Pneumococcal genetic variability influences age-dependent bacterial carriage

Kremer PHC[1], Ferwerda B[1,2], Bootsma HJ[3], Rots NY[3], Wijmega-Monsuur AJ[3], Sanders EAM[3,4], Trzciński K[4], Wyllie AL[4,5], Turner P[6,7], van der Ende A[8,9], Brouwer MC[1], Bentley SD[10], van de Beek D[1], Lees JA[11]

Affiliations

[1] Amsterdam UMC, University of Amsterdam, Department of Neurology, Amsterdam Neuroscience, Meibergdreef, The Netherlands

[2] Amsterdam UMC, University of Amsterdam, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, The Netherlands.

[3] Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, 3721 MA, The Netherlands.

[4] Department of Pediatric Immunology and Infectious Diseases, Wilhelmina Children's Hospital, University Medical Centre Utrecht, Utrecht, 3508 AB, The Netherlands.

[5] Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

[6] Cambodia Oxford Medical Research Unit, Angkor Hospital for Children, Siem Reap, Cambodia

[7] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[8] Amsterdam UMC, Department of Medical Microbiology and Infection Prevention, The Netherlands

[9] The Netherlands Reference Laboratory for Bacterial Meningitis, Amsterdam, The Netherlands

[10] Parasites and Microbes, Wellcome Sanger Institute, Hixton, Cambridge, United Kingdom

[11] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom

Funding statement

This work was supported by grants from the European Research Council (ERC Starting Grant, proposal/contract 281156; https://erc.europa.eu) and the Netherlands Organization for Health Research and Development (ZonMw; NWO-Vici grant, proposal/contract 91819627; www.zonmw.nl), both to DvdB. Work at the Wellcome Trust Sanger Institute was supported by Wellcome Trust core funding (098051; https://wellcome.ac.uk). JAL was funded by Wellcome [219699], and received support from the Medical Research Council (grant number MR/R015600/1). This award is jointly funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement and is also part of the EDCTP2 program supported by the European Union. PT was funded in part by the Wellcome Trust [Grant number 083735/Z/07/Z]. The Netherlands Reference Laboratory for Bacterial Meningitis was supported by the National Institute for Health and Environmental Protection, Bilthoven (www.rivm.nl). For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The pneumococcal conjugate vaccine (PCV) primarily reduces disease burden in adults through a reduction in carriage prevalence of invasive serotypes in children. Current vaccine formulations are the same for both adults and children, but tailoring these formulations to age category could optimize vaccine efficacy. Identification of specific pneumococcal genetic factors associated with carriage in younger or older age groups may suggest alternative formulations and contribute to a better mechanistic understanding of immunity. Here, we used whole genome sequencing to dissect pneumococcal variation associated with age. We performed genome sequencing in a large carriage cohort, and conducted a meta-analysis with an existing carriage study. We compiled a dictionary of pathogen genetic variation including serotype, sequence cluster, sequence elements, SNPs, burden combined rare variants, and clusters of orthologous genes (COGs) for each cohort – all of which used in a genome-wide association with host age. Age-dependent colonization had some heritability, though this varied between cohorts (h^2 = 0.10, 0.00 – 0.69 95% CI in the first; h^2 = 0.46, 0.33 – 0.60 95% CI in the second cohort). We found that serotypes and genetic background (strain) explained most of the heritability in each cohort ($h^2_{serotype}$ = 0.06 and h^2_{GPSC} = 0.04 in the first; $h^2_{serotype}$ = 0.20 and h^2_{GPSC} = 0.23 in the second cohort). We found one candidate association ($p = 1.2x10^{-9}$) upstream of an accessory Sec-dependent serine-rich glycoprotein adhesin. Overall, association with age was highly cohort and strain dependent, supporting proposals for a future vaccination strategy which is primarily targeted using serotypes rather than proteins, and is tailored towards specific pathogen populations.

Introduction

Streptococcus pneumoniae is a common commensal of the human upper respiratory tract and nasopharynx, but can also cause invasive diseases such as pneumonia, sepsis or meningitis.(1) Invasive pneumococcal disease (IPD) has a high mortality, and the overall mortality rate from IPD is higher in extreme age ranges, such as infants and the elderly.(2,3) In the Netherlands, pneumococcal carriage rates are higher in children than in adults, with a prevalence of up to 80% at 2 years of age.(4)

Host age is known to affect carriage prevalence and carriage duration of different serotypes(5,6), which is suggested to be driven by differences in immunity.(7) Studies in mice and humans showed evidence for age-dependent host-pathogen interactions involving interleukin (IL)-1 response in reaction to the pore-forming pneumoplysin (*ply*) toxin.(8) IgA secretion is important in clearing *S. pneumoniae* from host upper respiratory tract mucosa and this secretion more effective in previously exposed individuals, the adults.(9) Bacterial genetics has shown to explain over 60% of the variability in carriage duration, and specifically that presence of a bacteriophage inserted in a mediator of genomic competence was associated with a decreased carriage duration.(10)

Pneumococci are highly genetically variable, displaying over 100 diverse capsular serotypes(11), which are a major antigen and the strongest single predictor of carriage prevalence.(12) Pneumococcal conjugate vaccines, targeting up to thirteen capsule serotypes with high burden of invasive disease, cause decreased the rate of nasopharyngeal carriage and invasive disease.(13,14) Besides a direct effect of vaccination with an pneumococcal conjugate vaccine (PCV) on the disease burden in the target population, i.e. young children, it also reduces the disease burden caused by pneumococci with vaccine serotypes in the population not eligible for vaccination through indirect protection from colonization – reducing carriage rates in children reduces overall transmission of the most invasive serotypes.(12,15,16) However, the introduction of PCV has resulted in the replacement of serotypes not covered by the vaccine(17,18), which in some countries reaches levels of invasive disease return towards pre-vaccine levels.(19,20)

As not all serotypes can be included in a conjugate vaccine, three perspectives will lead to improved pneumococcal vaccination have been proposed: whole-cell vaccines(21,22), protein vaccines(23), or changing components in the conjugate vaccine in response to the circulating population.(24) Whole-cell vaccination trials are ongoing, but efficacy remains unproven in human populations.(25) Protein vaccines contain antigens which illicit a strong mucosal immune response, with their targets chosen to be common or conserved in the target population, and ideally reducing onward transmission.(26) In their current form, protein vaccines are not thought to be effective on their own, but if administered with serotype conjugates they may help to reduce serotype replacement. Detailed modelling of the dynamics of pneumococcal population genetics has shown that targeting these vaccines towards serotypes prevalent in specific populations would likely be a superior strategy. This work further shows that providing age-specific vaccine design, using complementary adult-administered vaccines (CAVs) is predicted to have the greatest effect on total IPD burden.(24)

For a future pneumococcal vaccination strategy based on age, we should understand the differences between infant and adult carriage. Differences between host niches have been found, some with a potential effect on onward transmission.(27–29) Treating age as a niche in a systematic analysis of genetic variation would therefore be a powerful way to discover useful vaccine targets. Here, we aim to determine how pathogen variation affects colonization. The identification of age-associated genetic variation, could provide further targets for protein vaccination, whereas ruling these out could provide confidence in age-targeted vaccine formulations based on serotype differences alone.

