Clinical Infectious Diseases

MAJOR ARTICLE

# The Contribution of Genetic Variation of *Streptococcus pneumoniae* to the Clinical Manifestation of Invasive Pneumococcal Disease

**Amelieke J. H. Cremers,[1,2,3] Fredrick M. Mobegi,[1,2,4] Christa van der Gaast–de Jongh,[1,2] Michelle van Weert,[1,2] Fred J. van Opzeeland,[1,2] Minna Vehkala,[5] Mirjam J. Knol,[6] Hester J. Bootsma,[6] Niko Välimäki,[5] Nicholas J. Croucher,[7] Jacques F. Meis,[8] Stephen Bentley,[9] Sacha A. F. T. van Hijum,[2,4,10] Jukka Corander,[5,9,11] Aldert L. Zomer,[12] Gerben Ferwerda,[1,2] and Marien I. de Jonge[1,2]**

[1]Section of Pediatric Infectious Diseases, Laboratory of Medical Immunology, Radboud Institute for Molecular Life Sciences, [2]Radboud Center for Infectious Diseases, [3]Department of Medical Microbiology, and [4]Bacterial Genomics Group, Center for Molecular and Biomolecular Informatics, Radboudumc, Nijmegen, The Netherlands; [5]Department of Mathematics and Statistics, University of Helsinki, Finland; [6]Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; [7]Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, United Kingdom; [8]Department of Medical Microbiology and Infectious Diseases, Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands; [9]Wellcome Trust Sanger Institute, Pathogen Genomics Group, Hinxton, Cambridge, United Kingdom; [10]NIZO, Ede, The Netherlands; [11]Department of Biostatistics, University of Oslo, Norway; and [12]Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, The Netherlands

***Background.*** Different clinical manifestations of invasive pneumococcal disease (IPD) have thus far mainly been explained by patient characteristics. Here we studied the contribution of pneumococcal genetic variation to IPD phenotype.

***Methods.*** The index cohort consisted of 349 patients admitted to 2 Dutch hospitals between 2000–2011 with pneumococcal bacteremia. We performed genome-wide association studies to identify pneumococcal lineages, genes, and allelic variants associated with 23 clinical IPD phenotypes. The identified associations were validated in a nationwide (n = 482) and a post–pneumococcal vaccination cohort (n = 121). The contribution of confirmed pneumococcal genotypes to the clinical IPD phenotype, relative to known clinical predictors, was tested by regression analysis.

***Results.*** Among IPD patients, the presence of pneumococcal gene *slaA* was a nationwide confirmed independent predictor of meningitis (odds ratio [OR], 10.5; *P* = .001), as was sequence cluster 9 (serotype 7F: OR, 3.68; *P* = .057). A set of 4 pneumococcal genes co-located on a prophage was a confirmed independent predictor of 30-day mortality (OR, 3.4; *P* = .003). We could detect the pneumococcal variants of concern in these patients' blood samples.

***Conclusions.*** In this study, knowledge of pneumococcal genotypic variants improved the clinical risk assessment for detrimental manifestations of IPD. This provides us with novel opportunities to target, anticipate, or avert the pathogenic effects related to particular pneumococcal variants, and indicates that information on pneumococcal genotype is important for the diagnostic and treatment strategy in IPD. Ongoing surveillance is warranted to monitor the clinical value of information on pneumococcal variants in dynamic microbial and susceptible host populations.

**Keywords.** invasive pneumococcal disease; bacterial genomics; genome-wide association study; clinical prediction; molecular diagnostics.

*Streptococcus pneumoniae* is well equipped to colonize and transmit from the human nasopharynx. Invasive pneumococcal disease (IPD) is a threat to both the patient as well as the pneumococcus [1]. It occurs nonetheless, and is a major cause of morbidity and mortality worldwide [2]. The variety in clinical presentations across IPD patients is considerable and not fully explained by host factors alone [3]. It is therefore of interest to investigate whether it matters which pneumococcal variant happens to proliferate in the body.

Invasive disease includes ongoing presence of bacteria in blood and other sterile body sites such as the pleural cavity and cerebrospinal fluid (CSF), corresponding with bacteremia, empyema, and meningitis, respectively. The main reason for these phenomena to occur is an insufficient host defense [4–6]. Although "invasive" pneumococcal traits have been suggested as well [7, 8], the large variety of pneumococci retrieved from IPD and the replacement of serotypes observed after the introduction of pneumococcal conjugate vaccines (PCVs) temper the relative importance of pneumococcal variation as determinant of invasion [9, 10].

Yet, patients who have acquired pneumococci in their bloodstream do not always develop sepsis, and clinical presentations vary from mild respiratory disease to imminent death [11]. Aside from the classical vulnerable elderly patient who slowly recovers from pneumococcal pneumonia upon in-hospital treatment, IPD can manifest at all ages, in a range of body sites, with varying severity and sequelae. It is important to understand

Received 16 August 2017; editorial decision 5 May 2018; accepted 10 May 2018; published online May 17, 2018

Correspondence: A. J. H. Cremers, Department Medical Microbiology, Radboudumc, Internal post 777, PO box 9101, 6500 HB Nijmegen, The Netherlands (amelieke.cremers@radboudumc.nl).

**Clinical Infectious Diseases® 2019;68(1):61–9**

DOI: 10.1093/cid/ciy417

the origins of this diversity. Despite the introduction of uniform clinical guidelines and vaccines, the global pneumococcal disease burden remains high [9, 12, 13], and patients may benefit from more tailored adjunctive measures targeting the effects of specific pneumococcal variants once these have become involved in invasive disease [14].

The diversity in pneumococcal variants, illustrated by the many different capsular serotypes, has long been appreciated in pneumococcal vaccination and surveillance. Previous serotype-based studies indicate that there is a relationship between the type of pneumococcus and particular clinical manifestations of IPD [15]. However, *S. pneumoniae* is a naturally competent organism that fosters genetic recombination via transformation not only at the region that encodes the capsular serotype, but throughout its entire genome [16]. This mechanism has enabled pneumococci to accept and delete genetic variations over time, and has led to high genome-wide diversity across pneumococcal populations [17].

