Chapter 3

Variation in duration of asymptomatic pneumococcal carriage

Declaration of contributions

Stephen Bentley, Paul Turner, Nicholas Croucher and Julian Parkhill supervised this work. Paul and Claudia Turner designed and ran the Maela study on which this work is based, and provided the swabbing data from which carriage duration was inferred. Susannah Salter for provided data on non-typable culture positive rates. I performed all analyses.

Publication

The following has been submitted as:
Lees, J. A., Croucher N. J., Goldblatt D., Nosten F., Parkhill J., Turner C., Turner P., Bentley, S. D. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *eLife*.
Deposited on *bioRxiv*: 107086. https://doi.org/10.1101/107086

3.1 Introduction

In chapter 2 I developed a method and piece of software to perform GWAS on bacterial populations. The main test of SEER was finding known antibiotic resistance determinants. These are one of the easiest GWASs to perform in bacteria, as the effect size of these variants is so high (close to fully penetrant, hence the need to use Firth regression in some cases) and the selection pressure over time has led to the causal variants being homoplasic and broadly spread evenly across the population. In this chapter I test the method on a phenotype likely to be polygenic in origin, with causal variants that are both population stratified (lineage effects) and independent of population structure (locus effects) (Earle et al., 2016).

S. pneumoniae spends most of the transmission cycle in the nasopharynx, and so understanding and predicting the amount of time spent in this niche is critical for understanding this bacterium's epidemiology, and therefore controlling transmission (Abdullahi et al., 2012a; Melegaro et al., 2007). The nasopharynx is a complex niche in which each pneumococcal genotype must tackle a wide range of factors including host immune defence (McCool et al., 2002), other bacterial species (Pericone et al., 2000), and other pneumococcal lineages (Auranen et al., 2010; Cobey & Lipsitch, 2012) in order to maintain the genotype's population. The average nasopharyngeal duration period is therefore affected by a large number of factors, which may, themselves, interact.

A major potential advantage of GWAS in bacteria is the ability to test association with less well defined phenotypes, for example transmissibility (Nebenzahl-Guimaraes et al., 2016), or phenotypes which would be difficult to test in a lab. Here I assess genetic variation associated with pneumococcal carriage duration. Traditionally this would be difficult to assess due to the complexity of the nasopharyngeal niche, and the length of time experiments would need to be run for.

One factor that is known to strongly associate with carriage duration is serotype: as capsular polysaccharides are important in bacterial physiology and determining host immune response, different serotypes have different clearance and acquisition rates (Abdullahi et al., 2012a; P. C. Hill et al., 2010; Högberg et al., 2007; Melegaro et al., 2007; P. Turner et al., 2012). Additionally, a range of other proteins have been identified as critical to the colonisation process (Kadioglu et al., 2008), some of which exhibit similar levels of diversity to the capsule polysaccharide synthesis locus (Iannelli et al., 2002; Jedrzejas et al., 2001). However, the overall and relative contributions of these sequence variations to carriage rate have not yet been characterised. In addition variation of pathogen protein sequence, accessory genes and interaction effects between genetic elements may also have as yet unknown effects on carriage duration.

Changes in average carriage duration have been shown to be linked with recombination rate (Chaguza et al., 2016), which has been found to correlate with antibiotic resistance

(Hanage et al., 2009) and invasive potential (Chaguza et al., 2016). The carriage duration by different serotypes is widely used in models of pneumococcal epidemiology, and consequently is important in evaluating the efficacy of the PCV (Melegaro et al., 2007; Weinberger, Harboe et al., 2011). Additionally, modelling work has proposed that if alleles exist which alter carriage duration, these explain the long standing puzzle of how antibiotic-resistant and sensitive strains stably coexist in the population (Lehtinen et al., 2017). Measurement of carriage duration and the analysis of its variance beyond the resolution of serotype will have important consequences for these models.

I sought to determine the overall importance of the pathogen genotype in carriage duration in a human population, and to identify and quantify the elements of the genome responsible for the variation in carriage duration using GWAS. By combining epidemiological modelling of longitudinal swab data with and genome wide association study methods on the connected sequences, I made heritability estimates for carriage duration. I further partitioned the heritability into contributions from lineage and locus effects to quantify the variation caused by each individual factor.

3.2 Ascertainment of carriage episode duration using epidemiological modelling

I first estimated carriage duration from longitudinal swab data available for the study population. For 598 unvaccinated children up to 24 swabs taken over a two year period were available. The study population was a subset of infants from the Maela longitudinal birth cohort (C. Turner et al., 2013), and was split into two cohorts. In the 'routine' cohort, 364 infants were swabbed monthly from birth, 24 times in total. All swabs had been cultured and serotyped using the latex sweep method (P. Turner et al., 2013). In the 'immunology' cohort 234 infants were swabbed on the same time schedule, but cultured and serotyped following the World Health Organisation (WHO) method (P. Turner et al., 2012). NT pneumococci had been confirmed by bile solubility, optochin susceptibility and Omniserum Quellung negative.

I only considered swabs from infants in the study, as mothers did not have sufficient sampling resolution relative to their average length of carriage to determine carriage duration. Furthermore, the immune response of mothers to bacterial pathogens is different to children (Maródi, 2006), leading to shorter carriage durations (Gritzfeld et al., 2014).

To estimate carriage duration from the longitudinal swab data I constructed a set of hidden Markov models (HMMs) with hidden states corresponding to whether a child was carrying a serotype at a given time point, and observed states corresponding to whether a positive swab was observed for this serotype at this time point. The most general model for the swab data would be a vector with an entry of 0 or 1 for every possible serotype (of

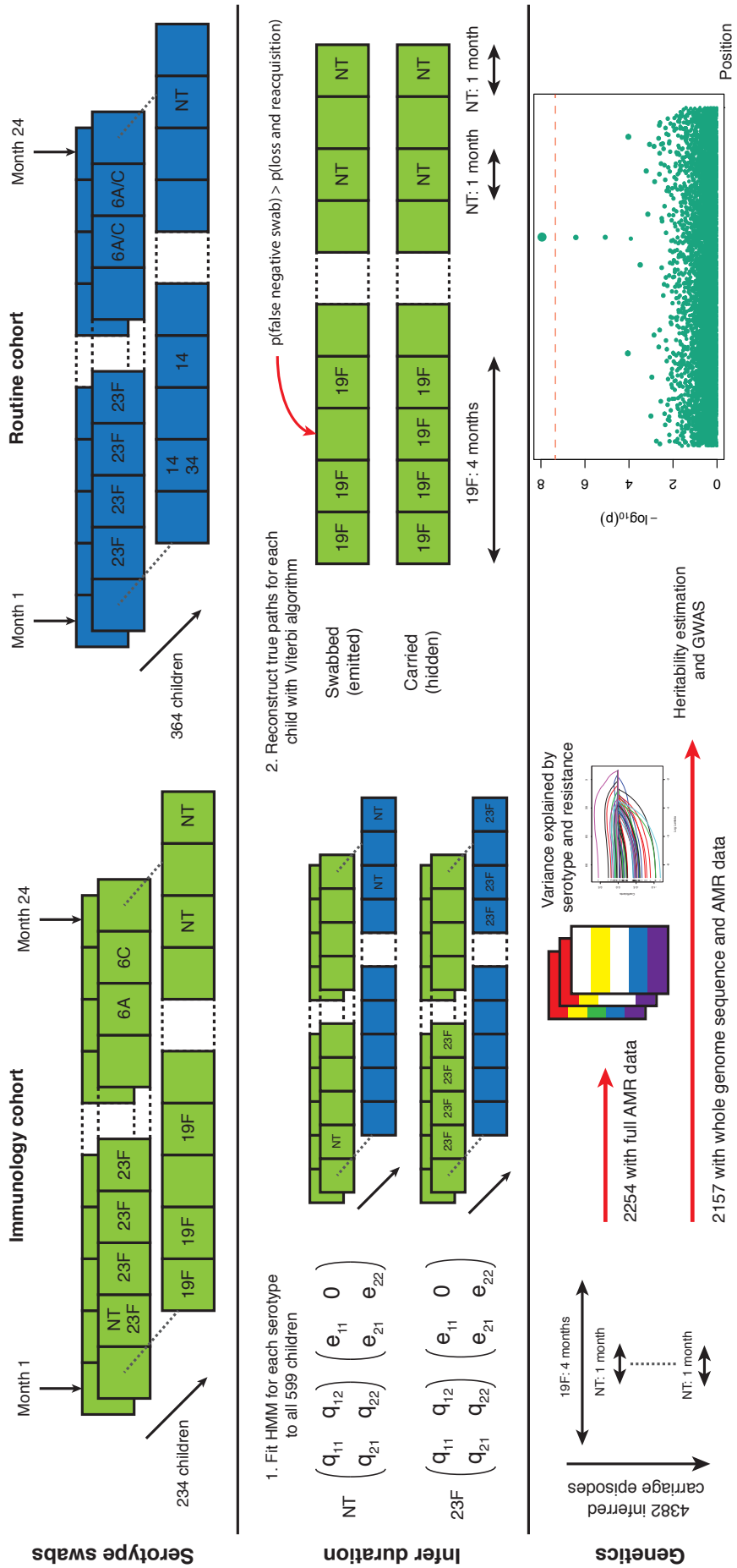

Figure 3.1: Swabbing and sequencing study design. I started with serotype swab data on 598 children from two cohorts, taken every month after birth for two years. For all samples I fitted the transition and emission probabilities of a continuous time hidden Markov model for each serotype. Then, for each child, I used these parameters were then used to infer the most likely carriage durations. I matched carriage episodes with resistance and genomic data for 2 157 episodes to draw conclusions on the basis of variation in this epidemiological parameter.

