Figures and figure supplements

Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration

John A Lees et al
Figure 1. Swabbing and sequencing study design. We start with serotype swab data on 598 children from two cohorts, taken every month after birth for two years. For all samples we fitted the transition and emission probabilities of a continuous time hidden Markov model for each serotype. Then, for each child, we used these parameters were then used to infer the most likely carriage durations. We matched carriage episodes with resistance and genomic data for 2157 episodes to draw conclusions on the basis of variation in this epidemiological parameter.

DOI: https://doi.org/10.7554/eLife.26255.003
Figure 2. Hidden Markov models of swab time series, and their goodness-of-fit. We fitted three different models to the processed time-series data with states, allowed transitions and emissions as shown. We refitted each model allowing the transitions probabilities to covary with the age of the child and whether the child had carried pneumococcus previously. For the converged model the Akaike information criterion (AIC) is shown for the original fit, and when including these covariates (AICcovar).

DOI: https://doi.org/10.7554/eLife.26255.004
Figure 3. Distribution of carriage duration, and effect of monotonic transformation. Panel (a) shows a histogram of the inferred carriage duration, (b) shows this result after the natural logarithm is taken, and (c) after the warping function is applied. DOI: https://doi.org/10.7554/eLife.26255.006
Figure 3—figure supplement 1. Regression diagnostics and outlier removal. Panel (a) shows prior to outlier removal, (b) after outlier removal as produced by plot.lm() in R. Points deviating from normal residuals (top right plot), and at high leverage (bottom right plot) were removed. These observations appeared to be due to swabs not taken at the prescribed monthly intervals.

DOI: https://doi.org/10.7554/eLife.26255.007
Figure 3—figure supplement 2. Monotonic warping function from warped-lmm. x-axis shows the centred and normalised input phenotype; y-axis shows corresponding warped value.

DOI: https://doi.org/10.7554/eLife.26255.008
Figure 3—figure supplement 3. Normal quantile-quantile plot of carriage length, and effect of monotonic transformation. Panel (a) the inferred carriage duration, (b) after the natural logarithm is taken, and (c) after the warping function is applied.
DOI: https://doi.org/10.7554/eLife.26255.009
Figure 4. Mapping of carriage duration onto phylogeny. Using the carriage duration as a continuous trait, the ancestral state at every node of the rooted phylogeny was reconstructed. Red branches are carriage for a short time, blue for a long time. Clusters identified in previous analysis have been labelled.

DOI: https://doi.org/10.7554/eLife.26255.012
Figure 4—figure supplement 1. Mapping of warped carriage duration onto phylogeny. As Figure 4, but using the warped carriage duration. Blue branches are carriage for a short time, yellow for a long time.
DOI: https://doi.org/10.7554/eLife.26255.013
Genetic distances between isolates

Figure 4—figure supplement 2. Histogram of pairwise patristic distances on the inferred phylogeny. A cut-off for heritability estimation was chosen at 0.04, under which a clear second maxima corresponds to closely related isolates on the tree.

DOI: https://doi.org/10.7554/eLife.26255.014
Figure 4—figure supplement 3. Change in carriage duration associated with capsule switching events. For each of the three events analysed the subtree containing the switch is shown on the left. For each isolate within the subtree, carriage duration (on a roughly exponential scale), warped carriage duration (on a roughly linear scale) and serotype are shown as coloured bars aligned with the tip.

DOI: https://doi.org/10.7554/eLife.26255.015
Figure 4-figure supplement 4. Lasso regression plots for lineage effects. Panel a) shows the value of each predictor on the y-axis for different values of the $\ell_1$ penalty $\lambda$ on the x-axis, which increases from left to right. The labels along the top are the number of predictors remaining in the model for Figure 4-figure supplement 4 continued on next page.
Figure 4—figure supplement 4 continued

Panel b) shows the results of leave-one-out cross validation on the mean-squared error, along the same x-scale. The λ at minimum error is shown by the left dashed line, and the λ within one standard error is shown by the right dashed line.

DOI: https://doi.org/10.7554/eLife.26255.016
Figure 5. Manhattan plot of SNPs associated with carriage duration. The significance of each SNP’s association with carriage duration against its position in the ATCC 700669 genome is shown. The red line denotes genome-wide significance ($\alpha<0.05$ Bonferroni corrected with 92487 unique tests), and the blue line suggestive significance (2.3 orders of magnitude below significant, following convention). Loci reaching suggestive significance are labelled with their nearest annotation, as in Table 4.

DOI: https://doi.org/10.7554/eLife.26255.020
Figure 5—figure supplement 1. Possible SNPs associated with lineage and carriage duration. The SNPs and p-values as shown in Figure 5, however the x-axis is now ordered by strength of association of lineage (defined by principal component) with carriage duration. The left most lineages are those most associated, those in black/grey were not significantly associated. SNPs are coloured by the lineage they are most associated with.

DOI: https://doi.org/10.7554/eLife.26255.021
Figure 5—figure supplement 2. Distribution of lengths of significant k-mers. The lengths of those k-mers reaching significance in the LMM analysis. Lengths below 20 bases were filtered from downstream analysis, due to having low specificity.

DOI: https://doi.org/10.7554/eLife.26255.022
Figure 5—figure supplement 3. Quantile-quantile plot of association p-values. For fast-lmm results on (a) SNPs passing quality filters and (b) k-mers of all lengths passing quality filters.

DOI: https://doi.org/10.7554/eLife.26255.023
Figure 5—figure supplement 4. Manhattan plots of phage-associated SNPs associated with carriage duration. As in Figure 5, but enlarging the phage region found to be significant. SNPs are coloured by their LD with the lead SNP (the highest P-value in the region plotted), and are crosses if they are predicted to cause a change in coding sequence. Panel (a) shows LD in relation to the lead SNP at position 1516350. Panel (b) plots genes in the region, with the start and end of the phage genes labelled. Panel (c) shows LD in relation to the second SNP signal at position 1517063.

DOI: https://doi.org/10.7554/eLife.26255.024
Figure 5—figure supplement 5. Identification of phage in assemblies by blastn hit length. Histogram of the length of top hits against a database of phage sequence by blastn. Isolates with >5000 bp hits were defined as having phage present.
DOI: https://doi.org/10.7554/eLife.26255.025
**Figure 6.** Predicted mean carriage duration as a function of child age. Fit is an exponential decay over the first two years of life, using the decay rate inferred from a linear regression of log(carriage duration).

DOI: https://doi.org/10.7554/eLife.26255.029