Here we studied whether genome-wide pneumococcal variants were associated with clinical manifestations of human IPD in naturally occurring nonselected patient populations.

## METHODS

### Three Clinical Cohorts

The index cohort consisted of 349 patients diagnosed with a pneumococcal bacteremia admitted to 2 Dutch hospitals between January 2000 and June 2011. For the geographical validation cohort, 482 adults with IPD admitted to 20 other Dutch hospitals (having blood cultures assessed in 9 sentinel laboratories) between June 2004 and December 2006 and June 2008 and May 2012 (periods for which clinical metadata were available) were randomly selected from the National Surveillance Database [18, 19]. The temporal validation cohort was collected from one index hospital, and consisted of 121 pneumococcal bacteremia patients hospitalized between November 2012 and February 2016. In the latter cohort, the distribution of pneumococcal serotypes had markedly changed since the introduction of PCVs in the Dutch national immunization program for infants (7-valent PCV in 2006, 10-valent PCV in 2011). This observational study was approved by the medical ethical committees of the participating hospitals.

Clinical data were collected from medical charts. Differences in comparison to the index cohort were tested with a 2-sided Student *t* test or Mann-Whitney *U* test depending on whether the variable was normally distributed. Differences in nominal variables were tested by 2-sided $\chi^2$ test (Fisher exact if <10 cases in any cell). Handling of pneumococcal blood isolates is described in Supplementary Methods 1.

### Clinical Variables

Clinical variables were collected for 2 purposes. Initially, their univariate association with genetic pneumococcal variants was tested. Finally, to assess the relative importance of a validated genetic pneumococcal variant, clinical determinants were adjusted for in a diagnostic model that predicts a particular manifestation of IPD. For all 3 cohorts, clinical data were retrospectively collected from medical charts using a standard data collection form, and registered with patient identifiers in a secured source file. In a separate working file, labeled clinical data were stored together with the nonidentifying study code assigned to each IPD case included. Cases were included if *S. pneumoniae* was isolated from culture of CSF or blood.

Clinical data collected for all 3 cohorts included several dates (of birth, admission, collection of first blood culture positive for *S. pneumoniae*, discharge, transfer, or in-hospital death), sex, diabetes mellitus, cancer (actual malignant neoplasm), immunocompromising therapy (use of systemic corticosteroids or chemotherapeutic agents), cough (as reported or observed at admission), clinical diagnosis (pneumonia, meningitis, and unknown focus of infection as reported by the attending physician; pleural effusion as reported by the attending radiologist; empyema if *S. pneumoniae* was isolated from pleural fluid culture; pneumonia, empyema, and meningitis were not mutually exclusive). The following clinical data were only collected for patients admitted to the index hospitals: Charlson comorbidity score (calculated for cases ≥18 years old), chronic obstructive pulmonary disease, cardiovascular disease (history of either hypertension, myocardial infarction, myocardial insufficiency, claudicatio intermittens, vasculitis, vascular stents, heart catheterization, atrial fibrillation, or hypercholesterolemia), antibiotics prior to admission (within preceding week either in context of separate medical issue or for current infection), systemic inflammatory response syndrome (SIRS) variables (percentage of immature neutrophils not accounted for), blood C-reactive protein (CRP) level and leukocyte count, and pneumonia severity index (PSI) risk class (calculated for cases ≥18 years old who suffered from pneumonia, formatted into PSI risk class, and stratified to PSI risk class 1–2 vs 3–5 for the genome-wide association study [GWAS] analyses). CRP, leukocytes, and other chemistry results included in clinical algorithms were required to be collected at the day of admission, except for bilirubin and albumin during the corresponding hospital stay. Derived clinical variables included age, date of infection as date of blood culture collection, influenza season as defined annually from the first to the last week with >5 reported influenza cases in the Netherlands as reported by the World Health Organization FluNet (http://apps.who.int/flumart/), sepsis if at least 2 SIRS criteria fulfilled, time to death in days from admission until in-hospital death, 30-day mortality as in hospital death within 30 days from admission, and early death as in hospital death within 48 hours from admission.

If within-hospital follow-up was incomplete due to transfer to a nonparticipating hospital, the case was excluded from analyses concerning clinical course during hospital stay. Missing data were not replaced. Data were considered to be missing if the corresponding section in the medical chart was not completed. PSI risk class was only considered valid and reported if ≥16 included variables were known. All other clinical algorithms were reported only if missing variables did not influence case classification.

### Genomic Analyses in the Index Cohort

Whole genome sequencing and assembly, as well as determination of orthologous genes (OGs), functional annotations, core genome, population phylogeny, and population structure (ie, the identification of genetically diverged subpopulations, which are called sequence clusters [SCs]) were performed for the 349 isolates from the index cohort as previously described [10]. In brief, isolates were sequenced on the Illumina HiSeq 2000 platform. Genome assemblies were created using the Sanger Institute genome assembly pipeline and were deposited at the European Nucleotide Archive under study number ERP001789. Isolate-specific sample accession number, serotype, sequence cluster, and multilocus sequence type are provided in Supplementary Data 1. After annotation using Prokka and identification of clusters of orthologous groups using TribeMCL, an alignment of ribosomal genes was established, supplemented with core OGs that followed the distribution of these ribosomal genes in their maximum likelihood phylogeny. From this alignment, we derived both a phylogenetic tree using the RAxML program as well as the SCs predicted using hierBAPS (Bayesian analysis of population structure).

The relationship between SC and 23 clinical characteristics of IPD patients was explored by stepwise regression analysis with SC variable entry set at 0.05, and removal at 0.05 for logistic and at 0.1 for linear regression.