56 observed in the population), corresponding to whether each serotype was observed in the swab at each time point. However, the number of parameters to estimate in this model (with over 6 million states) is much larger than the number of data points (around 14000), and in particular some serotypes have very few positive observations. Instead, I modelled each serotype separately.

The models fitted, and their permitted transitions and emissions are shown in fig. 3.2. In model one, observation i emits state 2 if positively swabbed for the serotype, and state 1 otherwise. The unobserved states correspond to the child 'carrying' and being 'clear' of the serotype respectively. I assumed swabs have a specificity of one, so do not show positive culture when the child is clear of the carried serotype; I therefore set the coefficient for the chance of observing positive culture when no bacteria are present to zero ($e_{21} = 0$ in the emission matrix). Model two added a third state of 'multiple carriage' which is occupied when the serotype and at least one other are being carried. Both models were compared with a version which allows the parameters to covary with whether the child has carried pneumococcus previously. In model three I accounted for this explicitly by having separate states and emissions based on whether carriage has previously been observed.

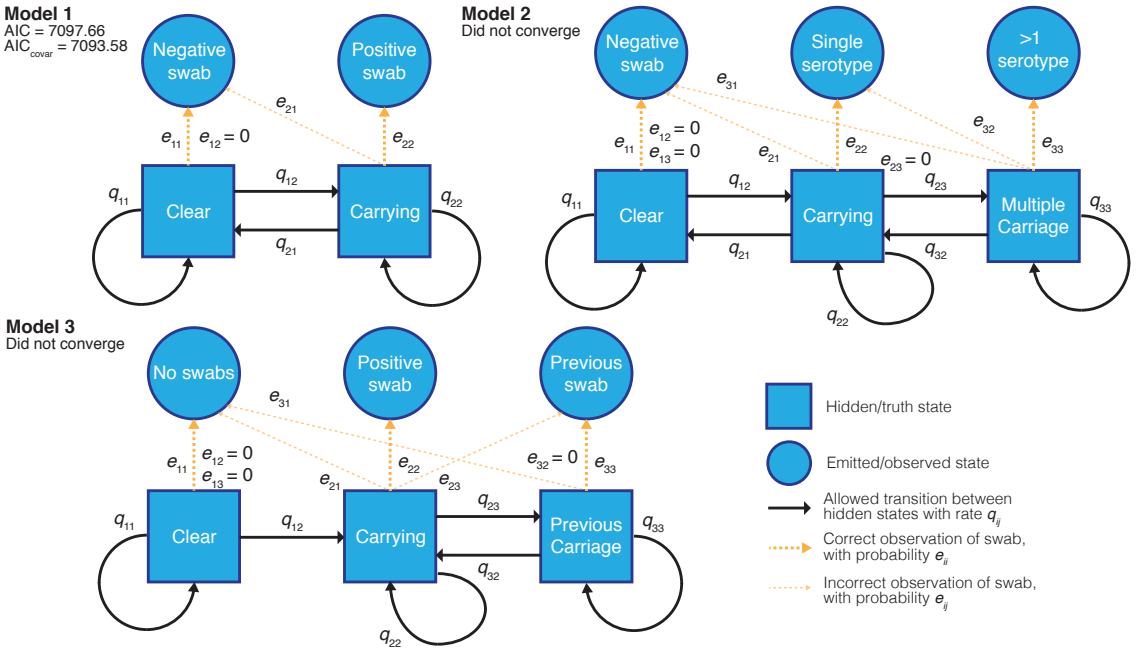

Figure 3.2: HMMs of swab time series, and their goodness-of-fit. I fitted three different models to the processed time-series data with states, allowed transitions and emissions as shown. I refitted each model allowing the transitions probabilities to covary with the age of the child and whether the 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 (AIC_{covar}).

I modelled the time series of swab data using a continuous-time HMM, as implemented in the R package `msm` (Jackson, 2011). Unobserved (true) states correspond to whether the child is carrying bacteria in their nasopharynx, and observed (emitted) states correspond to whether a positive swab was seen at each point. Transition probabilities between each state

Q and the emission probabilities **E** were jointly estimated by maximum likelihood using the BOBYQA algorithm. To get a good fit of the HMM, I normalised observation times for each sample. Defining infant birth as $t = 0$, subsequent sampling times t_i were measured in days, and normalised to have a variance of one. I then constructed the most likely path through the unobserved states for each child using the Viterbi algorithm (Forney, 1973) with the observed data and estimated model parameters. Assuming that continuous occupation of the carried state corresponded to a single carriage episode, I calculated the duration for each such episode from the inferred true states.

I applied all three models to 19F carriage episodes, as these had the most data available, and calculated the AIC (Akaike, 1974) for each model that converged. Only the simplest model (model one) converged, as judged by having a positive-definite Hessian and a converged BOBYQA run. The more complex models had lower log-likelihoods: as extensions of the simpler model they should have higher log-likelihoods, so this result was not consistent with model convergence. I tried fitting models two and three using a fixed false positive values slightly greater than zero: this lead to better log-likelihoods, but the models still didn't converge. This failure of the more complex models is probably because most children in the study immediately enter the carrying state, and episodes of dual carriage (when split up by serotype) are rare. Therefore there were not enough events between these carriage states to estimate to the transition and emission intensities, without sensitivity to initial conditions during the fitting.

I then fitted the best performing model in this test for all serotypes separately. Latex sweeps could not differentiate 6A and 6C serotypes, so I treated these as a single serotype (in WHO serotyping PCR was used to differentiate these serotypes, but I still combined them for consistency across the two cohorts). 15B and 15C serotypes spontaneously interconvert, so were combined. I also removed two duplicated swabs (08B09098 from the immunology cohort; 09B02164 from the routine observation cohort). The models for 19F, 23F, 6A/C, 6B, 14 and NT converged, but other serotypes did not have enough observations to successfully fit the parameters of the model. For these less prevalent serotypes I used the transition and emission parameters from the 19F model fitted with the correct observations when reconstructing the most likely route taken through the hidden states. I manually inspected the results to ensure this did not cause systematic overestimation when compared with previous studies.

I found that the fit for NT swabs produced results which overestimated carriage duration when compared to previously reported estimates. The best fit to the model overestimated the e_{21} parameter, which measures the false negative rate of swabbing, in favour of reduced transition intensities. I therefore fitted the model again, fixing this rate at 0.12. I based this figure on non-typable *S. pneumoniae* abundance as defined by 16S survey sequencing. At 1% proportional abundance in the sample, 12% came out as culture negative (table 3.1).

Abundance	Culture positive	Number
>1%	Cultured	361
>1%	Not cultured	44
<1%	Cultured	56
<1%	Not cultured	54

Table 3.1: Success of culturing unencapsulated *S. pneumoniae*. Based on having >1% abundance of 16S reads showing the bacteria as being present, 44/361 true positive swabs were not successfully cultured.

3.2.1 Combining epidemiological data with genomic data

From all the swab data, I estimated that there were a total of 4 382 carriage episodes (7.3 per child), of which 2 254 had a complete set of AMR data available (fig. 3.3). After removing ten outlier observations (fig. A.3) from swabs taken accidentally during disease, I was able to match 2 157 sequenced genomes with a carriage duration.

As I aimed to fit a multiple linear regression model to the carriage duration y against binary lineage associated predictors, I first ensured the data was appropriate for this model. The phenotype distribution was positively skewed, with an approximately exponential distribution. Residuals were therefore non-normally distributed, potentially reducing power (McCulloch, 2003). In the regression setting, a monotonic function can be applied to transform the response variable to avoid this problem. I first took the natural logarithm of the carriage duration

$$\hat{\mathbf{y}} = \ln(\mathbf{y})$$

which led to the residuals being much closer to being normally distributed (figs. 3.3 and A.2). I applied the same transformation to child age, when it was used as a covariate in association. For association with a LMM I instead took a monotonic transform of the carriage duration using `warped-lmm` (Fusi et al., 2014) to maximise the study's power to discover associations and estimate heritability (figs. A.1 and A.2). This used a sum over three nonlinear step functions, plus a linear term, to transform the residuals into Gaussians (Snelson et al., 2004).