We performed a pathogen genome-wide association study on pneumococci isolated from nasopharyngeal swabs of 4320 infants and adults from the Netherlands and Myanmar. To dissect pneumococcal variation associated with age we compare prevalence of pneumococcal strains and serotypes between infants and adults in different settings, calculate the contribution of pathogen genetic variation towards predilection for host age, and search for genetic regions associated with host age.

Methods

Cohort collection

The Dutch cohort consists of carriage samples from individuals obtained from three prospective carriage surveillance studies.(30–32) In these studies, carriage was assessed by conventional culture of nasopharyngeal or oropharyngeal swabs of vaccinated children (11 and 24 months of age) and their parents in 2009, in 2010/2011, in 2012 and 2013.(30) All children were vaccinated with PCV-7 or PHiD-CV10 according to the Dutch national immunization program at 2, 3, 4 and 11 months of age. Vaccination status of the parents was unknown. Exclusion criteria are described elsewhere.(30,31) Nasopharyngeal swabs were collected from all individuals and oropharyngeal swabs were collected from all adult subjects by trained study personnel using flexible, sterile swabs according to the standard procedures described by the World Health Organization.(33) After sampling, swabs were immediately placed in liquid Amies transport medium and transported to the microbiology laboratory at room temperature and cultured within 12 hours. Pneumococcal isolates were identified using conventional methods, as described previously.(34) The Maela cohort consists of samples from people from a camp for displaced persons on the Thailand-Myanmar border, where monthly nasopharyngeal sampling was performed in unvaccinated children (0 to 24 months old) and their mothers. Procedures for collecting samples and generating whole genome sequences have been previously described.(6,35)

Informed consent

Written informed consent was obtained from both parents of each child participant and from all adult participants. Approval for the 2009 and 2012/2013 studies in children and their parents (NL24116 and NL40288/NTR3613) were received from the National Ethics Committee in the Netherlands (CCMO and METC Noord-Holland). For the 2010/2011 study, a National Ethics Committee in The Netherlands (STEG-METC, Almere) waived the requirement for EC approval. Informed consent for the Maela cohort was described elsewhere.(6) Studies were conducted in accordance with the European Statements for Good Clinical Practice and the Declaration of Helsinki of the World Medical Association.

Host age distribution in sequenced carriage cohorts

In the Dutch cohort, children had a median age of 23 months (interquartile range (IQR) 10 – 24 months). Adults had a median age of 35 (IQR 32 – 38) years. In the Maela cohort, the median age of children was 13 months (IQR 6 – 19 months), and for mothers (women of childbearing age) the exact age was unknown (Supplementary Figure S1).(6,36) In the Dutch cohort, all children were vaccinated with PCV-7 or PHiD-CV10. None of the members of the Maela cohort had received PCV.

DNA extraction and whole genome sequencing

For the Dutch cohort, DNA extraction was performed with the Gentra Puregene Isolation Kit (Qiagen), and quality control procedures were performed to determine yield and purity. Sequencing was performed using multiplexed libraries on the Illumina HiSeq platform to produce paired end reads of 100 nucleotides in length (Illumina, San Diego, CA, USA). Quality control involved analysis of contamination with Kraken (version

1.1.1)(37), number and length of contigs, GC content and N50 parameter. Sequences for which one or more of these quality control parameters deviated by more than 3 standard deviations from the mean were excluded. Sequences were assembled using a standard assembly pipeline.(38) Assembly statistics can be found in the Supplementary Table S1. Genome sequences were annotated with PROKKA, version 1.11.(39) For the Maela cohort, DNA extraction, quality control and whole genome sequencing have been described elsewhere.(40) Serotypes were determined from the whole-genome sequence by in-house scripts.(41) Sequence clusters (strains) were defined as Global Pneumococcal Sequence Clusters (GPSC) using PopPUNK (version 2.2.0), using a previously published reference database.(42,43) For 114 and 401 sequences in the Dutch and Maela cohorts respectively, the GPSC couldn't be inferred due to low sequence quality.

Sequencing characteristics and quality control

A total of 1361 bacterial isolates were sequenced as part of the Dutch cohort. During quality control, 32 sequences were excluded. Of these, 8 belonged to a different pathogen species, 9 had contamination, 14 were excluded based on number of contigs or genome length and 1 sequence failed annotation. For 47 sequences, host age was missing. The association analyses were performed on 1282 sequences in the Dutch cohort. Of these, 1052 were isolated from children and 230 from adults. There were 3085 sequences available from the Maela cohort. Quality control for this cohort was described previously.(40) There were 2503 sequences isolated from children and 582 from adults. For the determination of the frequency and odds ratio of serotype and GPSCs in children and adults, only the first isolate from each carriage episode for each child was included in the analysis. This resulted in 964 serotypes and 799 GPSCs (165 missing) in children, and 582 serotypes and 508 GPSCs (74 missing) in adults. For adults Chi-squared tests to calculate the p-value for association between serotype and strain with age were performed in R (version 4.0.0).

Data availability

Fastq sequences of bacterial isolates from the Dutch cohort were deposited in the European Nucleotide Archive (ENA, study and accession numbers in Supplementary Table S2). Sequences of bacterial isolates in the Maela cohort are available at ENA under study numbers ERP000435, ERP000483, ERP000485, ERP000487, ERP000598 and ERP000599 (Supplementary Table S3).

Phylogenetic tree

A core genome for sequences from both cohorts together was generated with Roary (version 3.5.0, default parameters), using a 95% sequence identity threshold.(44) A maximum likelihood phylogeny of single-nucleotide polymorphisms (SNPs) in the core genome of all sequenced isolates from both cohorts together was produced with iqtree (version 1.6.5, including fast stochastic tree search algorithm, GTR+I+G) assuming a general time reversible model of nucleotide substitution with a discrete γ-distributed rate heterogeneity and the allowance of invariable sites.(45)

Heritability analysis

Based on the kinship matrix and phenotypes, a heritability estimate was performed in limix (version 3.0.4 with default parameters) for both cohorts separately.(46) A confidence interval around the heritability estimate was determined with Accurate LMM-based heritability Bootstrap confidence Intervals (ALBI) based on the eigenvalue decomposed distances in the kinship matrix and the heritability estimate with the gglim package (version 0.0.1) in R (version 4.0.0).(47) To estimate the proportion of heritability attributable to serotype or strain alone, we calculated the heritability based on a kinship matrix treating serotypes or strains as genetic variants.(10,48) This analysis was performed in Pyseer (version 1.1.1) using the linear mixed model.(48)

Determining bacterial genetic variation – unitigs, SNPs and COGs

Using the whole-genome sequence reads from both cohorts, we called SNPs, small insertions and deletions and SNPs clustered as rare variants (deleterious variants at an allele frequency < 0.01) based on the *S. pneumoniae* D39V reference (CP027540) sequence using the Snippy pipeline (version 4.4.0, default parameters). We determined non-redundant sequence elements (unitigs) from assembled sequences in the Dutch cohort by counting nodes on compacted De Bruijn graphs with Unitig-counter (version 1.0.5, default minimum k-mer length of 31).(49) These unitigs were called in an indexed set of sequences from the Maela cohort with Unitig-caller (version 1.0.0, default parameters).(50) This gave us the distribution of sequences from both cohorts with consistent k-mer definitions, making it possible to run predictive models across cohorts. The same Roary run as was used to generate the core-genome alignment was used to extract accessory clusters of orthologous genes (COGs).(44)