We performed 2 different GWASs in the index cohort. First, we investigated the relationship between the presence of each individual OG on the accessory pneumococcal genome and the clinical IPD phenotype. For this analysis, OGs present in <98% and >2% of cases were selected. Associations with binary clinical variables were assessed by Fisher exact test with cluster permutation and by Cochran-Mantel-Haenszel analyses implemented in PLINK [20]. The SC was introduced as a nominal covariate to adjust for population structure, and *P* values were false discovery rate–corrected to adjust for multiple testing by the Benjamini-Hochberg procedure. Second, we investigated the relationship between any allelic variant present anywhere on the genome and the clinical IPD phenotype. K-mers (DNA words of 10–99 base pairs) were identified from draft assemblies by distributed string mining, and subsequently filtered for adjacent bases having a different frequency support vector in the study cohort, and for being associated with each phenotype at $P < 1 \times 10^{-5}$ in univariate $\chi^2$ testing. Associations between the selected k-mers and binary clinical IPD phenotypes were assessed by sequence element enrichment (SEER) analysis [21], including correction for population structure by multidimensional scaling using a random subset of k-mers. The origin of k-mers was determined by alignment to the annotated draft genomes of the index cohort with complete coverage and identity using BLAST. To adjust for testing millions of anticipated k-mers, the significance threshold was set at $1 \times 10^{-8}$.

### Validation of Associations

We aimed to validate a selection of the identified OGs that were significantly and lineage-independently associated with a clinical IPD phenotype, in a nationwide cohort. In the temporal validation cohort, the number of identified genes evaluated was constrained by the number of cases in that collection period. The sample size of the validation cohorts was calculated to detect the index differences with a power of at least 0.8 and α of .05 in a 1-sided fashion. Because the similarity in distribution of phenotypes and OGs in the validation cohorts was uncertain, the significance threshold for validation was set at 0.1.

To determine the presence of the OGs of interest in pneumococcal genomes in the validation cohorts, primers were designed and validated based on the index cohort using a real-time fluorescent read-out as detailed in Supplementary Methods 2.

### Confirmed Pneumococcal Genotypes

Co-occurrence of confirmed genotypes with other sequence variants was determined by Pearson correlation. Co-localization of confirmed OGs with bacteriophages was assessed by identification of predicted prophage sequences in the draft genomes of the index cohort using the PHASTER (phage search tool enhanced release) program [22]. Sequence variation within the confirmed OGs and prophages was expressed in size, guanine–cytosine (GC) content, and pairwise distances calculated from amino acid alignments using MEGA (molecular evolutionary genetics analysis) 7 software.

### Clinical Relevance

The relative contribution of the identified pneumococcal genotypes (SCs, confirmed OGs, or k-mers) to the clinical IPD phenotype in relation to well-known clinical predictors was assessed by multivariate logistic regression analysis. Clinical variables entered into the model for meningitis were age (per year unit) and cough; for 30-day mortality: Charlson comorbidity score, meningitis, and pneumonia; for 30-day mortality among pneumonia cases: Charlson comorbidity score and the PSI risk class.

To explore clinical detection of pneumococcal variants during IPD, stored serum samples collected from IPD patients at day 0–3 of hospitalization were retrieved from −40°C. We selected those serum samples on which capsular sequence typing had previously been successful [23, 24]. DNA was isolated from 100 µL of serum using the Qiagen DNeasy Blood and tissue kit. The OG validation polymerase chain reaction (PCR) assays were performed in duplicate using 8 µL of template DNA and 50 amplification cycles.

Unless stated otherwise, significance thresholds were set at .05.

## RESULTS

### Three Clinical Cohorts

Although the geographical validation cohort largely overlapped with the study period of the index cohort, its serotype distribution was somewhat different (Supplementary Figure 1). The temporal validation cohort was included to monitor identified associations in changing populations. In this cohort, the patient characteristics of IPD cases had altered as compared to the index cohort, and serotypes clearly changed in response to pneumococcal vaccination. However, the distribution of IPD syndromes and outcomes had remained stable over time.

### Genomic Analyses in the Index Cohort

Of the 23 explored clinical manifestations of IPD, 87% appeared to be associated with 1 or more pneumococcal SCs (Supplementary Table 1), demonstrating the importance of correction for population structure in subsequent GWAS analyses.

In the first GWAS, we examined the relationship between the presence of individual OGs in the accessory pneumococcal genome and clinical IPD phenotype. Independently from SC, 68 of the 1127 selected pneumococcal OGs were associated with 9 different clinical IPD phenotypes (Supplementary Data 2, and most pronounced associations displayed in Figure 1).

Another method was used to identify genome-wide associations between any allelic variant, or k-mer, present anywhere on the genome and clinical IPD phenotype. The identified k-mers had nucleotide sequences that aligned with members of up to 6 different OGs. None of the $1.5 \times 10^7$ identified unique k-mers met the genome wide significance threshold in their SC-independent association with clinical IPD phenotypes (Supplementary Table 2). Despite this, certain OGs were overrepresented as they contained multiple variable regions related to a particular phenotype (Supplementary Table 3).

### Validation of Associations

Only associations with OGs were taken into validation, because validation of associations with SCs and k-mers would have required fully sequenced clinical validation cohorts. OG_17 was considered to be a proxy for OG_761 because of their consistency in the index cohort. PCR assays were successfully established for 8 of 10 OGs selected for validation (reasons detailed in Supplementary Methods 3). The sample size of the temporal validation cohort only allowed for validation of the OGs associated with 30-day mortality (Supplementary Table 4). All diluted templates from isolates in both validation cohorts (n = 603) were positive for the *gyrA* pneumococcal housekeeping gene with a cycle threshold value of 24 ± 2. Out of the 9 OG-phenotype combinations from the index cohort tested, 4 could be replicated in the geographical validation cohort (Figure 2). The presence of OG_2721 was again positively associated with meningitis ($P$ = .003; 8/27 vs 40/455), while the presence of OG_17, OG_675, and OG_58 (at $P$ values of .086, .055, and .090 respectively; on average 38/216 vs 33/265) were again positively associated with 30-day mortality.

