The Journal of Infectious Diseases

MAJOR ARTICLE

Correlates of Nonrandom Patterns of Serotype Switching in Pneumococcus

Shreyas S. Joshi, Mohammad A. Al-Mamun, and Daniel M. Weinberger

Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, Connecticut, USA

Background. Pneumococcus is a diverse pathogen, with >90 serotypes, each of which has a distinct polysaccharide capsule. Pneumococci can switch capsules, evading vaccine pressure. Certain serotype pairs are more likely to occur on the same genetic background as a results of serotype switching, but the drivers of these patterns are not well understood.

Methods. We used the PubMLST and Global Pneumococcal Sequencing Project databases to quantify the number of genetic lineages on which different serotype pairs occur together. We also quantified the genetic diversity of each serotype. Regression model were used to evaluate the relationship between shared polysaccharide components and the frequency of serotype co-occurrence and diversity.

Results. A number of serotype pairs occurred together on the same genetic lineage more commonly than expected. Co-occurrence of between-serogroup pairs was more common when both serotypes had glucose as a component of the capsule (and, potentially, glucuronic acid, any-N-acetylated sugar, or ribitol). Diversity also varied markedly by serotype and was associated with the presence of specific sugars in the capsule.

Conclusions. Certain pairs of serotypes are more likely to co-occur on the same genetic background. These patterns were correlated with shared polysaccharide components. This might reflect adaptation of strains to produce capsules with specific characteristics.

Keywords. Pneumococcus; serotype switching; polysaccharide capsule; vaccination.

Streptococcus pneumoniae (pneumococcus) is commonly found in the upper respiratory tract of healthy children and also causes a large burden of disease in children and adults. Pneumococci are grouped into serotypes based on the structure and antigenic properties of the extracellular polysaccharide capsule [1]. More than 90 serotypes have been identified, with a fraction of these serotypes causing the majority of invasive pneumococcal disease cases worldwide.

The capsule itself is the target of pneumococcal conjugate vaccines. Currently available conjugate vaccines target the capsules of 10 or 13 of these serotypes [2–4]. These vaccines have driven down rates of severe disease and have also disrupted transmission of vaccine-targeted serotypes among healthy children [5, 6]. Because of this disruption to the bacterial ecology, serotype replacement is observed, wherein serotypes not targeted by the vaccine increase in frequency among healthy carriers and, to a lesser degree, as causes of disease [7]. With next-generation vaccines under development that target additional serotypes, it is important to understand the factors that shape the pneumococcal population and the emergence of new strains.

Received 23 October 2019; editorial decision 18 December 2019; accepted 23 December 2019; published online December 25, 2019.

Correspondence: Daniel M. Weinberger, Department of Epidemiology of Microbial Diseases, Yale School of Public Health, PO Box 208034, New Haven, CT 06520 (daniel.weinberger@yale.edu).

The Journal of Infectious Diseases® 2020;221:1669–76

DOI: 10.1093/infdis/jiz687

The nonvaccine serotypes that emerge following vaccine introduction result from the tremendous diversity of pneumococcus. Beyond variation in the capsular polysaccharide, only about 70% of the genome is conserved across all pneumococcal strains, and there is a large complement of accessory genes [8, 9]. Pneumococci can be classified into genetic lineages using various classification schemes, and genetic lineages can acquire different capsules [10–12]. In a serotype-switching event, a strain producing a particular capsule acquires the genetic cassette required to synthesize a different capsular polysaccharide from another strain [13–16]. Recombination of these loci leads to the production of a different polysaccharide capsule or even the emergence of a novel serotype [17]. These capsule-switching events generate a pool of diversity that can contribute to serotype replacement. In particular, when a genetic lineage that predominantly expresses a vaccine-targeted capsule (a vaccine serotype, VT) has undergone capsule switching, then a variant of that lineage that expresses a capsule not targeted by the vaccine can increase in frequency [12].

Some pairs of serotypes tend to be found on the same genetic lineage more commonly than other pairs. For instance, structurally similar serotypes of the same serogroup (eg, serotypes 19A and 19F) tend to be detected together more commonly on the same genetic lineage than a pair of totally distinct serotypes in different serogroups [16, 18–21]. These patterns could result from direct capsule switching between the serotypes (eg, serotype 19A switching to 19F). Alternatively, a lineage producing a third serotype could independently switch to serotype 19A and

19F. In either case, the 19A and 19F strains would be genetically related and would have similar metabolic machinery. Therefore, co-occurrence patterns might reflect the metabolic suitability of the genetic background to produce particular capsular polysaccharides [21].