There were 966794 unitigs counted from combined sequences in the Dutch cohort. Of these, 303901 passed a minor allele frequency (MAF) of 0.05 filter and had association testing performed. The 9966794 unitigs from the Dutch cohort were called in sequences from the Maela cohort, to obtain 726040 unitigs. Association testing in this group was done for 323112 unitigs which were present at MAF 0.05 or more. Meta-analysis was performed on 251733 overlapping unitigs. There were 313143 SNPs called from sequences in the Dutch cohort, of these 43556 passed MAF filtering. For the Maela cohort, 382230 SNPs were called and 53553 passed the MAF filter. For meta-analysis, 20118 SNPs had overlapping positions and were included. There were 1997 rare variants called in the Dutch cohort, which were burdened in 538 genes. For the Maela cohort, these numbers were 1997 and 423. Together, 186 genes were included in the meta-analysis. Lastly, 2348 COGs were analyzed in the Dutch cohort and 4678 in the Maela cohort. In the meta-analysis there were 627 overlapping COGs.

Genome-wide association study

The association analysis for SNPs, unitigs, rare variants and COGs was run as a linear mixed model in Pyseer (version 1.1.1), with a minimum minor allele frequency of 0.05.(48) To correct for population structure, the model included a kinship matrix as covariates, which was calculated from the midpoint rooted phylogenetic tree. An association analysis not corrected for population structure was run with unitigs as sequence elements using a simple fixed effects model in Pyseer. Rare variants were clustered in their corresponding gene and

analyzed in a burden test. Meta-analysis was performed on summary statistics from the Pyseer results files with METAL (version released on August 28 2018, default parameters) for each variant.(51) A threshold for association of the phenotype with meta-analyzed variants was determined using a Bonferroni correction with alpha < 0.05 and the number of independent tests in the Dutch cohort, giving: $p < 1.0x10^{-7}$ for unitigs, $p < 1.0x10^{-6}$ for SNPs, $p < 2.0x10^{-5}$ for COGs and $p < 1.0x10^{-4}$ for rare variants. Unitigs were mapped to the *S. pneumoniae* D39V reference genome with bowtie-2 (version 2.2.3, with equal quality values and length of seed substrings 7 nucleotides). In accordance with the study populations in both cohorts, the phenotype was dichotomized as host age 0 to 24 months versus adult age (Supplementary Figure S1). Manhattan plots were generated in R, version 3.5.1, with the package ggplot2 (version 3.1.0). Presence or absence of pilus genes was detected by nucleotide BLAST (version 2.6.0, default parameters) analysis. Pilus gene presence association to carriage age was calculated with a likelihood-ratio test in Pyseer (version 1.1.1), corrected for population structure by including a kinship matrix as covariates.

The prediction analysis used the elastic-net mode of Pyseer. This fitted an elastic net model with a default mixing parameter (0.0069 L1/L2) to the unitigs counted in each cohort, using the strains from PopPUNK as folds to try and reduce overfitting.(50) ROC curves for each cohort were drawn using the linear link output, with the R package pROC (version 1.16.2) using smoothing. To test inter-cohort prediction, the called unitigs from the other cohorts were used as predictors with the model from the opposing cohort.

Results

We first analyzed the observed distribution of serotypes and strains in each of the two cohorts, to assess overall trends of differences between adults and children, and look at the genetic heterogeneity between the two cohorts. Although our cohorts were broadly matched in the primary phenotype, age, large differences between the pathogen population are expected due to different geographies, social backgrounds, and only children in one cohort being vaccinated.

Serotypes and strains are variable between age groups, and between cohorts

The Dutch cohort was made up of 1329 *S. pneumoniae* isolates comprising 41 unique serotypes (Figure 1, Supplementary Table S4). Of these isolates, 689 (52%) comprised 7 serotypes: 19A (225; 17%), 11A (111; 8%), 6C (97; 7%), 23B (84; 6%), 10A (67; 5%), 16F (54; 4%) and serotype 21 (51; 4%). In this cohort of which the children were vaccinated, a minority of isolates belonged to one of the vaccine serotypes, namely serotype 1 (6; <1%), serotype 4 (1; <1%), serotype 5 (1; <1%), serotype 6B (26; 2%), serotype 7F (11; 1%), serotype 9V (4; <1%), serotype 14 (4; <1%), serotype 18C (5; <1%), serotype 19F (31; 2%) and serotype 23F (12; 1%). The 3085 pneumococcal isolates of the Maela cohort comprised 64 unique serotype groups (Supplementary Table S5). Of these isolates, 1631 (53%) comprised five serotypes: Non-typable (511; 17%), 19F (402, 13%), 23F (307, 10%), 6B (236; 8%) and serotype 14 (175; 6%). In the Dutch cohort, there were 59 unique sequence clusters of which the four largest sequence clusters were GPSC 4 (171; 13%), GPSC 3 (156; 12%), GPSC 7 (131; 10%) and GPSC 11 (119; 9%) (Supplementary Table S6). The. There were 127 unique sequence clusters found in the Maela cohort (Supplementary Table S7). The four largest sequence clusters were GPSC 1 (352; 13%), GPSC 28 (190; 7%), GPSC 20 (168; 6%) and GPSC 42 (123; 5%).