### Confirmed Pneumococcal Genotypes

The confirmed OG_2721 related to meningitis was functionally annotated as *slaA* coding for phospholipase A2, and showed invariable anti-occurrence with OG_416 (a predicted membrane protein) and OG_679 (ABC transporter) in the genomes of the index cohort. The 3 confirmed OGs related to 30-day mortality were annotated as phage proteins (OG_17 specified as *pblB* encoding a prophage tail fiber protein), and showed high co-occurrence with each other in all 3 cohorts (Supplementary Figure 2). In the index cohort, all *pblB* homologues were located either within borders of predicted prophage elements, or located near contig breaks or on short contigs, thus representing circumstances under which prophage elements cannot be identified from draft genomes. While all sequences of *slaA* were identical, the other OGs showed large variation (Supplementary Figure 3). Within *pblB* and OG_58 the number of pairwise amino acid positions exceeded the number of amino acids encoded by their largest sequence member. This suggests genetic mosaicism, which is typical for bacteriophage genes. In the distribution of the confirmed OGs in the pneumococcal populations, *pblB* was taken as a proxy for its joint prophage vector shared with OG_761, OG_675, and OG_58 (Supplementary Figure 4). In addition to presence, the number of open reading frames per OG present in an isolate also strongly correlated between these 4 OGs. The distribution of *slaA* and *pblB* across various serotypes and SCs confirms that these are lineage-independent, which justifies the omission of correction for pneumococcal population structure in the validation studies. While all confirmed OGs were present in both vaccine and nonvaccine serotypes, the relative occurrence of *slaA* in IPD cases remained stable over time, yet *pblB* waned.

### Clinical Relevance

Relative to known clinical predictors of meningitis and 30-day mortality, pneumococcal SCs and OGs were still major independent determinants of these phenotypes (Table 1). To illustrate, for patients outside the high PSI risk class, 30-day mortality was 11% in the *pblB*-positive group compared to <1% in the *pblB*-negative group (odds ratio, 13.3).

*SlaA* associated with meningitis was correctly only detected by PCR in serum from patient PBCN0382 (Table 2). For 30-day mortality, OG_675 was most accurately and consistently identified in serum from patients with low pneumococcal DNA loads.

## DISCUSSION

Through comparative genomics, we identified genetic pneumococcal variants (SCs and OGs) to be contributory determinants

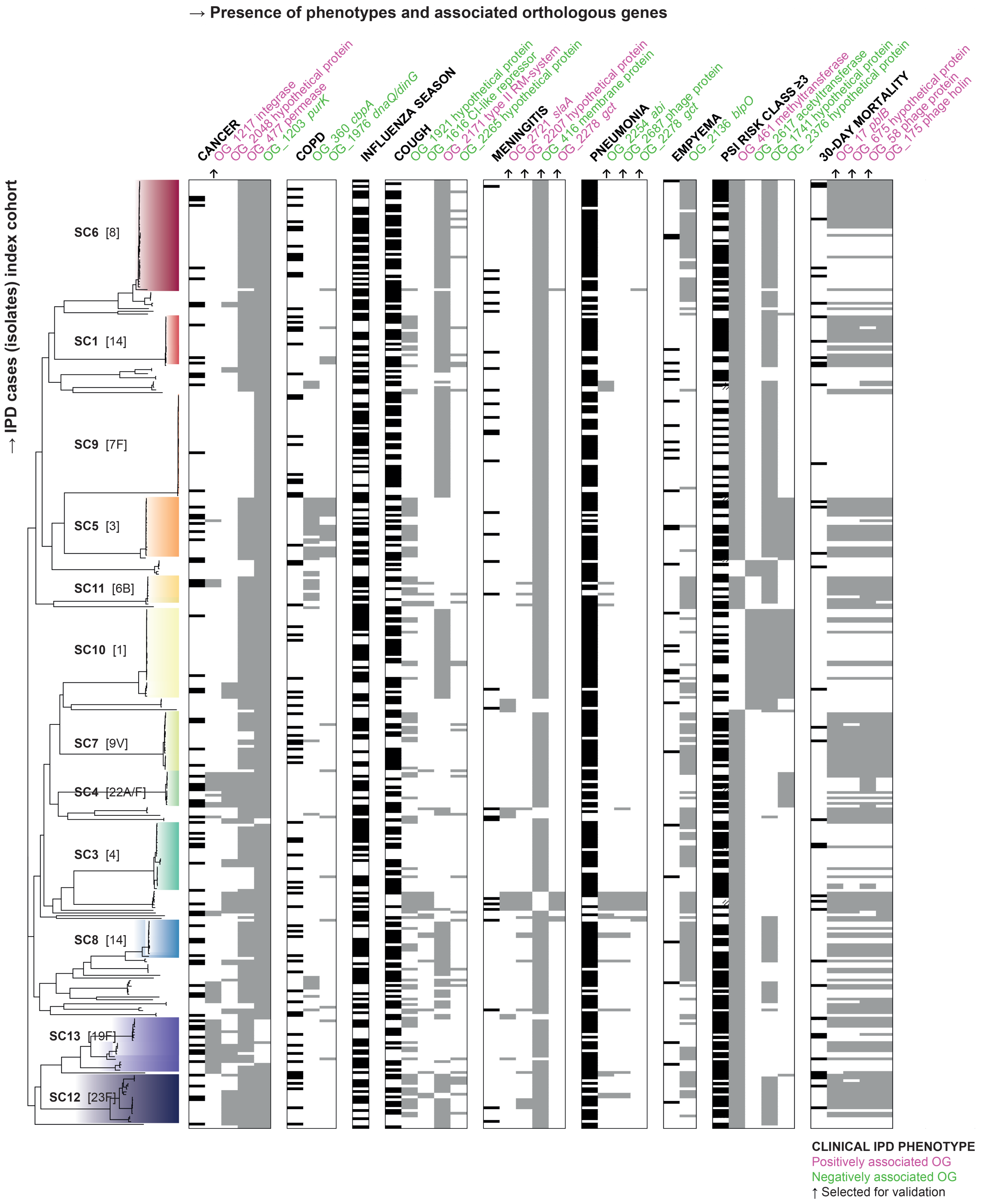

**Figure 1.** Clinical invasive pneumococcal disease (IPD) phenotypes with associated orthologous pneumococcal genes in the index cohort. Rows represent 349 IPD cases and corresponding pneumococcal blood isolates. The tree on the left represents their relative phylogenetic position based on single-nucleotide polymorphisms in the core genome, in which the identified sequence clusters (in bold) with their predominant serotypes (within square brackets) are highlighted. The columns represent the presence (filled) or absence (empty) of clinical IPD phenotypes and their associated pneumococcal orthologous genes (OGs) with annotation at the top. Maximally 4 associations that passed Fisher exact with $P < .01$ and were independent of population structure are displayed. The OGs selected for subsequent validation in 2 additional cohorts are indicated by an arrow. Abbreviations: COPD, chronic obstructive pulmonary disease; PSI, pneumonia severity index; SC, sequence cluster.

