown, relative to their pre-vaccination frequencies in each location. The predicted performance of these formulations are summarised in terms of the optimisation criteria in Supplementary Table 4.



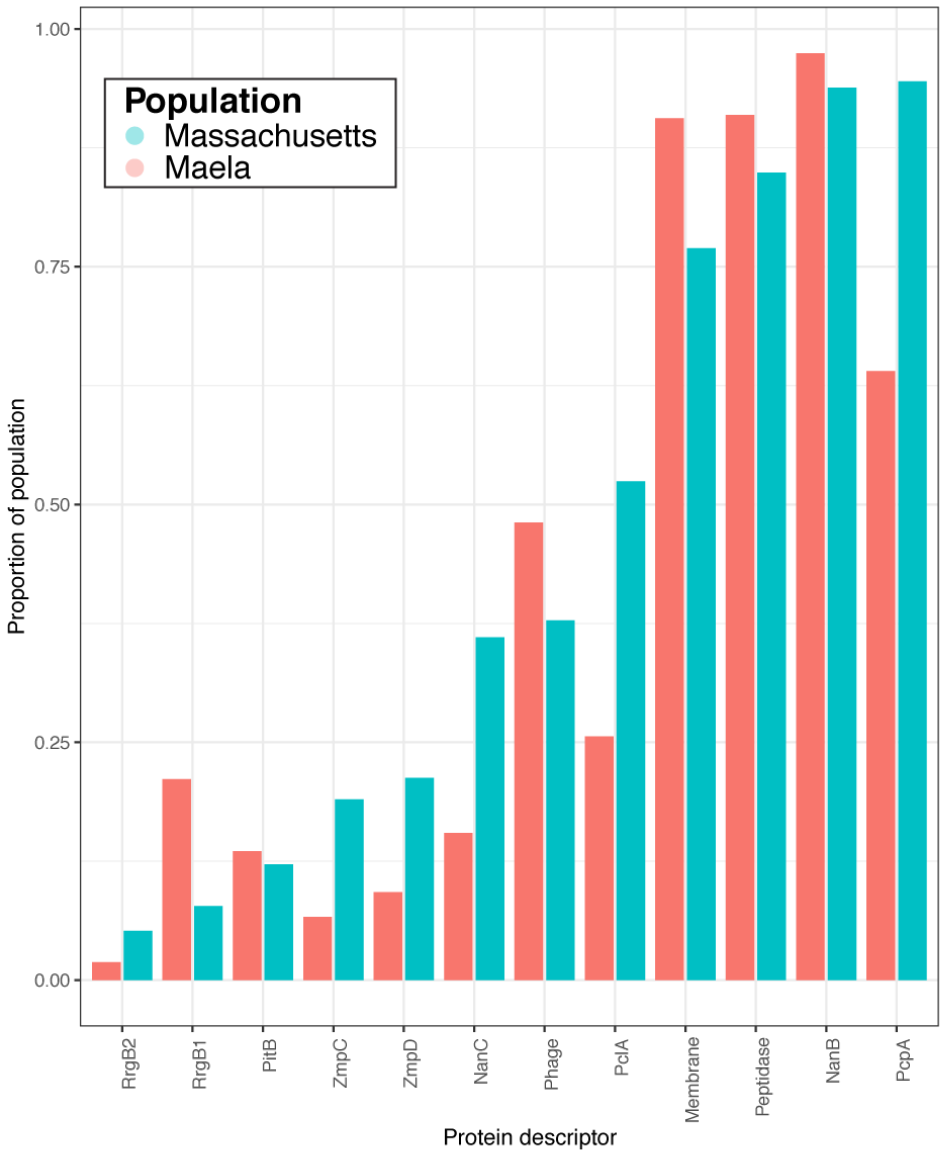
Supplementary Figure 4

Comparison of the pre-vaccination frequencies of serotypes in the Maela population with those predicted 10 years after the introduction of PCV13, or a 9-valent formulation consisting of a 'subset' of PCV13 serotypes (1, 3, 4, 5, 7F, 9V, 14, 18C and 19A). The serotypes are ordered by the licensed vaccine formulations in which they are present. The subset vaccine is forecast to reduce infant IPD to a greater extent, as it retains low invasiveness PCV13 serotypes (6A, 6B, 19F and 23F), and limits their replacement by high-invasiveness serotypes (e.g. 40 and 46).