In the current study, we used data from several large global databases of pneumococcal clinical isolates with information on genetic lineage and serotype to evaluate the degree of diversity of individual serotypes and to characterize patterns of co-occurrence of serotype pairs on the same genetic lineage. We accomplished this by analyzing the frequency of genotype/serotype co-occurrence using complementary analytic tools. We then considered the potential role of shared biochemical characteristics of the capsules in influencing these patterns.

METHODS

Data Sources

PubMLST (https://pubmlst.org/) is a large global database of pneumococcal clinical isolates with information about multilocus sequence types (MLSTs). Allele and MLST assignments for *S. pneumoniae* isolates in the PubMLST database were used in this analysis [20, 22]. Isolates for which the serotypes were nontypeable, or with errors in their names were excluded from the analysis. The final data set consisted of 35 898 isolates, representing 96 serotypes and 11 718 MLST lineages from the data available on 23 March 2019. Based on the allelic similarity between isolates, 581 clonal clusters (CCs) were generated using the eBURST [23] tool from the PubMLST database.

As a comparison, we analyzed the global pneumococcal sequence cluster (GPSC) data for 13 454 isolates (539 groupings) from the Global Pneumococcal Sequencing Project presented by Gladstone et al [24, 25]. Throughout the manuscript, "genetic lineage" refers to the MLST, CC, or GPSC group, as indicated. We used polysaccharide composition data from the previously published literature [26], and the carriage data were as described by Tothpal et al [27]. In sensitivity analyses, we used the GPSC phylogenetic data to create subsets of the GPSC groupings. These data were obtained from https://microreact.org/project/gpsGPSC*, where * indicates the GPSC number. These phylogenetic data, representing 10 455 isolates and 69 GPSCs, were compiled and saved in the GitHub repository for this project (https://github.com/weinbergerlab/sero_switch_paper). The GPSCs were divided into subgroups by cutting the phylogenetic tree at an arbitrary genetic distance of 300, 200, or 100 units, using the cutree function in R software version 3.6.1. This results in 1483, 2634, or 6490 subgroupings, respectively.

For all analyses, the key assumption is that the genetic lineages reflect a coherent grouping of isolates that share important characteristics (eg, metabolic machinery). GPSCs represent a broad grouping, and CCs and MLSTs are narrower (more homogenous) groupings. The GPSCs can be further subdivided based on phylogenetic distance.

Detecting Nonrandom Serotype-Co-occurrence Patterns

We sought to detect the frequency of co-occurrence of serotype pairs on the same lineage. We used 2 complementary methods to detect such capsule switching events: Monte Carlo (MC) [20, 28, 29] and market basket (MB) [30, 31] analysis (also called association rule mining). We used MC simulation to assess whether the number of genetic lineages associated with a serotype pair was more than what was expected by chance. We first counted the number of genetic lineages on which a serotype pair was detected—that is, the observed number of shared genetic lineages. These pairs were then reshuffled to produce 1000 random data sets, and for each pair, the number of shared genetic lineages were calculated. The 97.5th percentile was chosen as a significance threshold, and serotype pairs with shared genetic lineages higher than this threshold were considered significant candidates for potential serotype switching.

As a complementary analysis, we used MB analysis, which is used in commerce settings to identify items likely to occur together. This provides a computationally efficient way to detect associations between items and groups of items (serotypes in this instance). In MB analysis, several statistics are calculated that measure the strength of association. *Lift* is a measure of how much more common the pairing is than would be expected if the items were independent (values >1 indicate that the grouping is more common than expected), *confidence* is a measure of the proportion of instances in which, when serotype *A* is detected, serotype *B* is also detected, and *support* represents the proportion of all pairs that contain serotype pair *A-B* (among *N* total pairs). These 3 measures are typically used in tandem when performing MB analyses. Support for the pair *AB* is calculated as frequency (*AB*)/*N*, confidence as frequency (*AB*)/frequency (*A*) and lift as support (*AB*)/[frequency (*A*)/*N* * frequency (*B*)/*N*)].

Aggregating the data by MLST and CC has tradeoffs for detecting the frequency of co-occurrence of serotype pairs. To assign overall importance values to serotype associations from the MB and MC analysis results, we built an index by assigning values to the output of the different statistical analyses of our data and then summing them. Because there was not an a priori reason to prefer grouping by MLST or by CC, we gave equal weight to the 2 analyses. Serotype pairs with count of ≥3 were assigned a value of 1.