For each isolate with an inferred carriage duration I extracted SNPs from the previously generated alignment against the ATCC 700669 genome (Chewapreecha, Marttinen et al.,

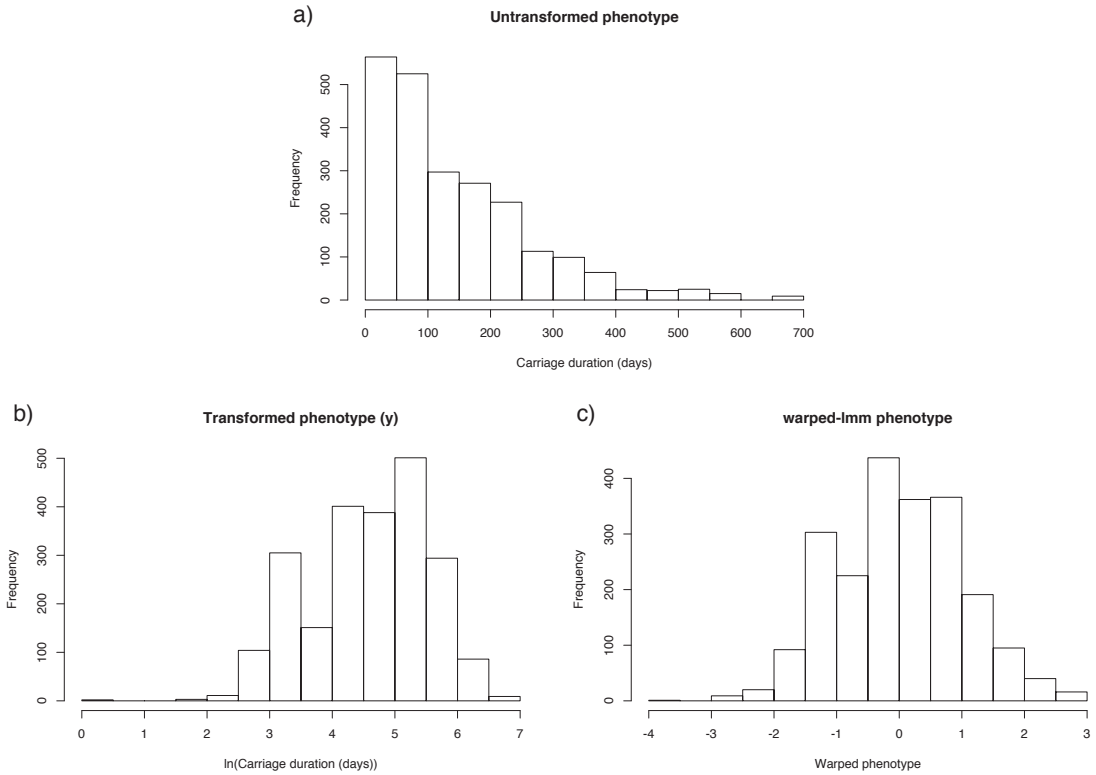

Figure 3.3: Distribution of carriage duration, and effect of monotonic transformation. Panel **a)** shows a histogram of the inferred carriage duration, **b)** shows this result after the natural logarithm is taken, and **c)** after the warping function is applied.

2014). Consequences of SNPs were annotated with `VEP`, using a manually prepared reference (McLaren et al., 2010). I generated a phylogenetic tree from this alignment using `FastTree` under the GTR+gamma model (M. N. Price et al., 2009). The carriage duration was mapped on to this phylogeny using `phytools` (Revell, 2013). I then filtered the sites in the alignment to remove any where the major allele was an 'N', any sites with a minor allele frequency lower than 1%, and any sites where over 5% of calls were missing. This left 115 210 sites for association testing and narrow-sense heritability estimation. I also used the 68M non-redundant k-mers with lengths 9-100 from the de novo assemblies of the genomes counted in section 2.2. I filtered out low frequency variants by removing any k-mers with a minor allele frequency below 2%, leaving 17M for association testing.

3.3 Overall heritability of carriage duration is high

To recap section 1.3.2, the variation in carriage duration σ^2_P is partly caused by variance in pneumococcal genetics, and variance in other potentially unknown factors such as host age and host genetics. It is common to write this sum as two components: genetic effects σ^2_G and environmental effects σ^2_E. The proportion of the overall variation which can be explained by the genetics of the bacterium is known as the broad-sense heritability

$H^2 = \frac{\sigma^2_G}{\sigma^2_G+\sigma^2_E}$. Variants which are directly associated with carriage duration independently of other variants (non-epistatic effects) contribute to the narrow-sense heritability h^2, which is smaller than the overall broad-sense heritability (Visscher et al., 2008).

H^2 can be estimated by linear regression on the phenotype of donor-recipient pairs which nearly share their genetics (Fraser et al., 2014). However in this dataset previous work was only able to confidently identify five transmission events, which was not enough to apply this method. Alternatively, analysis of variance of the phenotype between pathogens with similar genetics can be used to estimate heritability (T. J. C. Anderson et al., 2010). By applying this to phylogenetically similar bacteria (fig. 3.4), I estimated broad sense heritability H^2 with the ANOVA-CPP method in the `patherit` R package (Mitov & Stadler, 2016), using a patristic distance cutoff of 0.04 (fig. A.4). This estimated that $H^2 = 0.634$ (95% CI 0.592-0.686), implying that the genetics of *S. pneumoniae* is an important factor in determining carriage duration in this population. If environmental conditions are associated with streptococcal genotype between populations (such as host vaccination status) the heritability estimate may differ.

A lower bound on h^2 can be calculated by fitting a LMM through maximum likelihood to common SNPs (h^2_{SNP}) (S. H. Lee et al., 2011; Manolio et al., 2009). I used the 'GCTA' model implemented in `warped-lmm` (Fusi et al., 2014) to estimate h^2_{SNP} for carriage duration data, using the filtered SNPs and including child age and previous carriage as covariates. This yielded an estimate of 0.445, consistent with the estimate for H^2. I also estimated h^2_{SNP} using `LDAK` (Speed et al., 2012) with default settings, which gave an estimate of 0.437 ($<$1% difference from the `warped-lmm` estimate).

3.4 Lineage effects on carriage duration

After calculating the overall heritability, I wished to determine the amount that the specific variation in the pathogen genome contributes to changing carriage duration. However the strong LD present across the entire genome of *S. pneumoniae*, makes it difficult to pinpoint variants associated with carriage duration and not just present in the background of longer or shorter carried lineages (P. E. Chen & Shapiro, 2015). Serotype and antibiogram are correlated with the overall genome sequence (Brueggemann et al., 2003; Chewapreecha, Harris et al., 2014; Enright & Spratt, 1998), so if these factors are associated with carriage duration, large sets of variants which define long-carried and short-carried lineages will be correlated with carriage duration in a naive association test (P. E. Chen & Shapiro, 2015; T. D. Read & Massey, 2014).

I use the distinction between variants which evolve convergently and affect a phenotype independently of lineage – termed locus effects – to those which are collinear with a genotype which is associated with the phenotype, termed lineage effects (Earle et al., 2016).