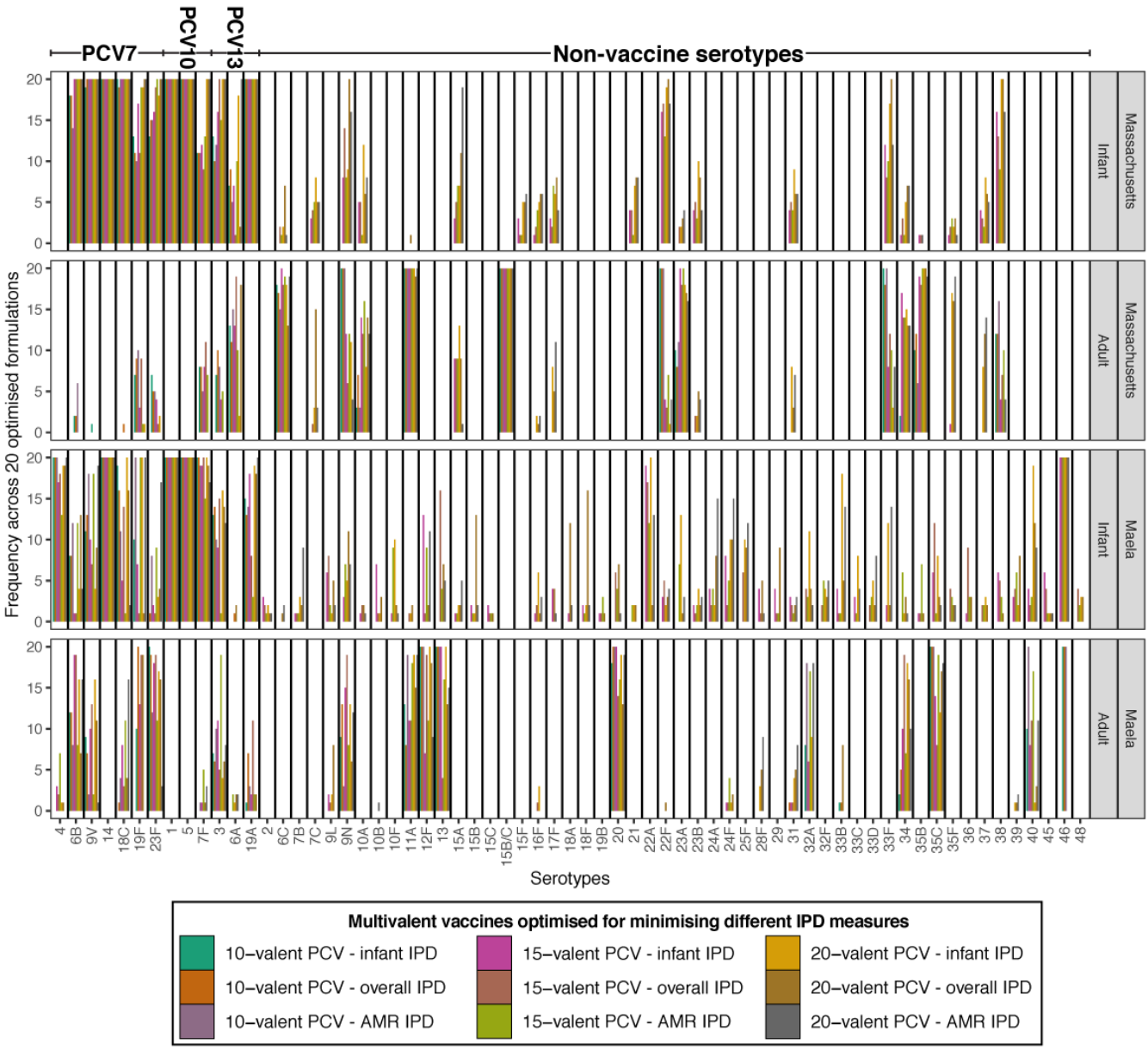
Supplementary Figure 5

Serotype composition of the post-vaccination populations for each of the infant-administered vaccination strategies generated by optimisation. The optimisation criterion and constraint are indicated by the column on the left. The heatmaps show the simulated frequency of each serotype after 10 years of either multi-locus NFDS, or neutral, evolution on a logarithmic scale for **a**, Massachusetts and **b**, Maela, assuming a total population size of 10^5.