For MC output, serotype pairs were assigned a value of 1 if observed values were greater than the threshold of the 97.5th percentile. For MB output, a value of 0.25 was assigned for pairs with confidence measure values in the top quartile, 0.25 if the χ^2 test result is significant, 0.25 for lift >1, and 0.25 if the Fisher exact test result for lift is significant. Essentially, we have assigned values of 1 to the count, 1 to MC, and 1 to MB (maximum value, 3). The values for both MLST and CC are then combined, thus giving a range of index scores from 0 to 6, with larger scores indicating stronger evidence of a link. Based on

the index output, serotype associations were visualized using a network plot. The network for serotype pairs with index scores >4 provided the best balance of complexity ability to visually interpret the plot.

Serotype Diversity

Serotype diversity was calculated based on how many isolates of that serotype were identified in different genetic lineages. We made a Serotype*Genetic Lineage matrix containing the number of isolates for each unique combination of serotype and genetic lineage. The Simpson diversity index (SDI) is a tool used in population genetics to measure diversity within populations [32, 33]. We used it to measure the diversity of the genetic lineage within individual serotypes. In this case, the SDI represents the probability that in a sample where all isolates belong to the same serotype, 2 randomly selected isolates have different genetic lineages. SDIs were calculated for each serotype using the formula $\Sigma n(n-1)/N(N-1)$,where N is the total number of isolates belonging to a serotype and n the number of isolates for a single genetic lineage within a serotype.

Correlation of Capsule Switch Patterns and Diversity With Polysaccharide Characteristics

We hypothesized that the inclusion of specific sugars in the capsule could influence the frequency of co-occurrence of pairs of serotypes on the same genetic lineage. The goal was to test whether serotypes that had sugar *X* in their capsule were more likely to be detected on the same genetic lineage as another serotype that also had sugar *X* in its capsule. For every potential pair of serotypes, we counted the number of genetic lineages that they were both detected on. We then fit a series of quasi-Poisson regressions in which the outcome variable was the number of genetic lineages on which both serotypes were detected.

The main covariate of interest was whether sugar *X* was present in both members of the serotype pair or just 1 member of the pair. A positive coefficient would indicate that the pairs in which both members have sugar *X* in the capsule are more common than serotype pairs in which only 1 member of the pair has sugar *X*. Serotypes found on more genetic lineages would be more likely to co-occur with each other by chance. Therefore, we adjusted for this in the regression by using the product of the proportion of genetic lineages on which serotype *A* was found and the proportion of genetic lineages on which serotype *B* was found. This product was normalized using a Box-Cox transformation ($\lambda = 0.01$). Each sugar was tested separately. The analyses were restricted to serotype pairs that were not in the same serogroup and in which ≥1 member of the pair had sugar *X* in the capsule. The analyses were repeated separately when grouping strains by MLST, CC, or GPSC. We also performed a similar set of analyses in which the outcome was binary (ie, the serotype pair does or does not co-occur on any MLST or CC). This was analyzed using a logistic regression model. The results from all 4 sets of analyses (CC and MLST; quasi-Poisson and logistic) are presented together for comparison.

As a complementary analysis, we evaluated the correlation between the presence of specific sugars in the capsule and the serotype-specific SDI. To evaluate the correlation between polysaccharide components and SDI, we performed linear regression analysis, in which the outcome variable was the serotype-specific SDI (logit transformed). We expected that SDI would be associated with the prevalence of the serotype (because more commons serotypes have more opportunity for recombination), so we controlled for carriage prevalence in the prevaccine period, as measured in the United Kingdom by Sleeman et al [27, 34]. We then tested whether the presence or absence of individual sugars in different serotypes was associated with SDI. Thirteen of 25 sugars were present in the capsules of ≥3 serotypes and were included in the analysis. Each sugar was tested individually in a series of regression models; each model included covariates for carriage prevalence and the presence or absence of 1 sugar at a time. Akaike information criterion values and regression coefficients with 95% confidence intervals have been reported for individual sugars.

The analyses were performed using RStudio IDE software [35] for programming in R [36]. Custom scripts were written for SDI, MC, and scraping the eBURST output. The arules package was used for performing MB [37]. The networks were generated using the igraph [38] and ggplot2 [39] packages. The igraph output was saved in a graphml file, which was then visualized using Cytoscape software [40]. Phylogenies based on polysaccharide composition and figures of the tree were made using the ape and phytools packages, respectively. The code and data used for the analysis are available at https://github.com/weinbergerlab/sero_switch_paper.