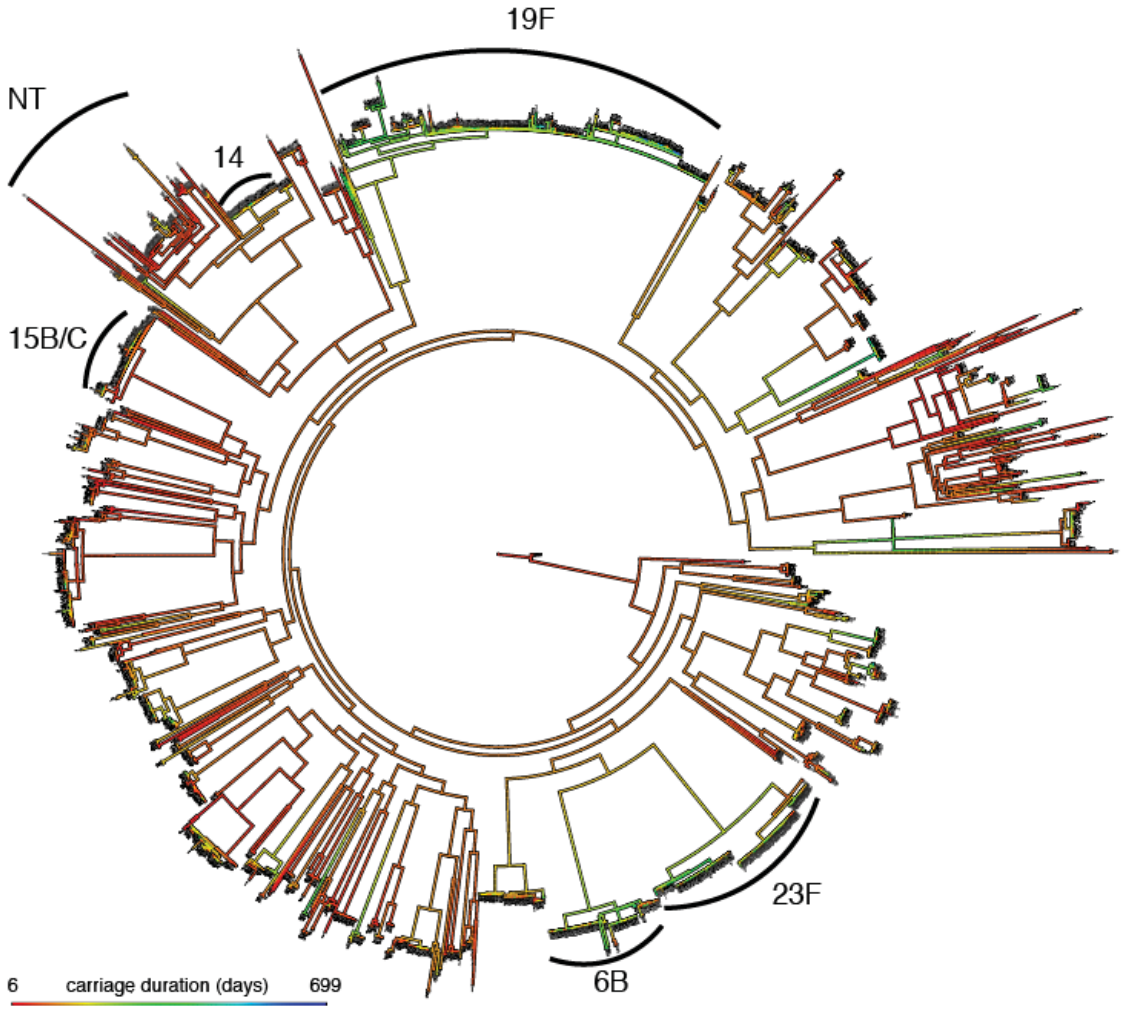

Figure 3.4: Mapping of carriage duration onto phylogeny. Using the carriage duration as a continuous trait, the ancestral state at every node of the rooted phylogeny was reconstructed. Red branches are carriage for a short time, blue for a long time. Clusters identified in previous analysis have been labelled.

Locus effects may be associated with a change in carriage duration due to convergent evolution (which may occur through recombination between lineages). In such regions, the causal loci and corresponding phenotypic effects are easier to identify (Power et al., 2016). The fixed effect model of SEER (chapter 2) or LMMs can be used to find these variants which are associated with a bacterial phenotype independent of lineage; discovery of homoplasic and polygenic variation associated with the phenotype across the entire tree is well powered (Earle et al., 2016).

While the high heritability suggests many pathogen variants do affect carriage duration, it does not give information on how many of these will be locus or lineage effects. I mapped carriage duration onto the phylogeny, reconstructing the ancestral state at each node. Consistent with the high heritability of carriage duration I found that carriage length was clearly stratified by lineage (fig. 3.4): I calculated Pagel's lambda as 0.56 ($p < 10^{-10}$) (Pagel, 1997). $\lambda = 0$ corresponds to a star-like tree, whereas $\lambda = 1$ is Brownian-motion evolution of the trait. I also modelled the evolution of carriage duration along the tree using

an Ornstein-Uhlenbeck model as implemented in `patherit`, and compared the likelihood of the full fit to that with no genetic effect on the trait ($\sigma^2_G = 0$) using a LRT with one degree of freedom. This also suggested that lineage genetics were significantly correlated with the trait (LRT = 952; $p < 10^{-10}$)

3.4.1 Serotype and drug resistance explain part of the narrow-sense heritability

I first tested for the association of serotype with carriage duration using lasso regression and with a LMM. Serotype is correlated with sequence type (Croucher, Harris, Fraser et al., 2011) and has previously been associated with differences in carriage duration (Abdullahi et al., 2012a; P. Turner et al., 2012). I also included resistance to six antibiotics, the causal element to some of which are known to be associated with specific lineages (section 2.6.2) . These are therefore possible lineage effects which would be unlikely to be found associated under a model which adjusts for population structure.

Not all serotypes and resistances may have an effect on carriage duration, or there may not be enough carriage episodes observed to reach significance. As including extra predictors in a linear regression always increases the variance explained, I first performed variable selection using lasso regression (Efron et al., 2004) to obtain a more reliable estimate of the amount of variation explained. Where a resistance and serotype are correlated and both associated with a change in carriage duration, this will produce a robust selection of the predictors (Hebiri & Lederer, 2012).

I encoded all 56 observed serotypes (including NT) and phenotypic resistance to the six antibiotics (chloramphenicol, β-lactams, clindamycin, erythromycin, trimethoprim and tetracycline) as dummy variables. I used serotype 6A/C as the reference level, as this had a mean carriage duration close to the grand mean in previous analysis. Orthogonal polynomial coding was used for the latter four antibiotics, where resistance could be intermediate or full. I then regressed this design matrix $\mathbf{X}$ against the transformed carriage duration $\hat{\mathbf{y}}$. I removed three observations with low carriage lengths due to a delayed initial swab, and seven observations with leverages of one (fig. A.3).

I performed variable selection using lasso regression (Efron et al., 2004), implemented in the R package `glmnet` (Friedman et al., 2010). I used leave-one-out cross-validation to choose a value for the ℓ_1 penalty; the value one standard error above the minimum cross-validated error (Tibshirani et al., 2001) was selected ($\lambda = 0.033$; fig. A.5). The 20 predictors with non-zero coefficients in the model at this value of λ were used in a linear regression to calculate the multiple R^2, which corresponds to the proportion of variance explained by these predictors.

I also estimated the variance components from serotype and resistance using genomic partitioning (J. Yang, Manolio et al., 2011), as implemented in `LDAK`. This estimates h^2

from a subset of the overall genetic loci, allowing for the heritability associated with a particular region of the genome to be tested. I used SNPs in the capsule locus to calculate a kinship matrix approximating the contribution from serotype variation. For antibiotic resistance I used SNPs in the *pbp* genes, *dyr* gene and ICE transposon to calculate a kinship matrix. Restricted maximum likelihood was used to estimate the variance explained by each of these components.

The selected predictors and their effect on carriage duration are shown in table 3.2. The total variance explained by these lineage factors was 0.19, 0.178 for serotype alone and 0.092 for resistance alone. When I used genomic partitioning of variance components these were instead estimated to be 0.253, 0.135 and 0.113, respectively. I applied the covariance test (Lockhart et al., 2014) to determine which lineage effects were significantly associated with carriage duration and found that 19F, erythromycin resistance, 23F, 6B caused significant ($\alpha < 0.05$) increase in carriage duration and being non-typable caused a significant decrease.

Factor	Effect on carriage duration (days)
Mean (intercept)	**59.5**
Erythromycin resistance	**+7.5**
Tetracycline resistance	+3.0
Trimethoprim resistance	+2.9
Clindamycin resistance	+1.8
Penicillin intermediate resistance	+1.3
Serotype 19F	**+46.9**
Serotype 23F	**+21.0**
Serotype 6B	**+16.2**
Serotype 14	+7.2
Serotype 21	+1.6
Serotype 19B	-0.1
Serotype 18C	-1.9
Serotype 29	-4.3
Serotype 3	-4.5
Serotype 4	-7.2
Serotype 24F	-8.5
Non-typable	**-12.3**
Serotype 5	-18.6

Table 3.2: Coefficients from lasso regression model of carriage duration. The mean (intercept) corresponds to a sensitive 6A/C carriage episode, and different serotypes and resistances are perturbations about this mean. Positive effects are expected to have a greater magnitude, due to the positive skew of carriage duration. Rows in bold were significant predictors in the covariance test.

3.4.2 Independent effects of serotype and genetic background

Previous studies have used isogenic strains to look for effects of serotype of colonisation and carriage duration independent of genetic background. Resistance to killing (Weinberger et al., 2009), growth phenotype (Hathaway et al., 2012) and resistance to complement (Melin et al., 2010) have all been shown to affect carriage through serotype rather than genetic background. Conversely, some bacterial genetic variation has been shown to be able to affect colonisation independent of serotype (Nadeem Khan et al., 2014).

I therefore wished to test whether the detected effect of serotype and resistance on carriage duration was entirely mediated through their covariance with lineage, or whether they are independently associated with carriage duration. I first looked for differences in duration over three recent capsule gain/loss events; if there is an effect of serotype independent of genetic background, these would be predicted have the largest difference

between serotypes while controlling for the relatedness of isolates. Capsule switch events had been previously identified by first reconstructing of the ancestral state of the serotype at each node through maximum parsimony (Chewapreecha, Harris et al., 2014). For each node involving loss or gain of the capsule, those with at least one child being a tip were selected to find recent switches (all were capsule gain). The carriage duration of all unencapsulated children (in the phylogenetic sense) of the identified node were used as the null distribution to calculate an empirical p-value for the switched isolate. P-values were combined using Fisher's method (Rosenthal, 1978). No significant difference in duration was seen between isolates with or without capsule within the same lineage (p = 0.39; fig. 3.5).