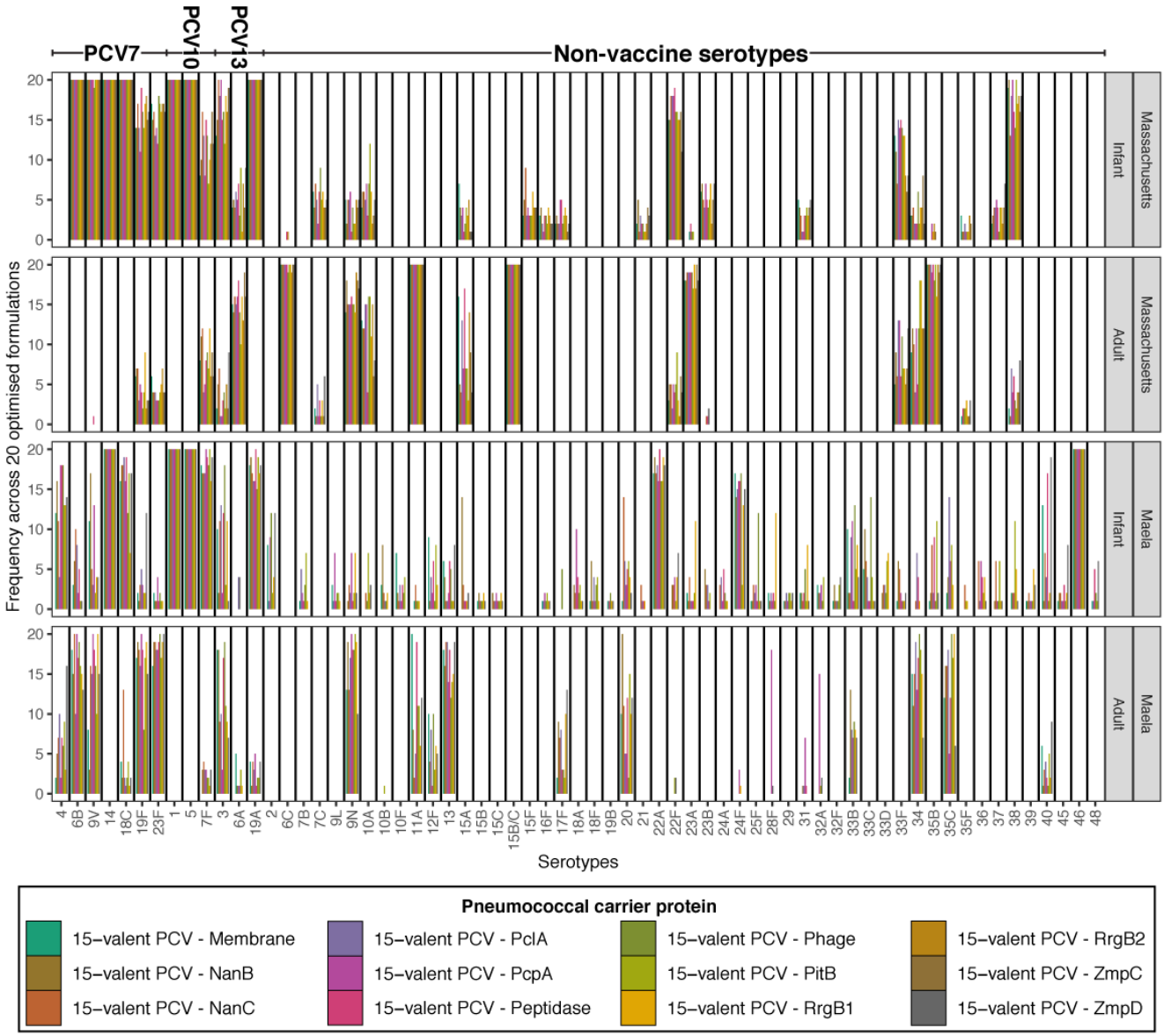
Supplementary Figure 6

Prevalences of the intermediate-frequency protein antigens in the two pneumococcal populations. These show isolates both possessing, and lacking, the antigen co-circulate in the same population. Therefore vaccine-induced immunity against these antigens might facilitate replacement by antigen-negative conspecific competitors.