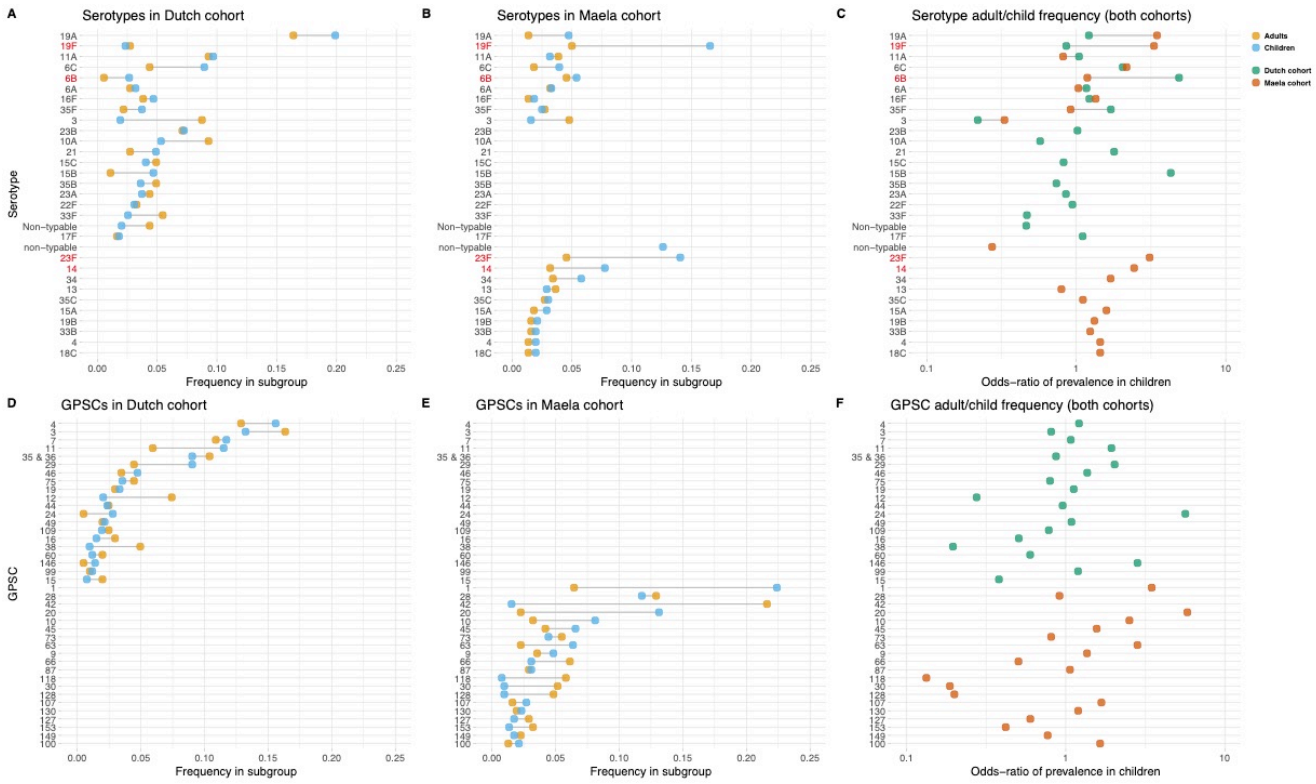

Figure 1. Serotype and strain (GPSC) distribution by age, and between cohorts. Blue dots represent frequency of serotype and strain in child carriage, yellow dots represent frequency in adult carriage. Red and green dots show odds-ratio of prevalence in children in the Dutch and Maela cohorts respectively, on a log scale for serotype. Lines show differences. Top row: dominant serotypes, ordered by presence in cohort, and internally by overall frequency. Vaccine serotypes shown in red. A: Serotype frequency in the Dutch cohort. B: Serotype frequency in the Maela cohort. C: Comparison of adult/child log odds in each cohort for serotype. Second row: dominants strains (GPSCs), ordered by presence in cohort, and internally by overall frequency. D: Strain frequency in Dutch cohort. E: Strain frequency in Maela cohort. F: Comparison of adult/child log odds in each cohort for strain.

Some serotypes exhibited a large difference in colonization frequency between the two age groups. In the Dutch cohort, serotype 6C was overrepresented in children relative to adults (chi-squared test, $p = 0.02$, not corrected for multiple testing), while in the Maela cohort, serotype groups overrepresented in children were serotype 19F (chi-squared test, $p = 3.1\times10^{-9}$), serotype 23F (chi-squared test, $p = 1.8\times10^{-7}$), serotype 14 (chi-squared test, $p = 1.3\times10^{-3}$); while non-typeable serogroup was overrepresented in adults (chi-squared test, $p < 1.0\times10^{-15}$) (Table 1). None of the 20 largest groups of sequence clusters overlapped between the cohorts. In the Dutch cohort only GPSC 11 was significantly associated with carriage in children (chi-squared test, $p = 0.03$, not corrected for multiple testing). In the Maela cohort, sequence clusters overrepresented in children were GPSC 1 (chi-squared test, $p = 1.8\times10^{-9}$) and GPSC 20 (chi-squared test, $p = 1.2\times10^{-7}$); while GPSC 42 (chi-squared test, $p < 1.0\times10^{-15}$), GPSC 118 (chi-squared test, $p = 7.9\times10^{-5}$), GPSC 30 (chi-squared test, $p = 9.3\times10^{-4}$) and GPSC 128 (chi-squared test, $p = 1.9\times10^{-3}$) were overrepresented in adults (Table 2).

A phylogenetic tree of pooled sequences from both cohorts, with serotype, sequence cluster, age group and cohort for each sequence, revealed clonal discrimination between cohorts (Figure 2). Combined with the effects shown in Figure 1, this highlighted a key feature of our analysis of these datasets, which was the genetic heterogeneity between the two cohorts. Individually, each dataset clearly has strains and serotypes with strong signals of host age differences, but the overall makeup of each dataset is very different (twelve common serotypes are shared, but only a single common GPSC), and where there are shared serotypes many have different effect directions between the two cohorts.

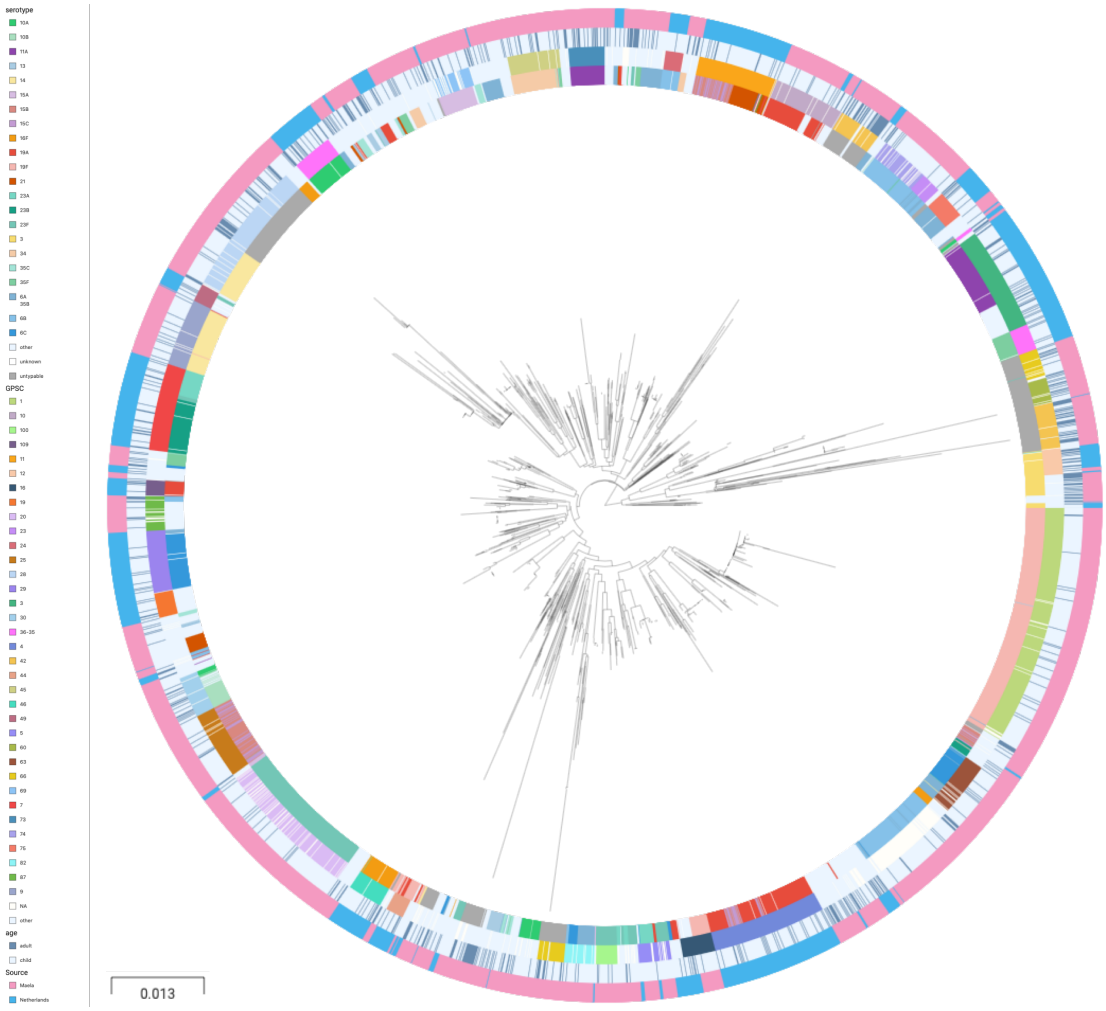

Figure 2. Phylogenetic tree of carriage samples from both cohorts. The rings show metadata for the samples. Depicted from inside to outside, these are serotype, sequence cluster (GPSC), age and source (Maela, Netherlands). Scale bar: 0.013 substitutions per site. An interactive version is available at https://microreact.org/project/f2MdBLZhSyU9eF8MBobHhA/e2a5ebd7 (project link https://microreact.org/project/f2MdBLZhSyU9eF8MBobHhA).