However, as these events were limited in number, assumed genetic independence within the clade and occurred only in part of the population, I also performed the same regression as above while also including lineage (defined by discrete population clusters) as a predictor. This therefore allows serotypes which appear in different population clusters to distinguish whether lineage or serotype had a greater effect on carriage duration. The covariance test found that 19F, erythromycin resistance and being non-typable had significant effects on the model (in that order). As these terms enter the model before any lineage specific effect, this suggested these serotypes and resistances are associated with variation in carriage duration independent of background genotype

This lasso-based analysis may be vulnerable to confounding from unmeasured variables which may be associated with the explanatory variables (serotype and resistance). To fully account for the effect of the bacterial genome rather than relying on discrete clusters as covariates in the regression, I then performed regression of these lineage effects under a LMM where the relatedness between strains was instead included as a random effect. The predictors had the same order of significance, but only serotype 19F reached genome-wide significance ($p = 3.8 \times 10^{-7}$).

Together, this suggested that the main lineage effect on carriage duration is the serotype, but only some serotypes (19F) have an association independent of genetic background. I also found that erythromycin resistance may be significantly associated with an increased carriage duration. While being a relatively uncommon treatment in this setting (3% of treatments captured), I did not find that other antibiotics were associated. This may be because erythromycin resistance would be expected to cause an almost four order magnitude increase in minimum inhibitory concentration (MIC), whereas other resistance acquisitions have a much smaller effect.

3.4.3 Average carriage duration by serotype

Additionally, I calculated the mean sojourn times (average length of time children are expected to remain in the carrying state of the model with the given serotype) and mean

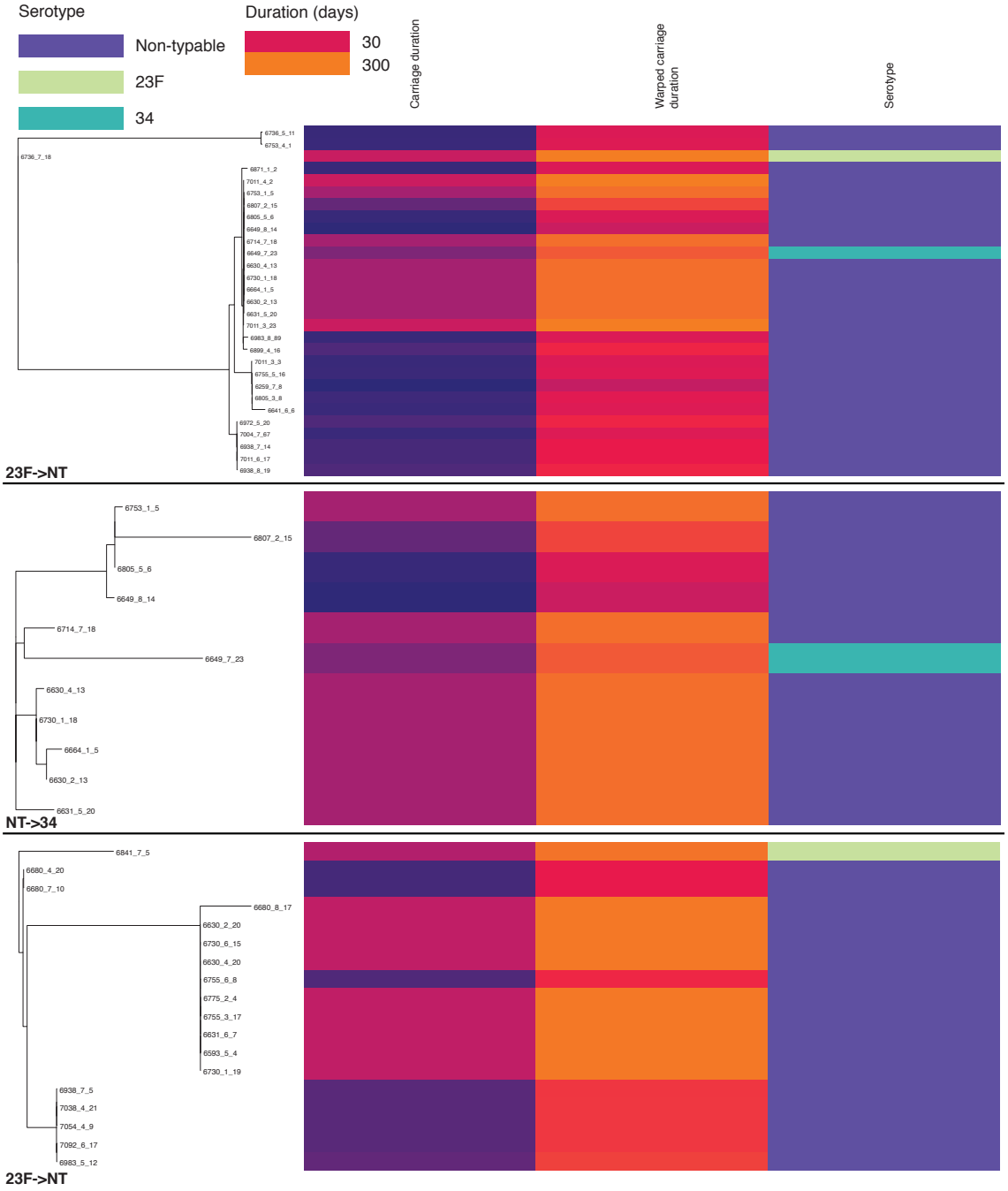

Figure 3.5: Change in carriage duration associated with capsule switching events. For each of the three events analysed the subtree containing the switch is shown on the left. For each isolate within the subtree, carriage duration (on a roughly exponential scale), warped carriage duration (on a roughly linear scale) and serotype are shown as coloured bars aligned with the tip.

number of carriage episodes from the fit to the HMM for commonly carried serotypes (table 3.3), which gave results similar to the regression performed above. These estimates are comparable to the previous analysis on a subset of these samples. The majority of carriage episodes were due to five of the seven paediatric serotypes (Shapiro & Austrian, 1994), or non-typeable isolates. The results show 19F, 23F and 14 were carried the longest, 6A/C and 6B for intermediate lengths, and NT the shortest.

Serotype	Sojourn time (days)	Expected number of infections
19F	292*	0.85
23F	112	0.83
6A/C	76.4	0.88
6B	114	0.75
14	137*	0.58
NT	40.6	2.05

Table 3.3: Mean length of carriage, and expected number of carriage episodes within the first two years of life. Only serotypes with enough data for the HMM fit to converge are shown. Starred observations have a standard error which is larger than the estimated value, indicating low confidence in the estimate.

The overall picture of the first two years of infant carriage is one containing one or two long (over 90 day) carriage episodes of a common serotype (6A/C, 6B, 14, 19F, 23F) and around two short (under a month) carriage episodes of non-typable *S. pneumoniae*. Colonisation by other serotypes seem to cause slightly shorter carriage episodes, though the relative rarity of these events naturally limits the confidence in this inference. That some serotypes are rarer and carried for shorter time periods may be evidence of competitive exclusion (Hardin, 1960; Trzciński et al., 2015), as fitter serotypes quickly replace less fit serotypes thus leading to reduced carriage duration. The calculated mean carriage duration of NT pneumococci is similar to the minimum resolution I was able to measure by the study design, which suggests carriage episodes may actually be shorter than one month. Unfortunately the only existing study with higher resolution did not check for colonisation by NT pneumococci (Abdullahi et al., 2012a).

These estimates are similar to previous longitudinal studies in different populations (P. C. Hill et al., 2010; Högberg et al., 2007; Melegaro et al., 2007), though against the Kilifi study these estimates are systematically larger. This may be due to the lower resolution swabbing we performed, or may be because the previous study was unable to resolve multiple carriage (11% of positive swabs). While the heritability estimates are specific to this population due to differences in host, vaccine deployment and transmission dynamics, the similarity of the estimates of serotype effect to those from different study populations suggests our results may be somewhat generalisable.

3.5 Additional loci identified by genome-wide association

To search for locus effects I used the linear mixed model implemented in `fast-lmm` (Lippert et al., 2011) to associate genetic elements with carriage duration, independent of overall lineage effects. I used the warped phenotype as the response, the kinship matrix (calculated from SNPs) as random effects, and variant presence, child age and previous carriage as fixed effects. For SNPs I used a Bonferroni correction with $\alpha < 0.05$ and an N of 92 487 phylogenetically independent sites to derive a genome-wide significance cutoff of $p < 5.4 \times 10^{-7}$, and a suggestive significance cutoff (Lander & Kruglyak, 1995; Stranger et al., 2011) of $p = 1.1 \times 10^{-4}$. I tested pairwise LD between the significant SNPs by calculating the R^2 between them. I removed those with $R^2 > 0.2$, assuming these represented the same underlying signal, to define the significant loci. For k-mers I counted 5 254 876 phylogenetically independent sites, giving a genome wide significance cutoff of 9.5×10^{-9}. I used `blastn` with default settings to map the significant k-mers to seven reference genomes (ATCC 700669, INV104B, OXC141, SPNA45, Taiwan19F, TIGR4 and NT_110_58), and the possible Tn*916* sequences (Croucher, Harris, Fraser et al., 2011).