Supplementary Figure 7

Distribution of capsular antigens, equating to serotypes, between vaccine formulations. Bar charts show the frequency of each capsular antigen in the 20 analysed formulations for the combinations of optimisation criterion and constraint listed in the key, which relates these conditions to the bar colours. Panels split the formulations by population (Massachusetts or Maela) and whether they were designed for infant administration through optimisation, or complementary vaccines designed for adult administration.

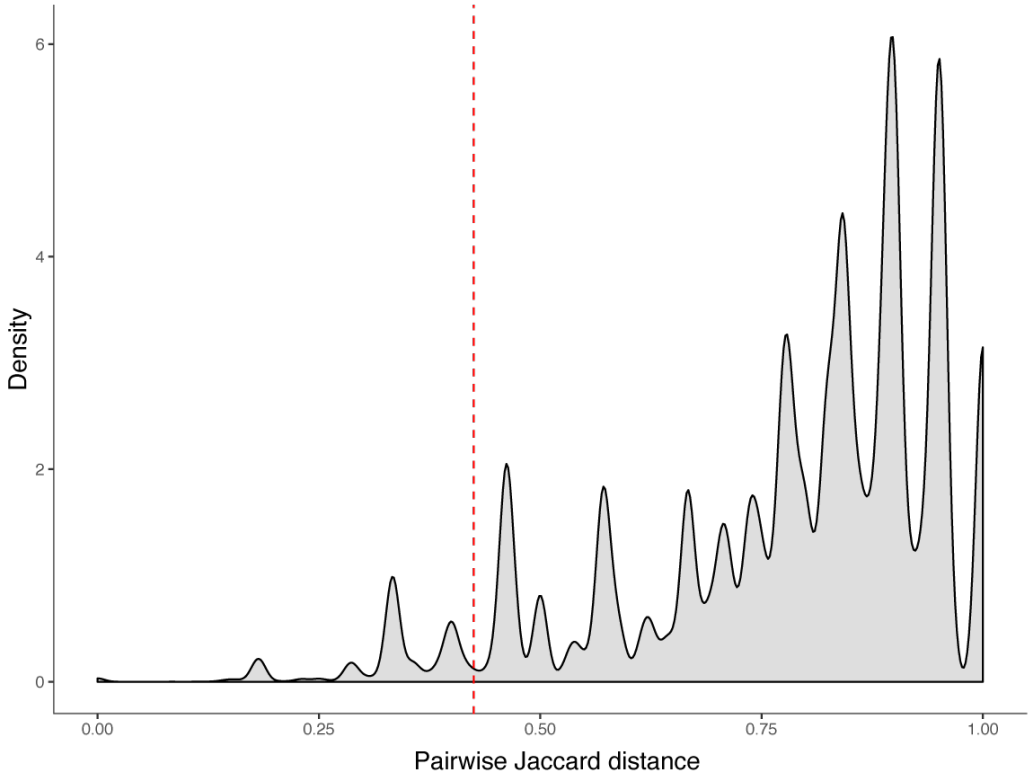


Supplementary Figure 8

Distribution of capsular antigens between vaccine formulations designed to feature pneumococcal carrier proteins. The bar charts show the frequency of each capsule type in the 20 optimised formulations for each pneumococcal carrier protein, as indicated by the bar colour. Panels split the formulations by population (Massachusetts or Maela) and whether they were designed for infant administration through optimisation, or complementary vaccines designed for adult administration.