RESULTS

Nonrandom Patterns of Within and Between-Serogroup Co-occurrence

We quantified the degree of serotype co-occurrence, both within and between serogroups, for every possible pair of serotypes. In agreement with previous studies [20], we found a higher-than-expected likelihood of detecting pairs of serotypes from the same serogroup on the same genetic lineage (Figure 1). A number of outside-serogroup pairings were also detected more frequently than expected by chance.

We hypothesized that serotype pairs with capsular polysaccharides that are more similar biochemically might be more likely to co-occur on the same genetic lineage. Serotype pairs in which both members had glucose in their capsules were more likely to be detected on the same genetic background (comparing between-serogroup pairs only) (Figure 2 and Supplementary Table 1). This result was consistent whether aggregating the isolates by CC, MLST, or GPSC and was confirmed by examining the top co-occurring serotype pairs from the MB/MC analyses (Table 1). Glucose was found in the structure of both serotypes

in all but 1 of the identified between-serogroup pairings for which structural information was available (overall, glucose is present in the capsules of 67 of 84 serotypes with known structure). There was also some evidence that serotypes that had glucuronic acid (GlcA), any N-acetylated sugar, or ribitol in the capsule were more likely to co-occur on the same genetic lineage as another serotype with the same component (Figure 2 and Supplementary Table 1).

However, the results varied depending on whether the isolates were grouped by MLST, CC, or GPSC. Subdividing the GPSC based on phylogeny yielded generally similar results, with the exception of galactose, which showed a strong negative association when the GPSCs were subdivided (Supplementary Figure 1). The association between the presence of any N-acetylated sugar or ribitol was stronger when the GPSCs were subdivided based on phylogeny (Supplementary Figure 1). Notable similarities were not identified between the between-serogroup pairs identified in the network plot and overall similarity of polysaccharide structures (Supplementary Figure 2).

Variation in Diversity by Serotype

Another measure of the amount of capsule switching occurring for each serotype is based on the diversity of CCs associated with a serotype—serotypes that switch more readily will be associated with more genetic lineages. There was considerable variability in diversity between serotypes (Figure 3). We considered whether the presence of certain sugars in the capsule was associated with higher or lower diversity. The hypothesis is that the inclusion of certain sugars in the capsule could influence the probability that the capsule could efficiently switch to a different genetic lineage, owing to epistatic interactions. In regression analysis, after controlling for prevaccine carriage frequency, diversity was lower for serotypes with GlcA in their capsules and higher for serotypes with galactose in the capsule (Figure 4 and Supplementary Table 2). The association between GlcA and diversity was apparent when analyzing the data by CC or by GPSC, but not when aggregating by MLST. The association with galactose was similar regardless of grouping, but for grouping by GPSC there was more uncertainty in the estimate. When the GPSCs were subdivided based on phylogeny, the association with galactose and GlcA weakened, whereas a negative association between N-acetylgalactosamine, an N-actetylated sugar, and ribitol strengthened.

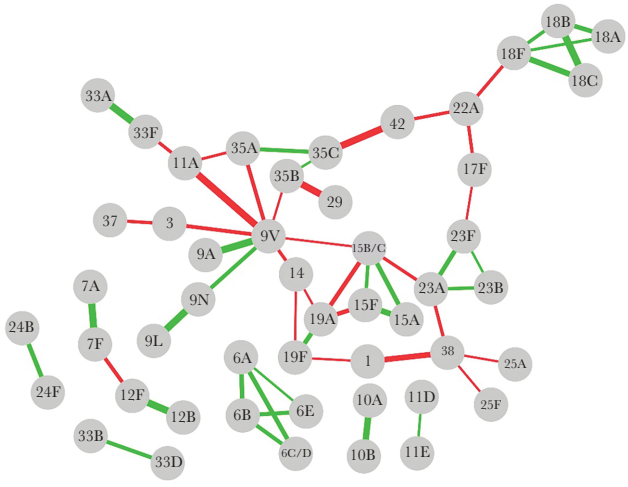

Figure 1. Consensus network of serotype associations. Pairs with index scores ≥4 are shown. The thickness of the connecting lines corresponds to the index score. Green lines indicate within-serotype switches; red lines, between-serotype switches.

DISCUSSION

In the current study, we demonstrated nonrandom patterns of serotypes co-occurring on the same genetic lineage, both within and between serogroups. Switching between serogroups was more common when both members of the pair had glucose (or possibly GlcA, N-acetylated sugars, or ribitol) in the capsule structure. Moreover, there was substantial variability in genetic diversity between serotypes. Variation in diversity was also correlated with the presence of specific components in the polysaccharide capsule. Together, these results support the notion that some capsules differ in their ability to switch to a different genetic background, and some genetic backgrounds could be well adapted to produce capsules with particular characteristics.