The results for SNPs are shown in fig. 3.6 and table 3.4, with 14 loci reaching suggestive significance and two reaching genome-wide significance (top hit $\beta = 0.17$; $p = 2.1 \times 10^{-7}$; MAF = 1%). I also found that 424 k-mers reached genome-wide significance (top hit $\beta = 0.11$; $p = 2.1 \times 10^{-12}$; MAF = 2%), which I filtered to 321 k-mers over 20 bases long to remove low specificity sequences (fig. A.7). To determine their function, I mapped these k-mers to the coordinates of reference sequences.

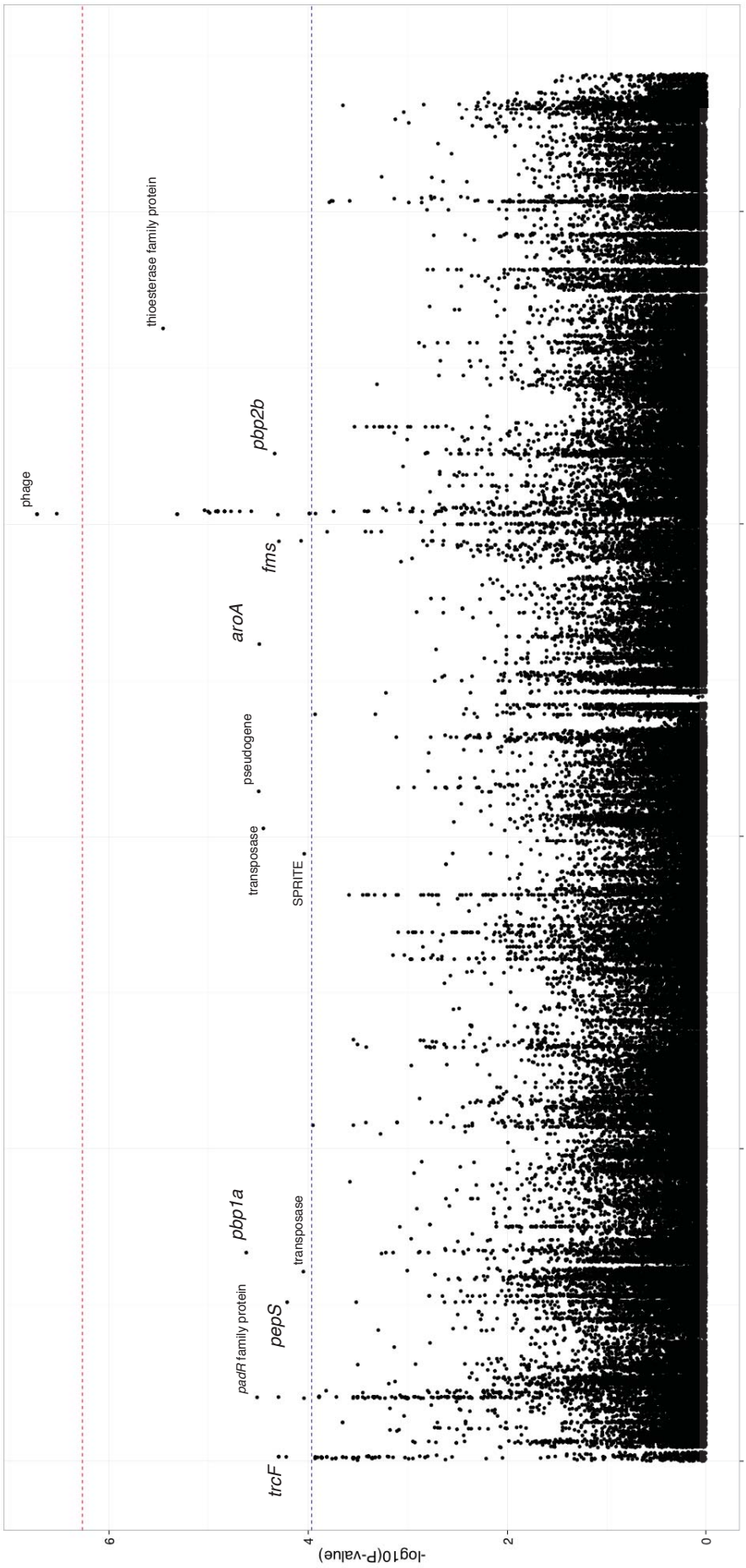

Figure 3.6: Manhattan plot of SNPs associated with carriage duration. The significance of each SNP's association with carriage duration against its position in the ATCC 700669 genome is shown. The red line denotes genome-wide significance ($\alpha < 0.05$ Bonferroni corrected with 92 487 unique tests), 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 3.4.

Co-ordinate	Nearest annotation	Effect size	P-value	Significance level
6753	*trcF*	-0.12	6.2×10^{-5}	Suggestive
254312	*pepS*	-0.11	6.4×10^{-5}	Suggestive
303239	IS630-Spn1 transposase	0.078	9.2×10^{-5}	Suggestive
333632	*pbp1a*	0.079	2.5×10^{-5}	Suggestive
971849	SPRITE repeat region	0.078	9.4×10^{-5}	Suggestive
1013978	IS630-Spn1 transposase	0.11	3.7×10^{-5}	Suggestive
1073185	FM211187.3435 (pseudogene)	0.086	3.3×10^{-5}	Suggestive
1308604	*aroA*	-0.27	3.8×10^{-5}	Suggestive
1472933	Upstream of *fms*	-0.23	5.3×10^{-5}	Suggestive
1473700	putative glutathione S-transferase	-0.16	8.8×10^{-5}	Suggestive
1515497	hypothetical phage protein	-0.099	5.2×10^{-5}	Suggestive
1516293	putative phage Holliday junction resolvase	-0.10	5.1×10^{-6}	Suggestive
1516350	putative phage Holliday junction resolvase	-0.12	2.1×10^{-7}	Genome-wide significant
1517063	phage protein	-0.11	3.3×10^{-7}	Genome-wide significant
1613197	*pbp2b*	-0.21	4.8×10^{-5}	Suggestive
1813192	thioesterase superfamily protein	-0.12	4.8×10^{-6}	Suggestive

Table 3.4: SNP locus effects at genome-wide and suggestive significance. Co-ordinates are with respect to the ATCC 700669 reference genome, and are for the lead SNP in each locus after LD-pruning. Effect sizes are for the warped phenotype.

3.5.1 Prophage sequences associated with reduced carriage duration

The only genome-wide significant SNP hits are synonymous changes in the replication module of the prophage in the ATCC 700669 genome (MAF = 1%), a highly variable component of the pneumococcal genome (Croucher, Coupland et al., 2014) (fig. 3.7). The LD structure suggested there were two separate significant signals found in this region. I therefore performed another GWAS conditioning on the top hit (using it as a fixed effect in the regression at other sites, and removing it from kinship estimation) to test if there was a second independent signal, but found that the second hit in this region was no longer significant (position 1526024; $p = 2.2 \times 10^{-4}$). The current data is therefore consistent

with only a single significant hit to prophage.

The most significant k-mer hits were also located in phage sequence (MAF 2%) and were associated with a reduced duration of carriage. As these mobile genetic elements are less weakly population stratified than other regions of the genome, they are easier to find as locus effects. The LD in this region is less than in the rest of the genome, as prophage sequence is highly variable within *S. pneumoniae* lineages (Croucher, Coupland et al., 2014). Multiple independent phage variants may therefore affect carriage duration, which will increase their significance using a LMM. Indeed, the significant results from the LMM (top SNP $p = 2.1 \times 10^{-7}$; top k-mer $p = 2.1 \times 10^{-12}$) are not significant (top SNP $p = 5.1 \times 10^{-6}$; top k-mer $p = 5.7 \times 10^{-8}$) under a model of association using a linear regression with the first 30 principal components as fixed effects to control for population structure rather than random effects, and are strongly associated with the population structure components of the model (highest association $p = 5.2 \times 10^{-75}$ with principal component 2).

I first postulated that presence of any phage in the genome may cause a reduction in carriage duration. I identified the presence of phage by performing a `blastn` of the de novo assemblies against a reference database of phage sequence (Croucher et al., 2016). If the length of the top hit was over 5000 I defined the isolate as having phage present (fig. A.6). I then used the presence of phage as a trait under the same linear mixed model, however I found no evidence of association when correcting for population structure (p = 0.35). These results are therefore evidence that infection with a specific phage sequence is associated with a slight decrease in carriage duration. A similar result has previously been found in a genome-wide screen in *N. meningitidis*, where a specific phage sequence was found to affect the virulence and epidemiology of strains (Bille et al., 2005; Bille et al., 2008). Additionally, previous in vivo tests have shown phage elements to cause a fitness decrease of *S. pneumoniae* during carriage (DeBardeleben et al., 2014).