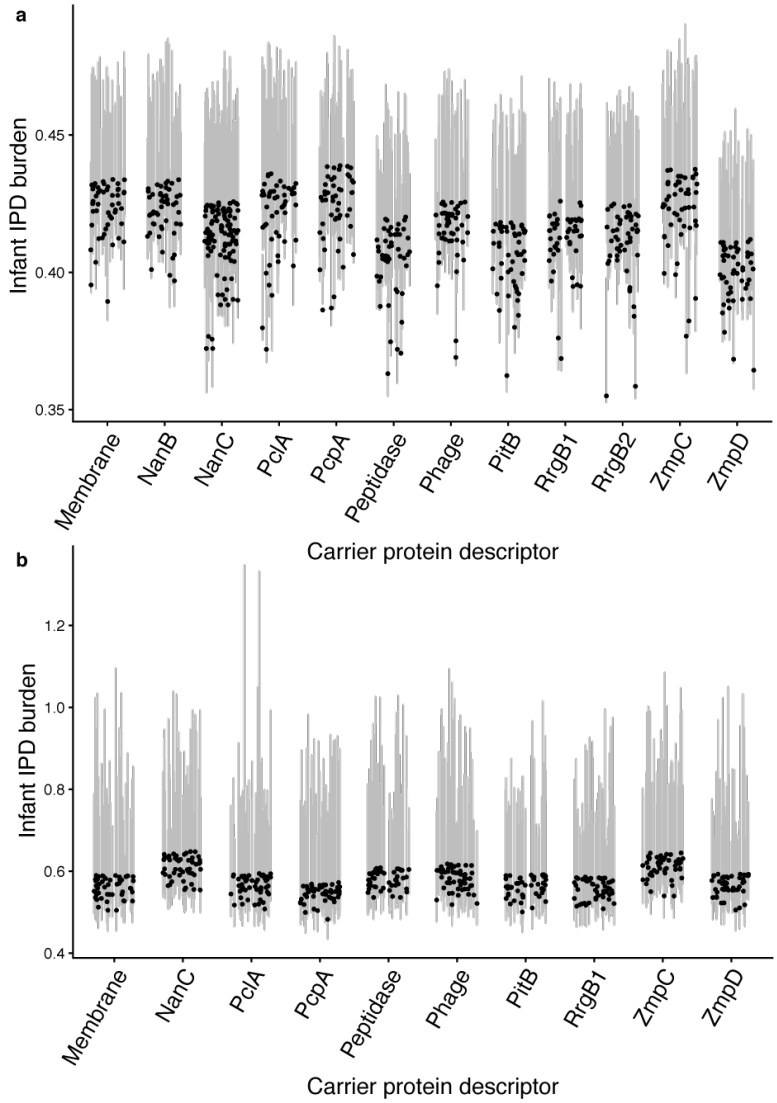
Supplementary Figure 9

Frequency of resistance loci within each serotype across the Massachusetts and Maela populations. To aid comparisons between the populations, the interchanging serotypes 15B and 15C are combined as serotype 15B/C in the Maela data. Bars are doubled in width if the corresponding serotype was only detected in one population.



Supplementary Figure 10

Density plot representing the distribution of pairwise Jaccard distances between PCV formulations generated by optimisation, calculated from the extent to which they shared capsular antigens. The vertical red dashed line shows the threshold similarity (0.425) used to define edges in the network displayed in Figure 6. This corresponds to a local minimum in the density plot that defines linked formulations as having a pairwise similarity in the highest 5.15% of the overall pairwise distance distribution.



Supplementary Figure 11

Variation in estimated IPD burden with resampling of serotypes' invasiveness from the distributions defined by the meta-analyses. The points represent the infant IPD burden predicted 10 years post-introduction for 15-valent PCVs containing a pneumococcal carrier protein in **a**, Massachusetts and **b**, Maela (n = 50 for each carrier protein in each population). These were calculated using the point estimates of serotype invasiveness from the meta-analysis summarised in Figure 1. The grey vertical bars quantify the uncertainty in the predicted post-vaccine infant IPD burden as the inter-quartile ranges calculated from 100 analyses with the same formulation. Each analysis independently resampled serotypes' invasiveness logarithmic odds ratio from a Gaussian distribution defined by the 95% confidence intervals calculated from the epidemiological meta-analysis. Consequently, the vertical bars are positively skewed relative to the point estimates, as the IPD burdens are calculated using non-logarithmic odds ratios. The uncertainty is greatest for serotypes rarely detected in epidemiological studies, with the consequence that the Maela estimates are associated with much greater uncertainty than the Massachusetts estimates.

Supplementary Table 4: Comparison of alternative approaches to rational vaccine design.

Three heuristics were used to design 15-valent formulations based on the pre-vaccine populations in both Massachusetts and Maela. These all had to include serotypes 1, 5 and 14, as for those identified through optimisation. The 'invasiveness' heuristic selected the most invasive serotypes in the pre-vaccine population; the 'virulence' heuristic selected the serotypes with the greatest product of invasiveness and pre-vaccine prevalence (i.e., those expected to be most prevalent in pre-vaccine IPD); the 'Nurhonen & Auranen' method was published previously, and was run assuming complete replacement of vaccine serotypes by non-vaccine serotypes in carriage. For all three heuristics in both populations, formulations were designed for both purely serotype-defined optimisation criteria (infant or overall IPD). For each of these formulations, the three optimisation criteria were calculated 10 years post-vaccine introduction using the multi-locus NFDS model. These approaches were compared to the predicted impact of licensed formulations (all analysed as if introduced into the pre-PCV7 population in each location), and the best-performing 15-valent formulations identified by this analysis when optimising for different criteria.