We hypothesized that the observed serotype-genotype patterns could be driven by the sugar composition of the capsule itself owing to epistatic interactions between capsular genes and components of central metabolism [41]. In this scenario, a capsule switch event that that results in an antigenic shift in the capsule structure but that would have similar metabolic properties would be most successful. Metabolic tradeoffs could explain why serotypes with certain components (eg, GlcA, ribitol, or any N-acetylated sugar) in the capsule were less diverse on average—if other noncapsular factors are needed to efficiently produce a capsule containing that component, there would be a lower probability of switching. For instance, enzyme efficiency on certain metabolic pathways could influence the metabolic flux that results from shunting key sugars away from central metabolism [42]. Experimental studies would be needed to determine the mechanism underlying these patterns.

Aside from metabolic explanations, serotype-genotype co-occurrence patterns could also be influenced by the structure of the capsular biosynthesis locus. In capsule switching, the genetic cassette from a new serotype is acquired by the bacterium through transformation and integrated into the chromosome through homologous recombination. During this process, the genetic cassette producing the original capsule is lost. Having more homology in the cassette or in the flanking regions could influence the probability of having a successful homologous recombination event [21]. Finally, the mechanism

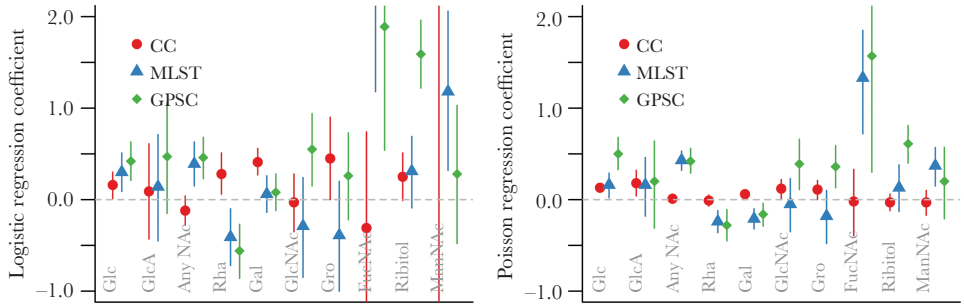

Figure 2. Correlation between the presence of shared sugars between any 2 serotypes and the probability or the number of times that the serotype pair was observed to occur on the same genetic background. *A,* Results from the logistic regression (whether or not the pair was observed). *B,* Results from a Poisson regression (how many times the pair was observed). Values >1 indicate that serotype pairs where both members of the pair had the indicated polysaccharide were more likely to be observed on the same clonal cluster (CC), multilocus sequence type (MLST), or global pneumococcal sequence cluster (GPSC) lineage. Abbreviations: FucNAc, N-acetyl fucosamine; Gal, galactose; Glc, glucose; GlcA, glucuronic acid; GlcNAc, N-acetyl glucosamine; Gro, Glycerol; ManNAc, N-acetyl mannosamine; NAc, any N-acetylated sugar; Rha, Rhamnose.

of initiation of polysaccharide chain formation could influence capsule switching. Most serotypes initiate their polysaccharide chain with glucose [43]. Characteristics of the cell surface on certain strains could be better adapted to link to glucose than to other possible initiating sugars.

In our analyses, we evaluated the frequency of co-occurrence of serotype pairs and the diversity of individual serotypes using genetic groupings of various breadths (eg, GPSC and MLST). The key is to define groupings that are homogenous in terms of their noncapsular metabolic pathways. A more refined way of making these groupings would involve using the whole-genome data to define subgroups of the GPSCs on the basis of both phylogenetic and metabolic similarity [44].

