The contribution of genetic variation of *Streptococcus pneumoniae* to the clinical manifestation of invasive pneumococcal disease

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

[1] Laboratory of Pediatric Infectious Diseases, Radboudumc, Nijmegen, the Netherlands

[2] Department of Medical Microbiology, Radboudumc, Nijmegen, the Netherlands

[3] Bacterial Genomics Group, Center for Molecular and Biomolecular Informatics; Radboudumc, Nijmegen, the Netherlands

[4] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[5] Centre for Infectious Disease Control (CIb), RIVM National Institute for Public Health and the Environment, Bilthoven, the Netherlands

[6] Department of Medical Microbiology and Infectious Diseases, Canisius-Wilhelmina Hospital, Nijmegen, the Netherlands

[7] Wellcome Trust Sanger Institute, Pathogen Genomics group, Hinxton Cambridge, United Kingdom

[8] NIZO, Ede, the Netherlands

[9] Department of Biostatistics, University of Oslo, Oslo, Norway

[10] Department of Infectious Diseases and Immunology, Veterinary Faculty, Utrecht University, Utrecht, The Netherlands

*Corresponding author: Amelieke JH Cremers, MD PhD

Email: Amelieke.cremers@radboudumc.nl Phone: 0031-24-3619041

Abstract

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 two Dutch hospitals between 2000-2011 with pneumococcal bacteraemia. 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.

Findings The presence of pneumococcal gene *slaA* was a nationwide confirmed independent predictor of meningitis (OR=10.5, p=0.001), as was sequence cluster 9 (OR=3.68, p=0.057). A set of 4 pneumococcal genes co-located on a prophage was a confirmed independent predictor of 30-day mortality (OR=3.4, p=0.003). We could detect the pneumococcal variants of concern in these patients' blood samples by molecular amplification. In the post-vaccination cohort where the distribution of both patient characteristics and pneumococcal serotypes had changed, the relative importance of the prophage was no longer supported.

Interpretation Knowledge of pneumococcal genotypic variants improved our clinical risk assessment for detrimental manifestations of IPD. This provides us with novel opportunities to target, anticipate or avert the pathogenic effects that are related to particular pneumococcal variants. Therefore, future diagnostics should facilitate prompt appreciation of pathogen diversity in clinical sepsis management. Ongoing surveillance is warranted to monitor the clinical value of information on pathogen variants in dynamic microbial and susceptible host populations.

Funding None.

Background

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 further sterile body sites like the pleural cavity and cerebrospinal fluid, corresponding with the clinical syndromes bacteraemia, empyema, and meningitis respectively. The reasons for these phenomena to occur mainly include flaws in host defence.[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 importance of pneumococcal variation as determinant of invasive disease.[9,10]

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 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. [14]

The diversity in pneumococcal variants, illustrated by over 95 different capsular serotypes, has long been appreciated in pneumococcal vaccination and surveillance. *S. pneumoniae* is a naturally competent organism that fosters genetic recombination via transformation throughout its entire genome.[15] Although pneumococcal serotypes have been related to particular clinical manifestations of IPD,[16] it is unsure if the capsule is solely responsible.

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

Methods

Three clinical cohorts

The index cohort consisted of 349 patients diagnosed with a pneumococcal bacteraemia admitted to two 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 - December 2006 and June 2008 - May 2012 (periods for which clinical metadata were available) were randomly selected from the National Surveillance Database.[17,18] The temporal validation cohort was collected from one index hospital, and consisted of 121 pneumococcal bacteraemia 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 Immunisation Programme 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. Details on handling and formatting of clinical variables for each analysis are described in Supplementary methods 1. Normality of continuous variables was tested by Shapiro-Wilk test, and differences in characteristics in comparison to the index cohort were tested with a 2-sided Student's t-test or Mann-Whitney U test accordingly. Differences in nominal variables were tested by 2-sided Chi-square testing (Fisher's exact if less than 10 cases in any cell).

Pneumococcal blood isolates were stored in 10% glycerol 10% skim milk at -80^{o}C. Serotypes were determined by capsular PCR and Quellung reaction, confirmed by molecular capsular typing in whole genome sequenced strains. From isolates at the index hospitals, DNA was isolated with Qiagen Genomic-tip 20/G after culture to OD_{620} 0.2-0.3 in 10ml Todd Hewitt broth with 5% yeast extract at 37^{o}C. DNA template from the National Surveillance isolates was prepared as a lysate by heating a 2ml overnight culture at 90^{o}C for 10 minutes.

Genome-wide 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 (i.e. the identification of genetically diverged subpopulations which are called sequence clusters) were performed for the 349 isolates from the index cohort as previously described.[10]

The relationship between sequence cluster (SC) and 23 clinical IPD phenotypes was explored by stepwise regression analysis. Dependent variables were clinical IPD phenotypes, and each sequence cluster was entered as a binary independent variable. The models included a constant, with variable entry set at 0.05, and removal at 0.05 for logistic and at 0.1 for linear regression.

We performed two different genome wide association studies 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's exact with cluster permutation and by Cochran-Mantel-Haenszel analyses implemented in PLINK.[19] Sequence cluster (SC) was introduced as the 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 to 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-values < 1e-5 in univariate chi-square testing. Associations between the selected k-mers and binary clinical IPD phenotypes were assessed by sequence element enrichment (SEER) analysis,[20] including correction for population structure by multi-dimensional 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 multiple testing, the significance threshold was set at 1e-8.