Vaccine design strategy	Formulation	Infant IPD	Overall IPD	AMR IPD
Massachusetts				
PCV10	1, 4, 5, 6B, 7F, 9V, 14, 18C, 19F, 23F	0.68	0.64	0.20
PCV13	1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 23F	0.42	0.44	0.078
PCV15	1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 22F, 23F, 33F	0.38	0.39	0.07
PCV20	1, 3, 4, 5, 6A, 6B, 7F, 8, 9V, 10A, 11A, 12F, 14, 15B/C, 18C, 19A, 19F, 22F, 23F, 33F	0.45	0.47	0.11
Infant virulence	1, 3, 5, 6A, 6B, 9V, 14, 15B/C, 18C, 19A, 19F, 22F, 23F, 33F, 38	0.45	0.51	0.10
Overall virulence	1, 3, 5, 6A, 6B, 9N, 9V, 11A, 14, 18C, 19A, 19F, 22F, 23F, 38	0.52	0.52	0.10
Infant invasiveness	1, 3, 5, 6B, 7C, 7F, 9V, 14, 18C, 19A, 19F, 22F, 31, 33F, 38	0.38	0.38	0.097
Overall invasiveness	1, 3, 5, 6B, 7F, 9N, 9V, 14, 17F, 18C, 19A, 22F, 31, 33F, 38	0.41	0.34	0.11
Nurhonen & Auranen (minimising infant IPD)	1, 3, 5, 7C, 7F, 9V, 14, 15F, 16F, 18C, 19A, 22F, 31, 33F, 38	0.46	0.42	0.12
Nurhonen & Auranen (minimising overall IPD)	1, 3, 5, 7F, 9N, 9V, 14, 16F, 17F, 18C, 19A, 22F, 31, 33F, 38	0.46	0.38	0.12
15-valent PCV (optimised to minimise infant IPD)	1, 5, 6A, 6B, 7F, 9V, 14, 17F, 18C, 19A, 19F, 22F, 23F, 33F, 38	0.37	0.42	0.07
15-valent PCV (optimised to minimise overall IPD)	1, 3, 5, 6B, 7C, 7F, 9N, 9V, 14, 18C, 19A, 22F, 23B, 23F, 38	0.40	0.35	0.09
15-valent PCV (optimised to minimise AMR IPD)	1, 3, 5, 6A, 6B, 7F, 9V, 14, 15A, 18C, 19A, 19F, 22F, 23F	0.40	0.41	0.070
Maela				
PCV10	1, 4, 5, 6B, 7F, 9V, 14, 18C, 19F, 23F	0.79	0.82	0.14
PCV13	1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 23F	0.88	0.93	0.14
PCV15	1, 3, 4, 5, 6A, 6B, 7F, 9V, 14, 18C, 19A, 19F, 22F, 23F, 33F	0.88	0.92	0.14
PCV20	1, 3, 4, 5, 6A, 6B, 7F, 8, 9V, 10A, 11A, 12F, 14, 15B/C, 18C, 19A, 19F, 22F, 23F, 33F	0.88	0.92	0.15
Infant virulence	1, 3, 4, 5, 6A, 6B, 7F, 14, 18C, 19A, 19F, 22A, 23F, 35C, 46	0.53	0.62	0.11
Overall virulence	1, 3, 4, 5, 6B, 7F, 9V, 13, 14, 18C, 19A, 19F, 23F, 35C, 46	0.53	0.54	0.096