Table 1. Serotype Pairs Found on the Same Genetic Background More Commonly Than Expected, and Their Shared Polysaccharide Components

Serotype Pair[a]	Index Score	Shared Polysaccharide Components
11A-9V	6	Ac, Gal, Glc
29-35B	6	Gal, GalNAc, ribitol
35C-42	6	Gal, Glc, mannitol
1-38	5.5	Structure unknown for 38
15B/C-19A	5	Glc
15F-19A	5	Structure unknown for 15F
12F-7F	4.75	Gal, GalNAc, Glc
3-9V	4.75	Glc, GlcA
35A-9V	4.75	Ac, Gal, Glc
11A-33F	4.5	Ac, Gal, Glc
14-9V	4.5	Gal, Glc
15B/C-23A	4.5	Structure unknown for 23A
17F-22A	4.5	Structure unknown for 22A
18F-22A	4.5	Structure unknown for 22A
22A-42	4.5	Structure unknown for 22A
23A-38	4.5	Structure unknown for 38
3-37	4.5	Glc
11A-35A	4.25	Ac, Gal, Glc
1-19F	4	NA
14-19A	4	Glc
14-19F	4	Glc
15B/C-9V	4	Ac, Gal, Glc
17F-23F	4	Gal, Glc
25A-38	4	Both structures unknown
25F-38	4	Structure unknown for 38
35B-9V	4	Ac, Gal, Glc

Abbreviations: Ac, acetate; Gal, galactose; GalNAc, N-acetylgalactosamine; Glc. glucose; GlcA, glucuronic acid; Man-ol, mannitol; Rib-ol, ribitol.

[a]Between-serogroup pairs with index scores ≥4. Polysaccharide components shared between the serotypes in a pair. Polysaccharide composition data were not available for serotypes 38, 23A, 25A, 22A, and 15F.

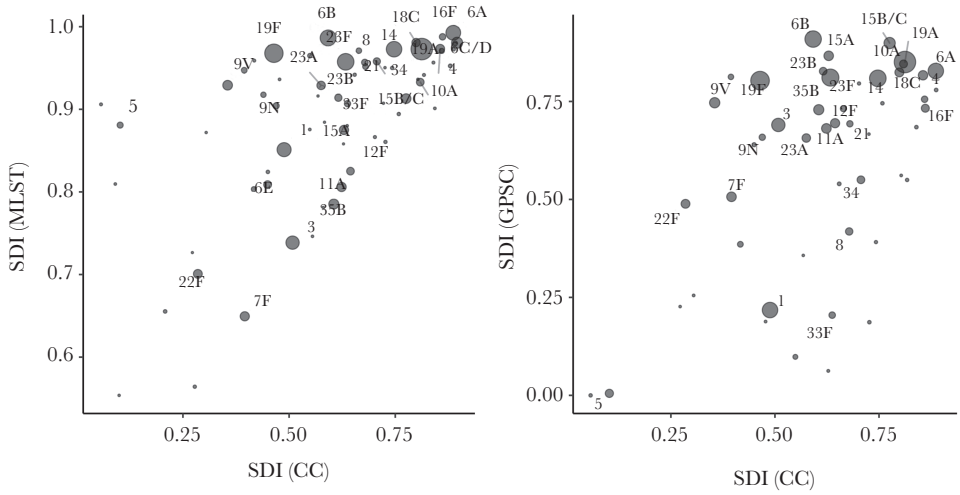

Figure 3. Simpson diversity index (SDI) for each serotype, based on the diversity of associated clonal clusters (CCs), multilocus sequence types (MLSTs), or global pneumococcal sequence clusters (GPSCs). Bubble size is proportional to the number of isolates of the serotype in the respective database.

Our analyses have certain limitations. We cannot determine the direction of the serotype switch from the MLST data, which would be possible with more detailed whole genome sequencing (WGS) data [45]. Tens of thousands of pneumococcal genomes have now been sequenced [14–16, 24, 46], providing an opportunity to further interrogate some of the serotype-genotype co-occurrence patterns we have identified here. The MLST database has other limitations as well. For instance, we were not able to verify the validity of the serotype assignments, and for certain pairs of serotypes (29-35B), the association could be due to misclassification of the serotype by the original investigators. These 2 serotypes are structurally similar and can sometimes be difficult to distinguish with routine methods. However, the data from the Global Pneumococcal Sequencing Project, which are based entirely on DNA sequence rather than laboratory-based serotyping, affirmed the main results of our analyses.

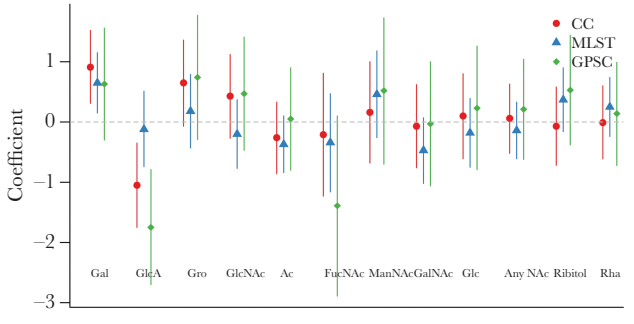

Figure 4. Association between the presence of specific sugars in the capsule and the diversity of that serotype, for isolates grouped by clonal cluster (CC), multilocus sequence type (MLST), or global pneumococcal sequence cluster (GPSC). Abbreviations: FucNAc, N-acetyl fucosamine; Gal, galactose; GalNAc, N-acetyle galactosamine; Glc, glucose; GlcA, glucuronic acid; GlcNAc, N-acetyle glucosamine; Gro, Glycerol; ManNAc, N-acetyl mannosamine; NAc, any N-acetylated sugar; Rha, Rhamnose.