The genetic polymorphisms in the prophage associated with changes in carriage duration, found in 2% of viral sequences, were within coding sequences inside the phage replication module (Romero et al., 2009). It is unlikely the specific variants of these proteins cause a significant difference in phenotype, because they are only highly expressed after the prophage is activated, and cell lysis usually happens shortly afterwards. One explanation for these results is that some subpopulations of prophage do not cause a significant decrease in their host bacterium's carriage duration, which could be due to beneficial 'cargo' genes. However previous surveys of pneumococcal prophage have found little evidence of these elements carrying such sequences (Croucher, Coupland et al., 2014; Romero et al., 2009). One phage protein that has been found to alter the bacterial phenotype is PblB, a phage structural protein that can also mediate bacterial adhesion to platelets (Loeffler & Fischetti, 2006). However, *pblB* is within the morphology module (Romero et al., 2009) and as an adhesin might if anything be expected to increase carriage

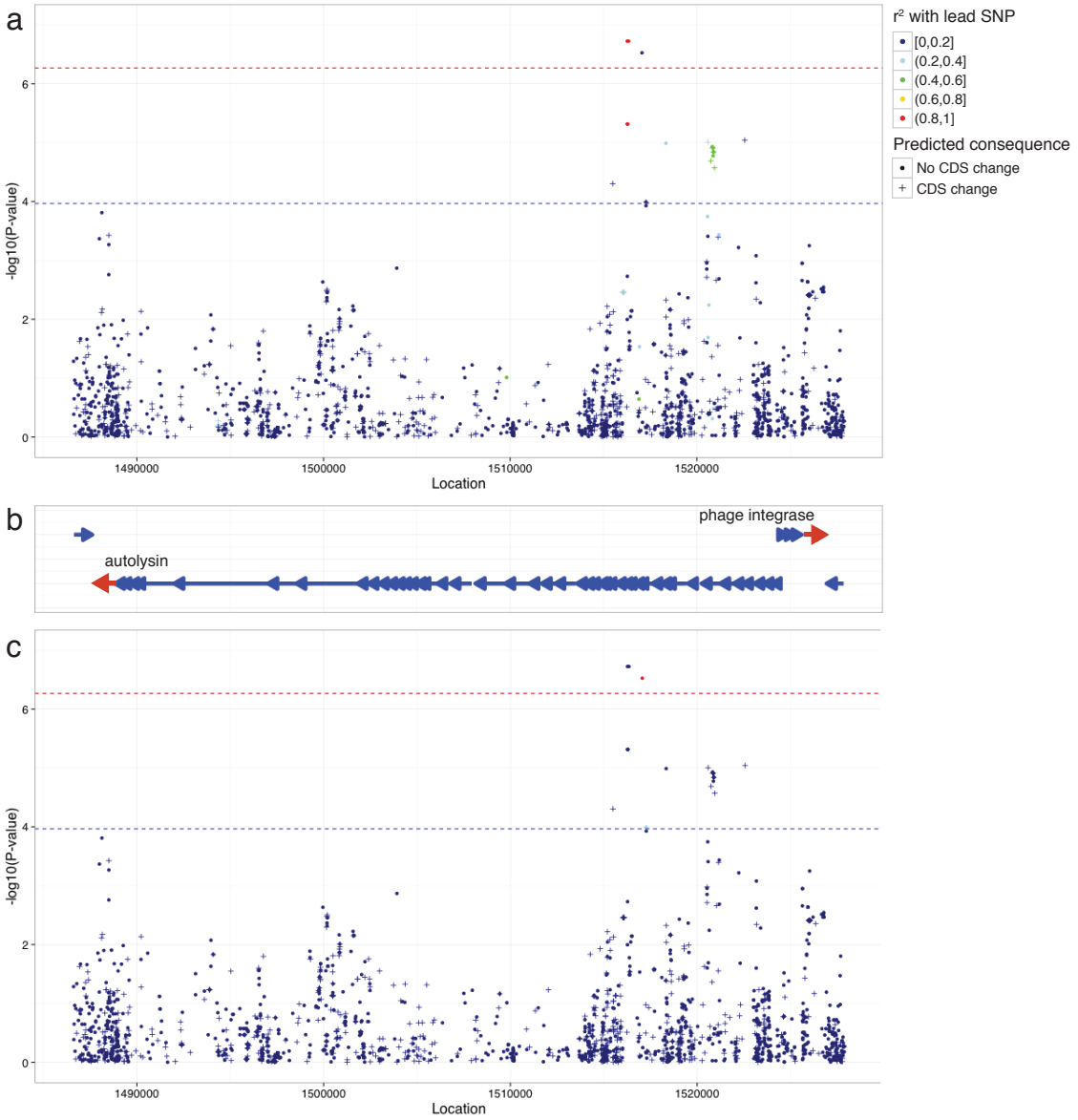

Figure 3.7: Manhattan plots of phage-associated SNPs associated with carriage duration. As in fig. 3.6, but enlarging the phage region found to be significant. SNPs are coloured by their LD with the lead SNP (the highest P-value in the region plotted), and are crosses if they are predicted to cause a 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.

duration, and was not detected in the association analysis. Hence the detected association is unlikely to represent expression of viral machinery or cargo genes in the host cell while the prophage is dormant.

Alternatively, the association with only a subset of prophage may have been a consequence of the study sampling design. Using a monthly swabbing approach, it was only possible to infer changes in the carriage duration of genotypes that colonise hosts for relative long periods. Therefore any prophage variant that enhances a virus' ability to infect long carriage duration pneumococci may have an increased association with the variation in the observed phenotype. As phage commonly exhibit high levels of strain specificity (Duplessis & Moineau, 2001), this is a plausible mechanism, although the role of the replication module in such host preference is unclear.

An additional mechanism by which prophage can affect host phenotype is by inserting into, and thereby disrupting, functional genes. Pneumococcal prophage frequently insert into *comYC*, thereby preventing the host cell undergoing transformation (Croucher, Harris, Fraser et al., 2011; Croucher, Hanage et al., 2014). Using previous categorisation of the *comYC* gene in this collection into intact versus interrupted or missing (Croucher et al., 2016), I found that having an intact *comYC* gene (23% of isolates) was significantly associated with an increased carriage duration ($\beta = 0.29$; $p = 1.4 \times 10^{-44}$). The effect size is similar to the associated phage k-mers, but has at a higher allele frequency (hence the increased significance of the result). An interpretation consistent with these findings would be that the effect of phage k-mers is actually through interrupting *comYC*. The k-mers themselves were spread out to lower frequencies due to their sequence variability, and none of the references I used allowed mapping to find the *comYC* interruption directly.

3.5.2 Other loci associated with altered carriage duration

Signals at the suggestive level included *pbp1a* and *pbp2b*, which suggested as above that penicillin resistance may slightly increase carriage duration, but there are not enough samples in this analysis to confirm or refute this. Other signals near genes at a suggestive level included SNPs in *trcF* (transcription coupled DNA repair), *padR* (repressor of phenolic acid stress response), *pepS* (aminopeptidase), *aroA* (aromatic amino acid synthesis), *fms* (peptide deformylase) and a thioesterase superfamily protein. K-mers from erythromycin resistance genes (*ermB*, *mel*, *mef*) were expected to reach significance from the above analysis, but did not: however I showed in section 2.6.2 that the power to detect these elements in a larger sample set taken from the same population is limited due to the multiple resistance mechanisms and stratification of resistance with lineage.

The test statistic from `fast-lmm` roughly followed the null-hypothesis, with the exception of the significant phage k-mers (fig. A.8). However there is limited power to detect effects associated with both the lineage and phenotype. This effect has been previously

noted, and while LMMs have improved power for detecting locus specific effects they lose power when detecting associated variants which segregate with background genotype (Earle et al., 2016). To search for candidate regions which may be independently associated with both a lineage and increased carriage duration, I ran an association test with SEER (chapter 2) using a set number of fixed effects as the population structure correction. In this case I used the patristic distances from the phylogenetic tree as the kinship matrix, which I then projected into 30 dimensions using metric multidimensional scaling to obtain covariates. This may be expected to have higher power than an LMM for true associated variants on ancestral branches as some association with population structure is permitted, but will also increase the number of false positives (variants co-occurring on these branches which do not directly affect the carriage duration themselves). To reduce the number of false positives I used a strict threshold of $p < 10^{-14}$. I separately tested SNPs for their association with those principal components which were themselves significantly associated with carriage duration, and therefore may be driving the lineage associations using the model of Earle et al. (2016).

The most highly associated SNPs were in all three *pbp* regions associated with β-lactam resistance, the capsule locus, *recA* (DNA repair and homologous recombination), *bgaA* (beta-galactosidase), *phoH*-like protein (phosphate starvation-inducible protein), *ftsZ* (cell division protein) and *groEL* (chaperonin). As 19F, the serotype most associated with carriage duration, is predominantly the β-lactam resistant PMEN14 lineage the *pbp* association may be driven through strong LD between with this serotype. Figure A.9 shows the analysis of SNPs which may be driving significant lineage associations – this also suggested *dnaB* (DNA replication) may be associated with altered carriage duration. Associated k-mers were also found in *phtD* (host cell surface adhesion), *mraY* (cell wall biosynthesis), *tlyA* (rRNA methylase), *zinT* (zinc recruitment), *adcA* (zinc recruitment) and *recJ* (DNA repair). Additionally I found k-mers in the bacteriocin *blpZ* and immunity protein *pncM* (Bogaardt et al., 2015) to be associated with variability in carriage duration. This could be evidence that intra-strain competition occurs within host via this mechanism, consistent with previous in vitro mouse models (Dawid et al., 2007).