Host age is heritable and mostly explained by strain and serotype

To quantify the amount of variability in carriage age explained by variability in the genome, we calculated a heritability estimate (h^2) for each cohort. For isolates in the Dutch cohort, we did not find strong evidence that genetic variability in bacteria was related to variance in host age (h^2 = 0.10, 0.00 – 0.69 95% CI). In the Maela cohort, we found significant evidence that affinity with host age was heritable (h^2= 0.46 0.33 – 0.60 95% CI) and thus genetic variation in this cohort explained variation in carriage age to a greater degree. In both cohorts pan-genomic variation could be used to predict host age to some degree of accuracy (area under the ROC curve 0.82 [Dutch cohort]; 0.91 [Maela cohort]), suggestive of some level of heritability and association of host age with strain (Figure 3). Prediction between cohorts using a simple linear model failed, as the genetic variants chosen as predictors were not found in the other cohort – again highlighting the high level of genetic heterogeneity between cohorts.

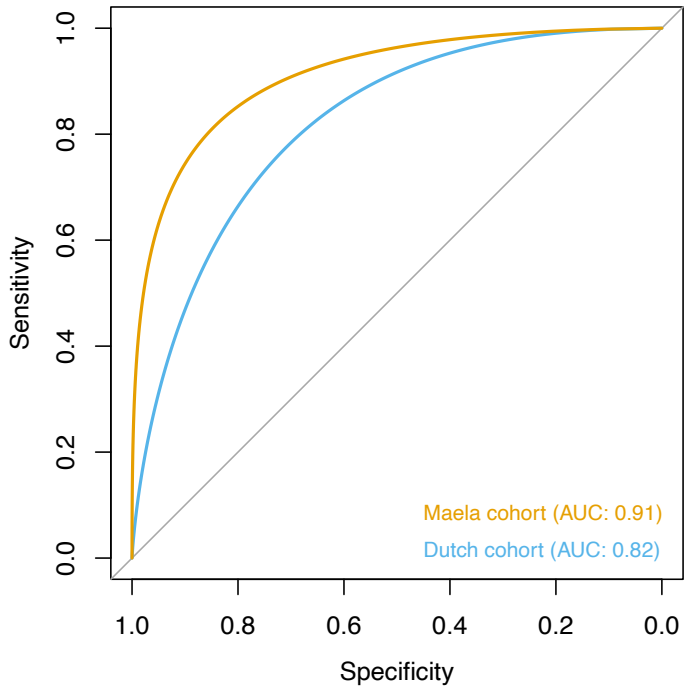

Figure 3. Prediction of host age from pan-genomic variation in each cohort. The smoothed ROC curve based on a linear predictor (elastic net fitted to unitigs, with strains used as folds for cross-validation) is shown. Area under the curve (AUC) is 0.5 for no predictive ability and 1 for perfect prediction.

To further investigate the association of serotype and sequence cluster to carriage age, we determined the proportion of variation in carriage age explained by serotype and sequence cluster alone. Here, we estimated $h^2_{serotype}$ = 0.06 and h^2_{GPSC} = 0.04 for the Dutch cohort and $h^2_{serotype}$ = 0.20 and h^2_{GPSC} = 0.23 for the Maela cohort, confirming the larger contribution of serotype and sequence cluster to carriage age in sequences from the Maela cohort. We also performed a genome-wide association analysis, but without controlling for population structure. This reveals genetic variants specific to serotype as determinants for carriage age (p-values < 5.0×10^{-8}) in both cohorts (Supplementary Table S8 and Supplementary Table S9). Among the genetic variants with the lowest p-values were variants in capsule locus genes (Cps) in both cohorts. This further supports a role of strain and serotype in association with host age, but does not distinguish between the two.

Genome-wide association analysis does not find genetic variants independent of strain

Following these observations that serotype and strain do not explain the full heritability, we performed a pathogen genome-wide association analysis to investigate whether we can detect genetic variants irrespective of the genetic background that are associated with carriage in children or adults. Though the cohorts have little genetic overlap in terms of genetic background, we would be well-powered to detect genetic variation independent of background ('locus' associations).(52,53) In the Dutch cohort, none of the unitigs, SNPs, COGs or rare variants surpassed the threshold for multiple testing correction (Supplementary Figure 1). The burden (sum) of rare variants in a gene for tryptophan synthase, *trpB*, approach the multiple testing threshold, but was not significant. In the Maela cohort unitigs in the *ugpA* gene surpassed the threshold for statistical

significance (Supplementary Figure 2). After meta-analysis, there were two hits which surpass the threshold for statistical significance (Figure 4).