Nontypeable (unencapsulated) strains are thought to play an important role as a source for genetic exchange in pneumococcus [46]. However, because the current analyses were focused on the role of capsule structure, they did not include nontypeable strains.

An assumption from these analyses is that serotype co-occurrence patterns and diversity of genetic lineages within serotypes arise from recombination events where the capsule biosynthetic cassette is exchanged. This is a reasonable assumption, given that much of the diversity in pneumococcus results from recombination rather than mutation. However, mutation could lead to divergence of genetic lineages at the MLST or CC level. The MLST/CC groupings themselves have varying levels of diversity, also complicating the interpretation. Phylogenetic analyses could help disentangle this.

In conclusion, we use information on the co-occurrence of serotypes on the same genetic background to evaluate capsule switching patterns in pneumococcus and to explore possible barriers to capsule switching. The associations of serotype diversity and specific capsular components suggests that there could be important interactions between the capsule and the genetic background on which the capsule is expressed. Such interactions could help shape how pneumococci respond to vaccine-related selective pressure. This information could help predict which serotypes are most likely to emerge in the future.

Supplementary Data

Supplementary materials are available at The *Journal of Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Author contributions. D. M. W. conceived of the study. S. J., D. M. W. conducted analyses; S. J., M. M., D. M. W. designed

analyses. S. J. and D. M. W. drafted the paper. All authors contributed to editing of the paper. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Disclaimer. The funding agency was not involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. had final responsibility for the decision to submit for publication.

Financial support. This work was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health (grant R01AI123208).

Potential conflicts of interest. D. M. W. is principal investigator on a research grant from Pfizer to Yale University and has received personal consulting fees from Pfizer, Merck, GlaxoSmithKline, and Affinivax. All other authors report no potential conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Satzke C, Turner P, Virolainen-Julkunen A, et al; WHO Pneumococcal Carriage Working Group. Standard method for detecting upper respiratory carriage of *Streptococcus pneumoniae*: updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. Vaccine **2013**; 32:165–79.
2. Advisory Committee on Immunization Practices. Preventing pneumococcal disease among infants and young children. Recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep **2000**; 49:1–35.
3. O'Brien KL, Wolfson LJ, Watt JP, et al. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Lancet **2009**; 374:893–902.
4. Kobayashi M, Bennett NM, Gierke R, et al. Intervals between PCV13 and PPSV23 vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Morb Mortal Wkly Rep **2015**; 64:944–7.
5. Hausdorff WP, Bryant J, Paradiso PR, Siber GR. Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I. Clin Infect Dis **2000**; 30:100–21.
6. Harboe ZB, Benfield TL, Valentiner-Branth P, et al. Temporal trends in invasive pneumococcal disease and pneumococcal serotypes over 7 decades. Clin Infect Dis **2010**; 50:329–37.
7. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. Lancet **2011**; 378:1962–73.
8. Donati C, Hiller NL, Tettelin H, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol **2010**; 11:R107.
9. Obert C, Sublett J, Kaushal D, et al. Identification of a candidate *Streptococcus pneumoniae* core genome and regions of diversity correlated with invasive pneumococcal disease. Infect Immun **2006**; 74:4766–77.
10. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology **1998**; 144:3049–60.
11. Costa NS, Pinto TCA, Merquior VLC, et al. MLVA Typing of *Streptococcus pneumoniae* isolates with emphasis on serotypes 14, 9N and 9V: comparison of previously described panels and proposal of a novel 7 VNTR loci-based simplified scheme. PLoS One **2016**; 11:e0158651.
12. Lo SW, Gladstone RA, van Tonder AJ, et al; Global Pneumococcal Sequencing Consortium. 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. Lancet Infect Dis **2019**; 19:759–69.
13. Beall BW, Gertz RE, Hulkower RL, Whitney CG, Moore MR, Brueggemann AB. Shifting genetic structure of invasive serotype 19A pneumococci in the United States. J Infect Dis **2011**; 203:1360–68.
14. Brueggemann AB, Pai R, Crook DW, Beall B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. PLoS Pathog **2007**; 3:e168.
15. Wyres KL, Lambertsen LM, Croucher NJ, et al. Pneumococcal capsular switching: a historical perspective. J Infect Dis **2013**; 207:439–49.
16. Chaguza C, Cornick JE, Andam CP, et al. Population genetic structure, antibiotic resistance, capsule switching and evolution of invasive pneumococci before conjugate vaccination in Malawi. Vaccine **2017**; 35:4594–602.
17. Pillai DR, Shahinas D, Buzina A, et al. Genome-wide dissection of globally emergent multi-drug resistant serotype 19A *Streptococcus pneumoniae*. BMC Genomics **2009**; 10:642.
18. Hanage WP, Bishop CJ, Lee GM, et al. Clonal replacement among 19A *Streptococcus pneumoniae* in Massachusetts, prior to 13 valent conjugate vaccination. Vaccine **2011**; 29:8877–81.
19. Croucher NJ, Harris SR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. Science **2011**; 331:430–4.
20. Hanage WP, Bishop CJ, Huang SS, et al. Carried pneumococci in Massachusetts children: the contribution of clonal expansion and serotype switching. Pediatr Infect Dis J **2011**; 30:302–8.