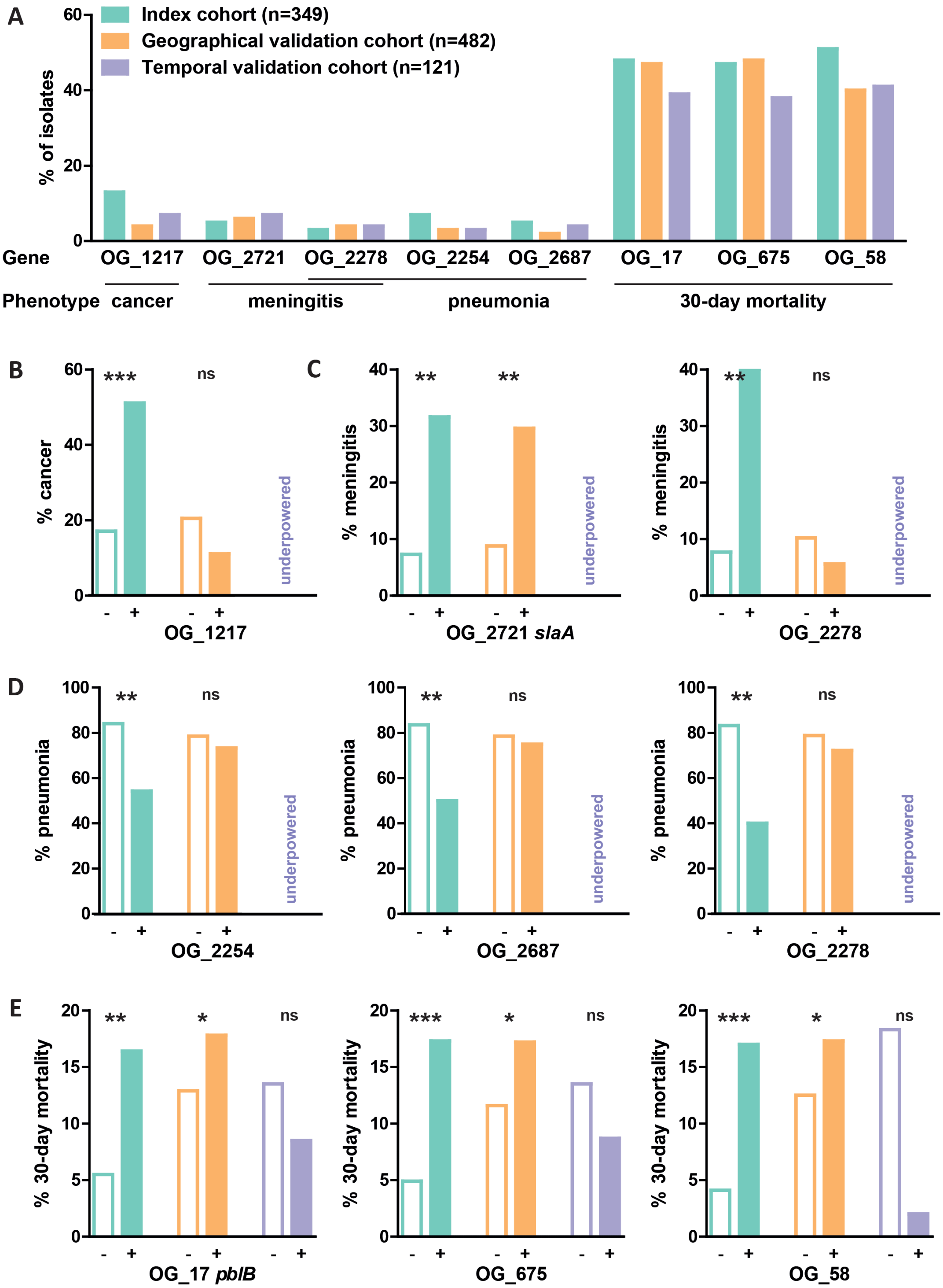

**Figure 2.** Geographical and temporal validation of orthologous gene (OG) associations. *A*, Occurrence of pneumococcal OGs across the genomes of the 3 invasive pneumococcal disease (IPD) cohorts. The associations between absence (–, empty bar) or presence (+, filled bar) of these OGs and the proportion of patients affected by a particular IPD phenotype in the index and 2 validation cohorts are displayed for cancer (*B*), meningitis (*C*), pneumonia (*D*), and 30-day mortality (*E*). **P* < .1; ***P* < .01; ****P* < .001. *P* values indicate differences in the proportion of patients affected by a particular IPD phenotype. Underpowered means the sample size did not meet the validation requirement. Abbreviation: OG, orthologous gene.

**Table 1. Optimized Prediction Models for Meningitis and 30-Day Mortality**

| Phenotype | Determinant | OR | (95% CI) | *P* Value |
|---|---|---|---|---|
| Meningitis | Cough | 0.06 | (.02–.22) | $9 \times 10^{-6}$ |
| | OG_2721 *slaA* | 10.5 | (2.61–42.30) | .001 |
| | Age | 0.97 | (.94–.99) | .006 |
| | SC9 (serotype 7F) | 3.68 | (.96–14.05) | .057 |
| 30-d mortality | Charlson comorbidity score | 1.44 | (1.22–1.70) | $1 \times 10^{-5}$ |
| | OG_17 *pblB* | 3.43 | (1.54–7.65) | .003 |
| | Meningitis | 5.05 | (1.68–15.15) | .004 |
| 30-d mortality among pneumonia cases | Charlson comorbidity score | 1.34 | (1.06–1.68) | .013 |
| | OG_17 *pblB* | 3.28 | (1.18–9.11) | .023 |
| | PSI risk class | 2.22 | (1.07–4.63) | .033 |

Sequence clusters 9 and 10 made no relative contribution to the models for 30-day mortality.