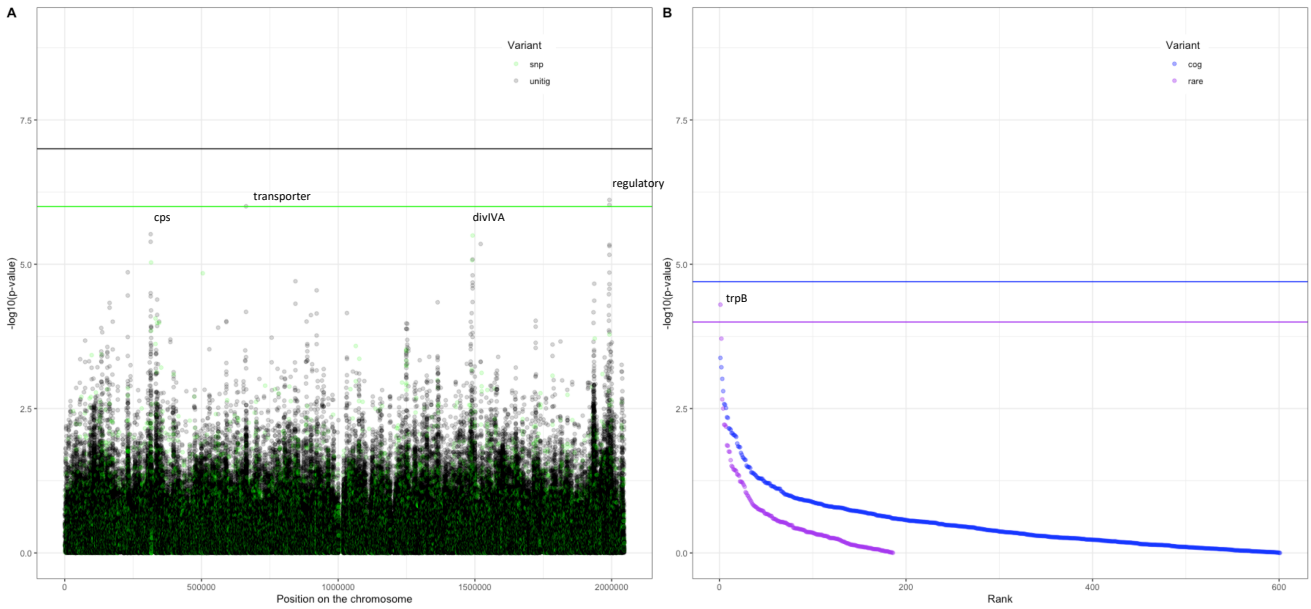

Figure 4. Association of variants after meta-analysis with carriage age 0 – 24 months, A. minus log transformed p-value on the y-axis and position of unitig and snp variants on the *S. pneumoniae* genome on the x-axis (Manhattan plot); B. minus log transformed p-value on the y-axis and sorted lowest to highest p-value for rare variant burden in genes (purple) and clusters of orthologous genes (COGs, blue) on the x-axis.

The first, a nucleotide sequence marked by multiple unitigs of which the lowest has a p-value of 1.2×10^{-9} (Supplementary Table S10). This sequence does not map to the *S. pneumoniae* D39V reference sequence.(54) For this reason, it is not visualized on the Manhattan plot, for which unitigs were mapped to the *S. pneumoniae* D39V reference sequence. Upon inspection of the individual sequences these unitigs are called from, we find them to map in the intergenic region between open reading frames encoding the accessory Sec-dependent serine-rich glycoprotein adhesin and a MarR like regulator, respectively. This region contains sequences resembling transposable elements and an open reading frame encoding a transposase. The unitigs map upstream of the start codon of the accessory Sec-dependent serine-rich glycoprotein adhesin. The sequence is present in 169 out of 1282 (13%) sequences in the Dutch cohort and in 241 out of 3085 (8%) in the Maela cohort. The sequence is present in isolates dispersed over the phylogenetic tree, and associated with carriage in children (Supplementary Figure S4). This protein is involved in adhesion to epithelial cells and biofilm formation.(55–57) Given that this sequence lies just upstream of the start codon, it is plausible that variation of this sequence alters the expression of the Sec-dependent adhesin protein, and therefore affects carriage.

The second hit is a burden of rare variants in a gene for tryptophan synthase, *trpB*, that surpass the threshold for statistical significance at a p-value of 5.0×10^{-5}. The variants are 2 frameshift variants of very low frequency. These result in a predicted dysfunctional *trpB* gene in 9 out of 1282 (1%) sequences in the Dutch cohort and in

12 out of 3073 (0.4%) sequences in the Maela cohort. This association of the *trpB* gene is likely to be an artefact of low allele frequency, as we estimate we are only powered to detect variation in at least 5% of isolates.

Pilus gene presence does not determine carriage age independent of genetic background

Finally, we investigated whether pneumococcal isolates containing a pilus gene preferentially colonize children in the Dutch cohort, as has been previously described in the Maela cohort.(9,58) This study analyzed the Maela cohort and found that 934 out of 2557 (37%) isolates in children versus 95 out of 592 (16%) isolates in adults had pilus genes present. However, this association of pilus gene presence to carriage age was dependent on lineages within the population.(9) In the Dutch cohort, we found no evidence that host age was dependent on pilus gene presence (22 out of 208 (10%) in adults versus 129 out of 1099 (12%) in children). This was the case whether or not the genetic background was adjusted for ($p = 0.35$, uncorrected for population structure and $p = 0.69$, corrected for population structure). Based on these findings, we suggest that the previously reported pilus-IgA1 association is not a universal explanation for difference in colonization between hosts of different ages.

Discussion

The age of the host is known to have an important effect on pneumococcal colonization.(12) Observational studies have demonstrated variation in serotype prevalence and carriage duration between infants and adults. Mechanistic studies in mice and humans have shown examples of differing immune responses depending both on host factors and pathogen factors. Findings from these studies include the observation that capsular polysaccharides (determinants of serotype) inhibit phagocytic clearance in animal models of upper respiratory tract colonization.(59) A pneumolysin-induced IL-1 response determined colonization persistence in an age-dependent manner(8); and pilus expressing strains were found to preferentially colonize children, because of immune exclusion via secretory IgA in non-naïve hosts.(9)

Building upon these observations, we sought to investigate and quantify the contribution of pathogen genetic variation to carriage in infant versus adult hosts, using a top-down approach. Through whole genome sequencing and application of statistical genetic methods to two large *S. pneumoniae* carriage cohorts, we show evidence that bacterial genetic variability influences predilection for host age, though this appears to be highly variable between populations. One important difference between our study cohorts was that children from the Dutch cohort were vaccinated, while children from the Maela cohort were not. While our findings demonstrate that vaccinated versus unvaccinated children were colonized with different bacterial serotypes and different sequence clusters, we observed differences in prevalence beyond just the serotypes included in the vaccine. Another difference between the cohorts was that adults from the Dutch cohort were males and females, while adults from the Maela cohort were female only.

Strain, or genetic background, appears to be the main effect, explaining roughly half of the total heritability in each cohort. We found sequences in one region which map closely to the start codon of the accessory Sec-dependent serine-rich glycoprotein adhesin to be associated with carriage age independent of genetic background, in a meta-analysis of the two cohorts. Our results our suggestive of a polygenic architecture of many variants with low effect sizes, along with larger effects between strains. Three reasons can contribute to this: the proportion of the heritability which is caused by lineage effects; rare locus effects which could not be detected with the current sample size; and by sampling from a cohort with vaccinated children and unvaccinated adults and comparing with a cohort of unvaccinated children and adults, we had lower power due to the reduced overlap within and between cohorts in pan-genome content. Although differences in vaccination status between cohorts is a plausible explanation for our findings, we were unable to rule out other factors, for example a population-specific host effect, or the broad effects of different socio-economic status between these cohorts.

In previous bacterial GWAS studies of antimicrobial resistance (such as a single gene which causes antibiotic resistance), large monogenic effects have typically been found to have high heritabilities close to one, and the GWAS identify the causal variant precisely.(50,52,60) When applied to virulence and carriage duration phenotypes, heritable effects have also been found, but these only explained some of the variation in the

phenotype. These appeared to be caused by weaker polygenic effects, not all of which could be detected using the relatively small cohorts available.(10,61) We found similar results for host-age heritability in these two cohorts. Within this genetic architecture, our finding of a sequence just upstream of the start codon of the accessory Sec-dependent serine-rich glycoprotein adhesin being associated with carriage in children, is notable.

We could not distinguish between genetic background or serotype being the primary effect due to their correlation. We did note a difference in effect size of serotype between the two cohorts, which may make it unlikely to be the single largest effect on host age. This difference in cohorts could be explained by strain/GPSC being the main and consistent effect on host age. As strains are different between cohorts and each serotype appears in multiple strains, combining them in different amounts would create different directions of effect for serotype. We did not replicate the association of piliated genomes in infant hosts in our newly sequenced cohort, further demonstrating important differences between populations.