Infant invasiveness	1, 2, 4, 5, 7F, 12F, 14, 18C, 19A, 22A, 24F, 36, 39, 40, 46	0.52	0.53	0.15
Overall invasiveness	1, 2, 4, 5, 7F, 9L, 12F, 14, 18C, 22A, 24F, 36, 39, 40, 46	0.56	0.56	0.16
Nurhonen & Auranen (minimising infant IPD)	1, 2, 4, 5, 7F, 12F, 14, 18C, 22A, 24F, 25F, 36, 39, 40, 46	0.56	0.56	0.16
Nurhonen & Auranen (minimising overall IPD)	1, 2, 4, 5, 7F, 9L, 12F, 14, 18C, 22A, 24F, 36, 39, 40, 46	0.56	0.56	0.16
15-valent PCV (optimised to minimise infant IPD)	1, 3, 4, 5, 7F, 9V, 14, 18C, 19A, 22A, 24F, 33B, 35C, 40, 46	0.48	0.48	0.14
15-valent PCV (optimised to minimise overall IPD)	1, 3, 4, 5, 7F, 9V, 13, 14, 18C, 19A, 22A, 24F, 35C, 40, 46	0.49	0.48	0.14
15-valent PCV (optimised to minimise AMR IPD)	1, 4, 5, 6B, 7F, 9N, 9V, 10F, 12F, 14, 19F, 22A, 23A, 23F, 46	0.62	0.65	0.11

Supplementary Table 5. Characteristics of the intermediate-frequency *S. pneumoniae* protein antigens

Each protein antigen is listed by its descriptor and the corresponding cluster of orthologous genes in Corander *et al* [1] and Croucher *et al* [2]; the sequences of all proteins in the latter study are available from https://doi.org/10.5061/dryad.t55gq. With the exception of the pilus proteins RrgB1 and RrgB2, these antigens were identified as antibody-binding targets using a panproteome array. The relative strength of vaccine-induced immunity, *a*, was the factor by which the immunity induced by protein-polysaccharide conjugates was multiplied to predict the effects of including these antigens in a vaccine.

Descriptor	Cluster of orthologous genes in Croucher *et al*	Cluster of orthologous genes in Corander *et al*	Function	Relative strength of vaccine-induced immunity, *a*
NanB	CLS01445	CLS00257	neuraminidase B	0.11235
ZmpD	CLS02608	CLS00476	zinc metalloprotease D variant	0.21889
PclA	CLS03178	CLS00440	pneumococcal collagen-like protein A variant	0.12243
RrgB1	CLS02942	CLS02709	type I pilus rrgB (clade 1) structural protein	0.016127
PitB	CLS02871	CLS01706	type 2 pilus structural protein PitB	0.19083
RrgB2	CLS02796	CLS03842	type I pilus rrgB (clade 2) structural protein	0.063344
Phage	CLS01887	CLS00695	Prophage protein	0.1393
Membrane	CLS00011	CLS01683	Membrane protein of unknown function	0.14118
ZmpC	CLS01991	CLS04319	zinc metalloprotease C	0.1832
NanC	CLS01160	CLS03670	neuraminidase C	0.15235
Peptidase	CLS01541	CLS01895	M50 peptidase family protein	0.11347
PcpA	CLS01852	CLS01587	choline binding protein PcpA	0.21002

Supplementary Table 6. Common features of optimized vaccine formulations

For each of the populations (Massachusetts and Maela), these descriptions define the common features of the optimised formulations identified when minimising the burden of infant, overall or AMR IPD through infant vaccination. Analogous definitions were also included for the complementary adult vaccines identified in each population. Each description was identified through logic regression against a random set of formulations, followed by manual curation to generate more intuitive descriptions.

Vaccinee demographic and region	Common features of formulations
Massachusetts infants	Contains a core of 1, 5, 18C, 14, and 19A; plus at least one of 6B or 9V; plus at least three of 19F, 6A, 23F, 3, 38, 7F, 33F, 22F
Massachusetts adults	Contains a core of 11A, 15B/C; plus one of 23A, 6C, 9N or 10A; plus one of 35B, 6A, 33F
Maela infants	Contains a core of 1, 14, 46 and 5; plus four of 24F, 22A, 40, 4, 10F, 7F, 19A, 18C, 9L, 19F, 35C, 3, 33C, 9V, 23B, 15A, 15B, 36, 32A, 45, 15A, 16F OR Contains a core of 1, 14, 4, 5; plus one of 18C, 19F, 7F, 9V, 19A, 6B, 3
Maela adults	One of 24A, 21, 40, 13, 45; plus four of 23F, 13, 9N, 19F, 35C, 6B, 20, 3, 9V, 34