Abbreviations: CI, confidence interval; OG, orthologous gene; OR, odds ratio; PSI, pneumonia severity index; SC, sequence cluster.

of clinical manifestations of IPD, supported by validation in a separate cohort. These pneumococcal sequence variants could be detected in serum samples from IPD patients by PCR.

Prediction of clinical phenotypes as performed in this study comes with two particular challenges. First, for the identification of certain clinical syndromes, one relies on the assessment and examinations performed by the attending physician. Missed diagnosis of, for example, meningitis cannot be ruled out, given that the absence of cough was one of its main predictors. While uncertainty in sensitivity is inherent to studying clinical phenotypes, the specificity of affected cases is robust as only laboratory-confirmed cases of meningitis were classified as such [25]. In fact, instant knowledge of pneumococcal genotype could be used to improve future recognition of particular disease manifestations. Second, although mortality from IPD is easier to establish, its determinants can vary widely across different clinical settings [26]. We observed in our temporal postvaccination validation cohort that the relative contribution of pneumococcal variants to mortality may also be influenced by an altered composition of the pneumococcal population itself. Therefore, validity of our findings in other settings should not be assumed, but tested. At the same time, it is difficult to estimate a sample size threshold at which to reject validity because other settings commonly differ in standards of care, population at risk [27], antibiotic resistance level, and serotype distribution [28]. Therefore, although targeted validation as performed in our relatively similar clinical cohorts seems appropriate, in very dissimilar populations de novo identification of relevant pneumococcal genotypes may be a more adequate approach.

Our nonselective method including genome-wide pneumococcal variants in naturally occurring IPD populations ensured the likelihood that an association being identified directly correlated with its clinical relevance. In a previous Malawian GWAS where no pneumococcal meningitis–related OGs were identified, not only did the human and pneumococcal populations differ from ours [29], but the selection for meningitis also likely altered the relative contribution of certain pneumococcal variants [30]. Vice versa, a determinant identified from an artificial distribution of cases (with unnatural pre-test odds), although it could be informative of disease mechanism, may no longer be valid among patient populations presenting to the hospital.

Pneumococcal genetic population structures are characterized by linkage disequilibrium, which means that particular sets of sequence variants (including the capsular locus) tend to co-occur across the pneumococcal genome. To prevent identification of sequences that actually represent a magnitude of co-occurring variants, we corrected for this population structure in our GWAS analyses. Also, we assessed whether these

**Table 2. Detection of Orthologous Gene Sequences in Serum From Patients With Invasive Pneumococcal Disease**

| | | Study Identifier | | | | |
|---|---|---|---|---|---|---|
| | | PBCN0382 | PBCN0389 | PBCN0420 | PBCN0480 | PBCN0442 |
| Meningitis | | Yes | No | No | Yes | No |
| 30-d mortality | | No | Yes | Yes | Yes | No |
| Serum pneumococcal DNA load, copies/mL | | $8 \times 10^2$ | $7 \times 10^3$ | $3 \times 10^3$ | $3 \times 10^3$ | $1 \times 10^4$ |
| OG_2721 *slaA* | Isolate WGS[a] | 1 | 0 | 0 | 0 | 0 |
| | Serum OG-PCR[b] | 2 | 0 | 0 | 0 | 0 |
| OG_17 *pblB* | Isolate WGS | 2 | 1 | 1 | 1 | 0 |
| | Serum OG-PCR | 2 | 2 | 1 | 1 | 0 |
| OG_675 | Isolate WGS | 2 | 1 | 1 | 1 | 0 |
| | Serum OG-PCR | 2 | 1 | 2 | 2 | 0 |
| OG_58 | Isolate WGS | 2 | 1 | 1 | 1 | 0 |
| | Serum OG-PCR | 2 | 2 | 2 | 2 | 2 |
| Prophage sequence | Isolate WGS | Partial | Partial | Complete | Partial | Absent |

Abbreviations: OG, orthologous gene; PBCN, pneumococcal bacteremia collection Nijmegen; PCR, polymerase chain reaction; WGS, whole genome sequencing.

[a] Number of open reading frames assigned.

[b] Times target detected in duplicate OG-PCR assays.

so-called "sequence clusters" as a whole were related to clinical IPD phenotypes, and we found a remarkable concordance to previous serotype-based studies [31–33].

Our correction for lineage-specific effects may explain why we have not identified variants of single genes that have previously been described to enhance transition from blood to CSF in laboratory models such as *nanA*, *cbpA*, *pCho*, *lytA*, *ply*, and *glpO* [34]. At the same time, this aspect of our approach may have favored the detection of bacterial genotypes located on prophage (ie, genomes of viruses that infect bacteria) to be associated with clinical phenotypes as found by us and by others previously [35]. Unlike many other genes in clonal populations, the distribution of bacteriophages is not as strictly determined by lineage [36]. On the other hand, an important example of a prophage sequence that was associated with the severity of invasive meningococcal disease was discovered by gene array despite no correction for population structure [37]. What we have learned from the k-mer-based GWAS is that the number of lineage-independent allelic variants present in a pneumococcal IPD population is too high for identification of robust associations with particular phenotypes at the current sample size. Although one may have expected *pblB*-fragments to be identified in relation to 30-day mortality, the sequences in this orthologous gene were too dispersed to meet the k-mer selection and association thresholds. On the other hand, despite all sequences of *slaA* being identical, k-mers originating from *slaA* were still included in the SEER analysis (and positively associated with meningitis), because these k-mers were also represented by a second OG that was more dissimilar and as such made the k-mer meet the inclusion criteria. These examples demonstrate the complementarity of the 2 different GWAS methods employed.