21. Croucher NJ, Kagedan L, Thompson CM, et al. Selective and genetic constraints on pneumococcal serotype switching. PLoS Genet **2015**; 11:e1005095.
22. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res **2018**; 3:124.
23. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol **2004**; 186:1518–30.
24. Gladstone RA, Lo SW, Lees JA, et al; Global Pneumococcal Sequencing Consortium. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine **2019**; 43:338–46.
25. Global Pneumococcal Sequencing Project. Global Pneumococcal Sequencing Project. https://www.pneumogen.net/gps/. Accessed 20 January 2020.
26. Geno KA, Gilbert GL, Song JY, et al. Pneumococcal capsules and their types: past, present, and future. Clin Microbiol Rev **2015**; 28:871–99.
27. Tóthpál A, Desobry K, Joshi SS, Wyllie AL, Weinberger DM. Variation of growth characteristics of pneumococcus with environmental conditions. BMC microbiology **2019**; 19:1–8.
28. Manly BFJ. Randomization, bootstrap and Monte Carlo methods in biology. New York: Chapman and Hall/CRC; **2018**:480.
29. Hanage WP, Finkelstein JA, Huang SS, et al. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. Epidemics **2010**; 2:80–4.
30. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. ACM SIGMOD Rec **1997**; 26:255–64.
31. Raeder T, Chawla NV. Market basket analysis with networks. Soc Netw Anal Min **2011**; 1:97–113.
32. Simpson EH. Measurement of diversity. Nature **1949**; 163:688.
33. Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol **1988**; 26:2465–6.
34. Sleeman KL, Griffiths D, Shackley F, et al. Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. J Infect Dis **2006**; 194:682–8.
35. Rstudio Team. RStudio: integrated development for R. Boston, MA: RStudio, **2015**.
36. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, **2018**.
37. Hahsler M, Grun B, Hornik K. arules—A computational environment for mining association rules and frequent item sets. J Stat Softw **2005**; 14:1–25.
38. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems **2006**; 1695:1–9.
39. Wickham H. Ggplot2: elegant graphics for data analysis. **2009**.
40. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res **2003**; 13:2498–504.
41. Skwark MJ, Croucher NJ, Puranen S, et al. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. PLoS Genet **2017**; 13:e1006508.
42. Mollerach M, García E. The *galU* gene of *Streptococcus pneumoniae* that codes for a UDP-glucose pyrophosphorylase is highly polymorphic and suitable for molecular typing and phylogenetic studies. Gene **2000**; 260:77–86.
43. Bentley SD, Aanensen DM, Mavroidi A, et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet **2006**; 2:0262–0269.
44. Watkins ER, Penman BS, Lourenço J, Buckee CO, Maiden MC, Gupta S. Vaccination drives changes in metabolic and virulence profiles of *Streptococcus pneumoniae*. PLoS Pathog **2015**; 11:e1005034.
45. Mostowy RJ, Croucher NJ, De Maio N, et al. Pneumococcal capsule synthesis locus *cps* as evolutionary hotspot with potential to generate novel serotypes by recombination. Mol Biol Evol **2017**; 34:2537–54.
46. Chewapreecha C, Harris SR, Croucher NJ, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet **2014**; 46:305–9.