It is not possible to determine whether variation in these genes is associated with a change in carriage duration or if the variation is present in longer carried, generally more prevalent lineages. For example, β-lactam resistance may appear associated as the long carried lineages 19F (dominated by PMEN14, as noted above) and 23F are more frequently resistant, or it may genuinely provide an advantage in the nasopharynx that extends carriage duration independent of other factors. Future studies of carriage duration, or further experimental evidence will be needed to determine which is the case for these regions.

Antigenic variation in known regions (of *pspA*, *pspC*, *zmpA* or *zmpB*) may be expected to cause a change in carriage duration (Lipsitch & O'Hagan, 2007), however I did not find

any of these to be associated with a change in carriage duration. This was likely due to stratification of variation in these regions with lineage, but may also be caused by a larger diversity of k-mers in the region reducing power to detect an association.

3.6 Child age independently affects variance in carriage duration

Finally, I wished to determine the importance of two environmental factors which are known to contribute to variance in this phenotype: child age and whether the carriage episode is the first the child has been exposed to (Abdullahi et al., 2012a, 2012b; P. Turner et al., 2012). These have been applied throughout the analysis as covariates, both in the estimation of carriage episodes and in associating genetic variation with change in carriage duration.

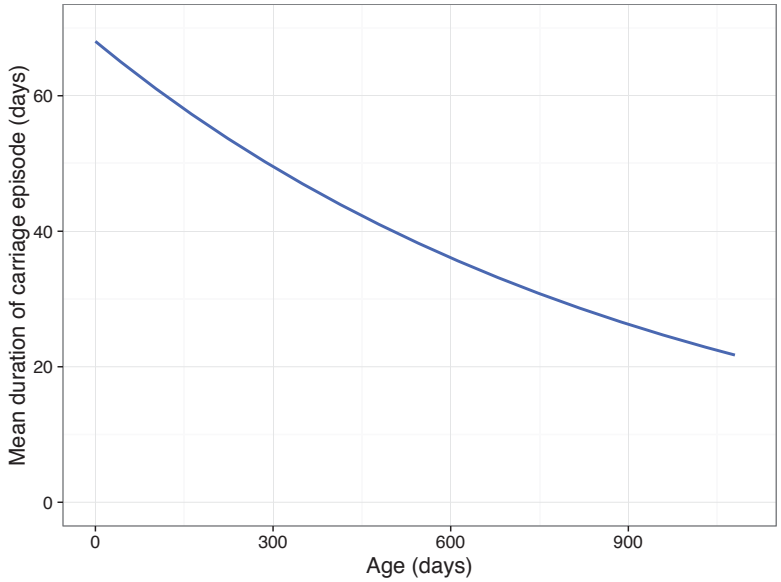

Figure 3.8: Predicted mean carriage duration as a function of child age. Fit is an exponential decay over the first two years of life, using the decay rate inferred from a linear regression of log(carriage duration).

I applied linear regression to these factors while using the first 30 PCs to correct for the effect of the bacterial genome, which showed they were both significantly associated with carriage duration as expected (age $p = 3.9 \times 10^{-7}$; previous carriage $p = 2.5 \times 10^{-8}$). Using the linear mixed model to control for bacterial genotype both factors were again significant (LRT = 26.4; $p = 1.8 \times 10^{-6}$). Together, they explained 0.046 of variation in carriage duration. As found previously, increasing child age contributes to a decrease in the duration of carriage episodes. From a mean of 68 days long, I calculated a drop of 19 days after a year, and 32 days after two years. Extrapolating, this causes carriage episodes longer than two days to cease by age 11 (fig. 3.8). Previous carriage of any serotype was

estimated to cause an increase in the duration of future carriage episodes, though previous studies have found no overall effect (Weinberger et al., 2008). It has previously been shown that prior exposure to non-typables in this cohort make colonisation by another non-typable occur later, and for a shorter time (P. Turner et al., 2012). The positive effect observed in this analysis is therefore likely to be an artefact due to subsequent carriage episodes being more likely to be due to typable pneumococci.

Additional environmental factors that explain some of the remainder of the variance may include the variation of the host immune response and interaction with other infections or co-colonisation. In particular, co-infection with influenza A was not recorded but is known to affect population dynamics within the nasopharynx (Kono et al., 2016). Fundamentally, imprecise inference of the carriage duration will limit the ability to fully explain its variance here.

3.7 Conclusions

Other than serotype, the genetic determinants of pneumococcal carriage duration were previously unknown. By developing models for longitudinal swab data and combining the results with whole genome sequence data I have quantified and mapped the genetic contribution to the carriage duration of *S. pneumoniae*. I found that despite a range of other factors such as host age which are known to cause carriage duration to differ, sequence variation of the pneumococcal genome explains most of this variability (63%). Common serotypes and resistance to erythromycin caused some of this effect (19% total), as does the presence or absence of particular prophage sequence in the genome. Table 3.5 summarises the sources I found to be significantly associated with variation in carriage duration.

Source	of which is	Total variance explained	Proportion of total heritability explained
Total heritability (H^2)		0.634 (CPP)	1.00
	Common SNP heritability (h^2_{SNP})	0.438 (LMM)	0.691
	Serotype and resistance	0.190 (R^2)	0.300 (R^2)
		0.253 (LMM)	0.399 (LMM)
	Serotype only	0.178 (R^2)	0.281 (R^2)
		0.135 (LMM)	0.213 (LMM)
	Resistance only	0.092 (R^2)	0.145 (R^2)
		0.113 (LMM)	0.178 (LMM)
	Phage k-mers	0.067 (LMM)	0.106
	Intact *comYC*	0.127 (LMM)	0.201
Measured environmental effects	Age and previous carriage	0.046 (R^2)	-

Table 3.5: Summary of variance of carriage duration explained by genetic and environmental factors. H^2 encompasses all rows, other than the measured environmental effects. For each variant component the method used to estimate it is reported: CPP - closest phylogenetic pairs; LMM - variance component using a linear mixed model with pathogen genotype as random effects; R^2 - linear regression using lasso to select predictors.

I have provided a quantitative estimate of how closely transmission pairs share their carriage duration, and show evidence for differences both between and within serotypes. The implication of phage as having a significant effect on carriage duration has interesting corollaries on pneumococcal genome diversification through frequent infection and loss of prophage, even during carriage episodes in this dataset.

Investigating a mechanism for the prophage association, I found that having an intact *comYC* gene, which is frequently interrupted by prophage causing loss of function of the competence system, was associated with increased carriage duration. While the competence system is observed to remain intact over the evolutionary history of the species, these disruptive mutations spread irreversibly through the population as competent bacteria can acquire the mutation, and non-competent bacteria can no longer reverse it through recombination (Croucher, Hanage et al., 2014). Selection must therefore maintain

the function at this locus over short timescales, and an increased carriage duration may be evidence of this. I therefore hypothesise that the associated prophage sequences may haved affected carriage duration through disruption of the competence system.

The results presented here have important implications for the modelling of pneumococcal transmission and their response to perturbation of the population by vaccine. Importantly, the analysis of heritability shows that variants other than serotype affect carriage duration, consistent with recent theoretical work (Lehtinen et al., 2017). Here I have shown that these alleles do exist in a natural population, and also identified candidates for the loci which fulfil this role. Together these studies suggest that variants exist in the pneumococcal genome which alter carriage duration, which in turn is linked to antibiotic resistance.

I was not able to fully explain the basis for heritability of carriage duration for a number of reasons. The close association of the phenotype with lineage limited our power to fine-map lineage associated variants other than capsule type which may affect carriage duration. Meta-analysis with more large studies with higher resolution may help to resolve these issues. Additional environmental factors that explain some of the remainder of the variance may include the variation of the host immune response and interaction with other infections or co-colonisation. In particular, co-infection with influenza A was not recorded but is known to affect population dynamics within the nasopharynx (Kono et al., 2016).

This is a phenotype which would have been difficult to assay by traditional methods such as in an animal model due to the cohort size needed and the length of time experiments would need to be run for. By using GWAS I have been able to quantitatively investigate a complex phenotype in a natural population. This chapter has also advanced the application of GWAS methods applied to bacteria started in chapter 2 by application to a more difficult to define phenotype, introducing heritability and genomic partitioning, and testing specifically for locus effects. I have also implicitly compared fixed effect and random effect models to control for population structure. In the next chapter I will continue using these approaches to identify pneumococcal genetic variation associated with bacterial meningitis, while developing a more thorough catalogue of variation within the pneumococcal genome.