While we identified pneumococcal genes to be associated with detrimental IPD phenotypes independently of pneumococcal lineage (serotype) and clinical predictors, this does not prove causality. We have not included potential confounders such as host genotype, a host factor that mediates susceptibility to meningitis [38] and may simultaneously induce a mucosal environment that fosters colonization by specific pneumococcal variants. Yet, evidence for a direct effect of pneumococcal variants in the human bloodstream was demonstrated by measurement of increased activation of human platelets upon interaction with an *S. pneumoniae* wild-type strain compared to an isogenic *pblB* knockout mutant [39]. The predicted function of the protein encoded by the *pblB* gene (ie, functional annotation based on homology) is a phage tail fiber, and its *Streptococcus mitis* orthologue was shown to bind to human platelets as well [40]. In addition, given that multiple pneumococcal genes co-located with *pblB* on a particular prophage were positively associated and validated with 30-day mortality, merely the presence of phages capable of lysing bacterial cells may have contributed to increased pathogenicity in human disease. The identified prophage also seems able to affect the pneumococcus' ability to take up foreign DNA via insertion into the *comYC* competence gene [41]. While we observed such disruption of the *comYC* gene among isolates carrying the prophage, this phenomenon itself was not related to mortality in IPD. Although we have studied pneumococcal blood isolates, it has been shown on the basis of genomic data that no adaptation is needed to cross the blood–brain barrier, so DNA sequences from blood isolates seem representative for pneumococci that reach the CSF and cause meningitis [42]. Human phospholipase A2, encoded by *slaA*, has been shown to reduce the integrity of the blood–brain barrier in vitro, thereby mediating penetration of endothelial cells by group B *Streptococcus* [43]. In group A *Streptococcus* the presence of *slaA* enhanced the bacterium's potential for epithelial adherence, colonization, and invasive disease [44]. Also for *slaA* further studies would be required to elucidate its effects in vivo and possibilities to avert these during disease. From an evolutionary point of view, there is no reason to believe that the pneumococcal genes identified in this study circulate for their pathogenic role in IPD, as invasion is rare, and their occurrence in our 3 IPD cohorts does not deviate from that in several pneumococcal populations isolated from nasopharyngeal carriage cohorts (Supplementary Figure 5). Recent advancements in methods to study epistatic interactions between genes separately located on pneumococcal genomes [45–47] may help to improve our understanding of why particular SCs are overrepresented in certain phenotypes.

Aside from possible disease mechanisms, this study suggests that, in an acute setting, pneumococcal genotypic information could identify patients who deserve additional attention but would have been missed based on the usual clinical cues. Further modeling and feasibility studies will be required to determine the exact positioning of bacterial genotypic testing in clinical management decisions.

This study provides evidence that it does matter which pneumococcal variant proliferates in the bloodstream, as it improves our risk assessment in patients affected by IPD. This suggests that the established value of microbial genomics in public health [48], outbreak management, and combating antimicrobial resistance [49] may now be extended to individual patient care. Increased appreciation of genotypic microbial variants could guide the development of tailored adjunctive measures that are heavily searched for in clinical sepsis care [50].

Because population dynamics are likely to affect their relative importance, the mapping of microbial variants of concern needs to be supported by strong interdisciplinary surveillance networks. Prompt molecular diagnostics at the emergency department could readily improve risk stratification and alertness for complicated infection in individual patient care [51].

## Supplementary Data

Supplementary materials are available at *Clinical 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

***Acknowledgments.*** We thank Dr A. van der Ende at the Reference Laboratory for Bacterial Meningitis and all hospitals involved in the Dutch national surveillance program for their concerted efforts, which made it possible to validate our findings.