In summary, we found an effect of pneumococcal genetics on carriage in children versus adult hosts, which varies between cohorts, and is likely primarily driven by strain (lineage) effects rather than large population-wide effects of individual genes. An important corollary of our work is on future pneumococcal vaccine optimization efforts. A promising approach for future vaccination strategies is to target the different age groups.(24) Whether these should consist of the dominant disease-causing serotypes overrepresented in carriage by each age group, or whether there are age-specific pathogen proteins that should be included is an open question. Our study suggests that targeting these age groups using serotype makeup alone would be sufficient, and supports previous observational and modelling studies which advise targeting the serotype makeup in the vaccine at specific populations to maximize their effect.

Acknowledgments

We would like to thank Dr. Nicholas Croucher from Imperial College London for commenting on the manuscript.

Table 1. Chi-squared values for serotypes in the Dutch and Maela cohorts and the age group that the serotype is affiliated with

Serotype	Dutch cohort		Maela cohort	
	χ^2 p-value	Age group	χ^2 p-value	Age group
19A	0.089	Children	$2.6x10^{-3}$	Children
Non-typeable	0.188	Adults	$<1.0x10^{-15}$	Adults
19F	1	Adults	$3.1x10^{-9}$	Children
11A	0.591	Children	0.728	Adults
6B	0.099	Children	0.079	Children
6C	0.022	Children	0.049	Children
6A	0.709	Children	0.952	Children
10A	0.155	Adults	-	-
15B	0.023	Children	-	-
15C	1.000	Adults	-	-
35F	0.279	Children	1.00	Adults
3	$2.5x10^{-5}$	Adults	$3.4x10^{-3}$	Adults
23B	0.727	Children	-	-
21	0.166	Children	-	-
16F	0.519	Children	-	-
35B	0.775	Adults	-	-
23A	1	Adults	-	-
22F	1	Adults	-	-
33F	0.132	Adults	-	-
17F	0.943	Children	-	-
23F	-	-	$1.8x10^{-7}$	Children
14	-	-	$1.3x10^{-3}$	Children
34	-	-	0.066	Children
13	-	-	0.686	Adults
35C	-	-	0.812	Children
15A	-	-	0.288	Children
19B	-	-	0.612	Children
33B	-	-	0.729	Children
18C	-	-	0.524	Children
4	-	-	0.524	Children
16F	-	-	0.633	Children

χ^2, chi-square; -, not applicable

Table 2. Chi-squared values for strains in the Dutch and Maela cohorts and the age group that the strain is affiliated with

GPSC	Dutch cohort		Maela cohort	
	χ^2 p-value	Age group	χ^2 p-value	Age group
60	0.568	Adults	-	-
4	0.298	Children	-	-
3	0.392	Adults	-	-
7	0.858	Children	-	-
11	0.03	Children	-	-
35 & 36	0.617	Adults	-	-
29	0.049	Children	-	-
46	0.563	Children	-	-
75	0.666	Adults	-	-
19	0.978	Children	-	-
12	$1.2x10^{-4}$	Adults	-	-
44	1	Adults	-	-
24	0.094	Children	-	-
49	1	Children	-	-
109	0.817	Adults	-	-
16	0.249	Adults	-	-
38	$2.1x10^{-4}$	Adults	-	-
146	0.489	Children	-	-
99	1	Children	-	-
15	0.22	Adults	-	-
1	-	-	$1.8x10^{-9}$	Children
28	-	-	0.959	Adults
20	-	-	$1.3x10^{-7}$	Children
42	-	-	$<1.0x10^{-15}$	Adults
10	-	-	0.005	Children
45	-	-	0.1462	Children
63	-	-	0.008	Children
73	-	-	0.754	Adults
9	-	-	0.388	Children
66	-	-	0.086	Adults
87	-	-	0.928	Children
118	-	-	$7.9x10^{-5}$	Adults
30	-	-	$9.3x10^{-4}$	Adults
128	-	-	$1.9x10^{-3}$	Adults
107	-	-	0.372	Children
127	-	-	0.464	Adults
130	-	-	0.809	Children
153	-	-	0.147	Adults
149	-	-	0.885	Adults
100	-	-	0.479	Children

χ^2, chi-square; -, not applicable

References

1. Bogaert D, van Belkum A, Sluijter M, Luijendijk A, de Groot R, Rümke HC, et al. Colonisation by Streptococcus pneumoniae and Staphylococcus aureus in healthy children. Lancet. 2004;363(9424):1871–2.
2. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, 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. The Lancet Global Health. 2018;6(7):e744–57.
3. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, et al. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. The Lancet. 2009;374(9693):893–902.
4. Wyllie AL, Wijmenga-Monsuur AJ, van Houten MA, Bosch AATM, Groot JA, van Engelsdorp Gastelaars J, et al. Molecular surveillance of nasopharyngeal carriage of Streptococcus pneumoniae in children vaccinated with conjugated polysaccharide pneumococcal vaccines. Scientific Reports. 2016;4(6):23809.
5. Stearns JC, Davidson CJ, Mckeon S, Whelan FJ, Fontes ME, Schryvers AB, et al. Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age. ISME Journal. 2015;9(5):1246–59.
6. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, et al. A longitudinal study of streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. PLoS ONE. 2012;7(5):e38271.
7. Wyllie AL, Warren JL, Regev-Yochay G, Givon-Lavi N, Dagan R, Weinberger DM. Serotype Patterns of Pneumococcal Disease in Adults Are Correlated With Carriage Patterns in Older Children. Clinical Infectious Diseases. 2020;ciaa1480.
8. Kuipers K, Lokken KL, Zangari T, Boyer MA, Shin S, Weiser JN. Age-related differences in IL-1 signaling and capsule serotype affect persistence of Streptococcus pneumoniae colonization. PLoS Pathogens. 2018;14(10):e1007396.
9. Binsker U, Lees JA, Hammond AJ, Weiser JN. Immune exclusion by naturally acquired secretory IgA against pneumococcal pilus-1. Journal of Clinical Investigation. 2020;130(2):927–41.
10. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. eLife. 2017;(6):e26255.
11. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, et al. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large cps fragment from an oral streptococcus. mBio. 2020;11(3):e00937-20.
12. Croucher NJ, Løchen A, Bentley SD. Pneumococcal Vaccines: Host Interactions, Population Dynamics, and Design Principles. Annual Review of Microbiology. 2018;(72):521–49.
13. Whitney CG, Farley MM, Hadler J, Harrison LH, Bennett NM, Lynfield R, et al. Decline in Invasive Pneumococcal Disease after the Introduction of Protein–Polysaccharide Conjugate Vaccine. New England Journal of Medicine. 2003;348(18):1737–46.
14. Poehling KA, Talbot TR, Griffin MR, Craig AS, Whitney CG, Zell E, et al. Invasive pneumococcal disease among infants before and after introduction of pneumococcal conjugate vaccine. Journal of the American Medical Association. 2006;295(14):1668–74.
15. Desai AP, Sharma D, Crispell EK, Baughman W, Thomas S, Tunali A, et al. Decline in pneumococcal nasopharyngeal carriage of vaccine serotypes after the introduction of the 13-valent pneumococcal conjugate vaccine in children in Atlanta, Georgia. Pediatric Infectious Disease Journal. 2015;34(11):1168–74.
16. von Gottberg A, de Gouveia L, Tempia S, Quan V, Meiring S, von Mollendorf C, et al. Effects of Vaccination on Invasive Pneumococcal Disease in South Africa. New England Journal of Medicine. 2014;371(20):1889–99.
17. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics. 2013;45(6):656–63.
18. Corander J, Fraser C, Gutmann MU, Arnold B, Hanage WP, Bentley SD, et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. Nature Ecology and Evolution. 2017;1(12):1950–60.