***Potential conflicts of interest.*** M. V. reports employment by Statfinn and EPID Research (a contract research organization performing commissioned clinical and pharmacoepidemiological studies in collaboration with pharmaceutical companies). N. J. C. reports consultancy fees from Antigen Discovery Inc. J. F. M. reports consultancy fees and grants from Astellas, Basilea, and Merck, and payments for lectures from Astellas, Gilead, Merck, United Medical, and Teva. F. M. M. reports travel grants from KNAW Academy Ter Meulen Grant 2014. A. L. Z. reports consultancy fees and patents from Janssen. 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. Weiser JN. The pneumococcus: why a commensal misbehaves. J Mol Med (Berl) **2010**; 88:97–102.
2. O'Brien KL, Wolfson LJ, Watt JP, et al; Hib and Pneumococcal Global Burden of Disease Study Team. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Lancet **2009**; 374:893–902.
3. Wunderink RG. CAP death: what goes wrong when everything is right? Lancet Infect Dis **2015**; 15:995–6.
4. Picard C, Puel A, Bustamante J, Ku CL, Casanova JL. Primary immunodeficiencies associated with pneumococcal disease. Curr Opin Allergy Clin Immunol **2003**; 3:451–9.
5. van der Poll T, Opal SM. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. Lancet **2009**; 374:1543–56.
6. Mook-Kanamori BB, Geldhoff M, van der Poll T, van de Beek D. Pathogenesis and pathophysiology of pneumococcal meningitis. Clin Microbiol Rev **2011**; 24:557–91.
7. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J Infect Dis **2003**; 187:1424–32.
8. Browall S, Backhaus E, Naucler P, et al. Clinical manifestations of invasive pneumococcal disease by vaccine and non-vaccine types. Eur Respir J **2014**; 44:1646–57.
9. Miller E, Andrews NJ, Waight PA, Slack MP, George RC. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: an observational cohort study. Lancet Infect Dis **2011**; 11:760–8.
10. Cremers AJ, Mobegi FM, de Jonge MI, et al. The post-vaccine microevolution of invasive *Streptococcus pneumoniae*. Sci Rep **2015**; 5:14952.
11. Cillóniz C, Gabarrús A, Almirall J, et al. Bacteraemia in outpatients with community-acquired pneumonia. Eur Respir J **2016**; 47:654–7.
12. Thigpen MC, Whitney CG, Messonnier NE, et al. Bacterial meningitis in the United States, 1998–2007. N Engl J Med **2011**; 364:2016–25.
13. Billings ME, Deloria-Knoll M, O'Brien KL. Global burden of neonatal invasive pneumococcal disease: a systematic review and meta-analysis. Pediatr Infect Dis J **2016**; 35:172–9.
14. McGill F, Heyderman RS, Panagiotou S, Tunkel AR, Solomon T. Acute bacterial meningitis in adults. Lancet **2016**; 388:3036–47.
15. Hausdorff WP, Feikin DR, Klugman KP. Epidemiological differences among pneumococcal serotypes. Lancet Infect Dis **2005**; 5:83–93.
16. Croucher NJ, Harris SR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. Science **2011**; 331:430–4.
17. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science **2009**; 324:1454–7.
18. Netherlands Reference Laboratory for Bacterial Meningitis (AMC/RIVM). Available at: https://www.amc.nl/web/Research/Overview/Departments/Medical-Microbiology/Medical-Microbiology/Current-research/Reference-Laboratory-for-Bacterial-Meningitis.htm?print=true. Accessed 24 July 2017.
19. Wagenvoort GH, Sanders EA, Vlaminckx BJ, et al. Invasive pneumococcal disease: clinical outcomes and patient characteristics 2-6 years after introduction of 7-valent pneumococcal conjugate vaccine compared to the pre-vaccine period, the Netherlands. Vaccine **2016**; 34:1077–85.
20. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **2007**; 81:559–75.
21. Lees JA, Vehkala M, Välimäki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun **2016**; 7:12797.
22. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res **2016**; 44:W16–21.
23. Elberse K, van Mens S, Cremers AJ, et al. Detection and serotyping of pneumococci in community acquired pneumonia patients without culture using blood and urine samples. BMC Infect Dis **2015**; 15:56.
24. Cremers AJ, Hagen F, Hermans PW, Meis JF, Ferwerda G. Diagnostic value of serum pneumococcal DNA load during invasive pneumococcal infections. Eur J Clin Microbiol Infect Dis **2014**; 33:1119–24.
25. Khatib U, van de Beek D, Lees JA, Brouwer MC. Adults with suspected central nervous system infection: a prospective study of diagnostic accuracy. J Infect **2017**; 74:1–9.
26. Aston SJ, Rylance J. Community-acquired pneumonia in sub-Saharan Africa. Semin Respir Crit Care Med **2016**; 37:855–67.
27. Carter R, Wolf J, van Opijnen T, et al. Genomic analyses of pneumococci from children with sickle cell disease expose host-specific bacterial adaptations and deficits in current interventions. Cell Host Microbe **2014**; 15:587–99.
28. Hausdorff WP, Hanage WP. Interim results of an ecological experiment—conjugate vaccination against the pneumococcus and serotype replacement. Hum Vaccin Immunother **2016**; 12:358–74.
29. Everett DB, Cornick J, Denis B, et al. Genetic characterisation of Malawian pneumococci prior to the roll-out of the PCV13 vaccine using a high-throughput whole genome sequencing approach. PLoS One **2012**; 7:e44250.
30. Kulohoma BW, Cornick JE, Chaguza C, et al. comparative genomic analysis of meningitis- and bacteremia-causing pneumococci identifies a common core genome. Infect Immun **2015**; 83:4165–73.
31. Luján M, Gallego M, Belmonte Y, et al. Influence of pneumococcal serotype group on outcome in adults with bacteraemic pneumonia. Eur Respir J **2010**; 36:1073–9.
32. Weinberger DM, Harboe ZB, Sanders EA, et al. Association of serotype with risk of death due to pneumococcal pneumonia: a meta-analysis. Clin Infect Dis **2010**; 51:692–9.
33. Fletcher MA, Schmitt HJ, Syrochkina M, Sylvester G. Pneumococcal empyema and complicated pneumonias: global trends in incidence, prevalence, and serotype epidemiology. Eur J Clin Microbiol Infect Dis **2014**; 33:879–910.
34. Nina Gratz LNL, Tuomanen E. *Streptococcus pneumoniae*. Molecular mechanisms of host-pathogen interactions. Chapter 23: pneumococcal invasion: development of bacteremia and meningitis. Cambridge, MA: Academic Press, **2015**.
35. Kremer PH, Lees JA, Koopmans MM, et al. Benzalkonium tolerance genes and outcome in *Listeria monocytogenes* meningitis. Clin Microbiol Infect **2017**; 23:265.e1–7.
36. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. Nat Commun **2014**; 5:5471.
37. Bille E, Ure R, Gray SJ, et al. Association of a bacteriophage with meningococcal disease in young adults. PLoS One **2008**; 3:e3885.
38. Kloek AT, van Setten J, van der Ende A, et al. Exome array analysis of susceptibility to pneumococcal meningitis. Sci Rep **2016**; 6:29351.
39. Tunjungputri RN, Mobegi FM, Cremers AJ, et al. Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. MBio **2017**; 8. doi:10.1128/mBio.01984-16.
40. Bensing BA, Rubens CE, Sullam PM. Genetic loci of *Streptococcus mitis* that mediate binding to human platelets. Infect Immun **2001**; 69:1373–80.
41. Lees JA, Croucher NJ, Goldblatt D, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife **2017**; 6. doi:10.7554/eLife.26255.
42. Lees JA, Kremer PH, Manso AS, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. Microb Genom **2017**; 3:e000103.
43. Maruvada R, Zhu L, Pearce D, Sapirstein A, Kim KS. Host cytosolic phospholipase $A_2\alpha$ contributes to group B *Streptococcus* penetration of the blood-brain barrier. Infect Immun **2011**; 79:4088–93.
44. Sitkiewicz I, Nagiec MJ, Sumby P, Butler SD, Cywes-Bentley C, Musser JM. Emergence of a bacterial clone with enhanced virulence by acquisition of a phage encoding a secreted phospholipase A2. Proc Natl Acad Sci U S A **2006**; 103:16009–14.
45. van Opijnen T, Lazinski DW, Camilli A. Genome-wide fitness and genetic interactions determined by Tn-seq, a high-throughput massively parallel sequencing method for microorganisms. Curr Protoc Microbiol **2015**; 36:1E.3 1–24.
46. 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.
47. Arnold BJ, Gutmann MU, Grad YH, et al. Weak epistasis may drive adaptation in recombining bacteria. Genetics **2018**; 208:1247–60.
48. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome Biol **2014**; 15:538.
49. Li Y, Metcalf BJ, Chochua S, et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting beta-lactam resistance levels in *Streptococcus pneumoniae*. MBio **2016**; 7.
50. Cohen J, Vincent JL, Adhikari NK, et al. Sepsis: a roadmap for future research. Lancet Infect Dis **2015**; 15:581–614.
51. Peacock S. Health care: bring microbial sequencing to hospitals. Nature **2014**; 509:557–9.