19. Ladhani SN, Collins S, Djennad A, Sheppard CL, Borrow R, Fry NK, et al. Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in England and Wales, 2000–17: a prospective national observational cohort study. The Lancet Infectious Diseases. 2018;18(4):441–51.
20. Koelman DiederikLH, Brouwer MatthijsC, van de Beek Diederik. Resurgence of pneumococcal meningitis in Europe and Northern America. Clinical Microbiology and Infection. 2019;26(2):199–204.
21. Malley R, Lipsitch M, Stack A, Saladino R, Fleisher G, Pelton S, et al. Intranasal immunization with killed unencapsulated whole cells prevents colonization and invasive disease by capsulated pneumococci. Infection and Immunity. 2001;69(8):4870–3.
22. Campo JJ, Le TQ, Pablo J v., Hung C, Teng AA, Tettelin H, et al. Panproteome-wide analysis of antibody responses to whole cell pneumococcal vaccination. eLife. 2018;12(7):e37015.
23. Moffitt K, Malley R. Rationale and prospects for novel pneumococcal vaccines. Human Vaccines and Immunotherapeutics. 2016;12(2):383–92.
24. Colijn C, Corander J, Croucher NJ. Designing ecologically optimized pneumococcal vaccines using population genomics. Nature Microbiology. 2020;5(3):473–85.
25. Morais V, Texeira E, Suarez N. Next-generation whole-cell pneumococcal vaccine. Vaccines. 2019;7(4):151.
26. Pichichero ME. Pneumococcal whole-cell and protein-based vaccines: changing the paradigm. Expert Review of Vaccines. 2017;16(12):1181–90.
27. Lees JA, Kremer PHC, Manso AS, Croucher NJ, Ferwerda B, Serón MV, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. Microbial Genomics. 2017;3(1):e000103.
28. Lees JA, Brouwer M, van der Ende A, Parkhill J, van de Beek D, Bentley SD. Within-Host sampling of a natural population shows signs of selection on Pde1 during bacterial meningitis. Infection and Immunity. 2017;85(3):e01061-16.
29. Zafar MA, Hammond AJ, Hamaguchi S, Wu W, Kono M, Zhao L, et al. Identification of pneumococcal factors affecting pneumococcal shedding shows that the dlt locus promotes inflammation and transmission. mBio. 2019;10(3):e01032-19.
30. Spijkerman J, Prevaes SMPJ, van Gils EJM, Veenhoven RH, Bruin JP, Bogaert D, et al. Long-term effects of pneumococcal conjugate vaccine on nasopharyngeal carriage of S. pneumoniae, S. aureus, H. influenzae and M. catarrhalis. PLoS ONE. 2012;7(6):e39730.
31. Bosch AATM, van Houten MA, Bruin JP, Wijmenga-Monsuur AJ, Trzciński K, Bogaert D, et al. Nasopharyngeal carriage of Streptococcus pneumoniae and other bacteria in the 7th year after implementation of the pneumococcal conjugate vaccine in the Netherlands. Vaccine. 2016;34(4):531–9.
32. van Beek J, Veenhoven RH, Bruin JP, van Boxtel RAJ, de Lange MMA, Meijer A, et al. Influenza-like illness incidence is not reduced by influenza vaccination in a cohort of older adults, despite effectively reducing laboratory-confirmed influenza virus infections. Journal of Infectious Diseases. 2017;216(4):415–24.
33. O'Brien KL, Nohynek H. Report from a WHO working group: standard method for detecting upper respiratory carriage of Streptococcus pneumoniae. The Pediatric Infectious Disease Journal. 2003;22(2):e1-11.
34. Trzciński K, Bogaert D, Wyllie A, Chu MLJN, van der Ende A, Bruin JP, et al. Superiority of Trans-Oral over Trans-Nasal Sampling in Detecting Streptococcus pneumoniae Colonization in Adults. PLoS ONE. 2013;8(3):e60520.
35. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genetics [Internet]. 2014;10(8):e1004547. Available from: http://dx.plos.org/10.1371/journal.pgen.1004547
36. Turner P, Turner C, Jankhot A, Phakaudom K, Nosten F, Goldblatt D. Field Evaluation of Culture plus Latex Sweep Serotyping for Detection of Multiple Pneumococcal Serotype Colonisation in Infants and Young Children. PLoS ONE. 2013;8(7):e67933.
37. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15(3):R46.
38. Page AJ, de Silva N, Hunt M, Quail MA, Parkhill J, Harris SR, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. Microbial genomics. 2016;2(8):e000083.
39. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.

40. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nature Genetics. 2014;46(3):305–9.
41. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331(6016):430–4.
42. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Research. 2019;29(2):304–16.
43. Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, et al. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. The Lancet Infectious Diseases. 2019;19(7):759–69.
44. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3.
45. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 2015;32(1):268–74.
46. Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. bioRxiv. 2014;
47. Schweiger R, Fisher E, Rahmani E, Shenhav L, Rosset S, Halperin E. Using stochastic approximation techniques to efficiently construct confidence intervals for heritability. Journal of Computational Biology. 2018;25(7):794–808.
48. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. Pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018;34(24):4310–2.
49. Jaillard M, Lima L, Tournoud M, Mahé P, Belkum A van, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between kmers and genetic events. PLoS Genet. 2018;14(11):e1007758.
50. Lees JA, Mai TT, Galardini M, Wheeler NE, Horsfield ST, Parkhill J, et al. Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. mBio. 2020;11(4):e01344-20.
51. Willer CJ, Li Y, Abecasis GR. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.
52. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nature Microbiology. 2016;4(1):16041.
53. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nature Communications. 2016;9(7):12797.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403–10.
55. Chan JM, Gori A, Nobbs AH, Heyderman RS. Streptococcal Serine-Rich Repeat Proteins in Colonization and Disease. Frontiers in Microbiology. 2020;10(11):593356.
56. Middleton DR, Aceil J, Mustafa S, Paschall A v., Avci FY. Glycosyltransferases within the psrP Locus Facilitate Pneumococcal Virulence. Journal of Bacteriology. 2021;203(7):e00389-20.
57. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: Transmission, colonization and invasion. Nature Reviews Microbiology. 2018;16(6):355–367.
58. Turner P, Melchiorre S, Moschioni M, Barocchi MA, Turner C, Watthanaworawit W, et al. Assessment of Streptococcus pneumoniae pilus islet-1 prevalence in carried and transmitted isolates from mother-infant pairs on the Thailand-Burma border. Clinical Microbiology and Infection. 2012;18(10):970–5.
59. Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. Infection and Immunity. 2007;75(1):83–90.
60. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genetics. 2014;10(8):e1004547.
61. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nature Communications. 2019;10(1):2176.