Validation of associations

We aimed to validate a selection of the identified OGs that were significantly associated with a clinical IPD phenotype, irrespective of their functional annotation, in a nation-wide cohort. In the temporal validation cohort the number of identified genes evaluated was constrained by the number of cases in that collection period. The size of the validation cohorts was calculated to detect the index differences with a power of at least 0.8 and alpha of 0.05 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 on pneumococcal genomes in the validation cohorts, primers were designed and validated based on the index cohort, using a real-time fluorescent read out. The 20µl reaction mix contained 1x SsoAdvanced universal SYBR Green supermix, 200nM of each primer and as template either 0.005ng Qiagen genomic tip DNA or 200 times diluted pneumococcal lysate which yielded similar Ct-values. Cycling conditions were 95°C 3min; 40 cycles of 95°C 10sec and 55°C 30sec; 95°C 10sec; melting curve 65 to 95°C with 0.5°C/sec increase. All diluted templates were tested for detection of the *gyrA* pneumococcal housekeeping gene. All PCR runs included positive and negative control samples from the index cohort, plus negative extraction and PCR controls, and the specificity of produced amplicon for the OG of interest was confirmed by its melting temperature (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 PHASTER.[21] Sequence variation within the confirmed OGs and prophages was expressed in size, GC-content and pairwise distances. Distances were calculated from amino acid alignments, using the MEGA7 p-distance metric assuming gamma distribution with pairwise deletion of ambiguous positions.

Clinical relevance

The relative contribution of the identified pneumococcal genotypes to the clinical IPD phenotype in relation to well-known clinical predictors was assessed by logistic regression analysis as described elsewhere,[22] with the addition of the significantly associated sequence clusters.

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. The pneumococcal genomic DNA load in the serum samples was assessed previously.[23] We selected those serum samples on which capsular sequence typing had previously been successful.[24] DNA was isolated from 100µl of serum using Qiagen's DNeasy Blood and tissue kit. The OG validation PCRs (not matching human DNA sequences) were performed in duplicate using 8µl of template DNA and 50 amplification cycles.

Unless stated otherwise, the significance threshold was set at 0.05.

Results

Three clinical cohorts

Although the geographical validation cohort largely overlapped with the study period of the index cohort, serotypes appeared to be not evenly distributed (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.

Genome-wide analyses in the index cohort

Of the 23 tested clinical manifestations of IPD 87% appeared to be associated with one or more pneumococcal sequence clusters (SCs) (Table 1).

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

Another method was used to identify genome-wide associations with any allelic variant, or k-mer, present anywhere on the genome and the clinical IPD phenotype. The identified k-mers had nucleotide sequences that aligned with members of up to 6 different OGs. This number of origins was inversely related to k-mer size as well as to the proportion of core (versus accessory) OG origins. None of the 15,249,832 identified unique k-mers met the genome-wide significance threshold in their SC-independent association with clinical IPD phenotypes (Supplementary table 1). Despite this, certain OGs were overrepresented as they contained multiple variable regions related to a particular phenotype (Supplementary table 2).

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 successful for 8 out of 10 OGs selected for validation (Supplementary methods 3). The size of the temporal validation cohort only allowed for validation of the OGs associated with 30-day mortality. All diluted templates from isolates in both validation cohorts (n=603) were positive for the *gyrA* pneumococcal housekeeping gene with a Ct-value of 24 ± 2. Out of the nine OG-phenotype combinations from the index cohort tested, four were confirmed in the geographical validation cohort (Figure 3) and further characterised as described below.

Confirmed pneumococcal genotypes

The confirmed OG_2721 related to meningitis was functionally annotated as *slaA* coding for phospholipase A2, and showed 100% anti-occurrence with OG_416 (a predicted membrane protein) and OG_679 (ABC transporter) in the index cohort. The three 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 three cohorts (Figure 4). 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 OG_2721 were identical, other OGs showed large variation (Supplementary Figure 1). Within OG_17 and OG_58 the number of pairwise amino acid positions exceeded the number expected from their largest sequence variant. This suggests genetic mosaicism which is typical for bacteriophage genes. In the distribution of the confirmed OGs in the pneumococcal populations, OG_17 was taken as a proxy for its joint prophage vector shared with OG_675 and OG_58 (Supplementary Figure 2). In addition to presence, also the number of open reading frames per OG present in an isolate strongly correlated between these three OGs. While all confirmed OGs were present in both vaccine and non-vaccine serotypes, the relative occurrence of OG_2721 in IPD cases remained stable over time, yet OG_17 waned.

Clinical relevance

Relative to clinical predictors of meningitis and 30-day mortality, pneumococcal sequence clusters and orthologous genes were still major independent determinants of these phenotypes (Table 2).

OG_2721 associated with meningitis was correctly only detected by PCR in serum from patient PBCN0382 (Table 3). 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 pneumococcal genetic variants (sequence clusters and orthologous genes) to be independent determinants 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. In fact, instant knowledge of pneumococcal genotype could be used to improve future recognition of particular disease manifestations. Second, although mortality from IPD is more easy to establish, its determinants can vary widely across different clinical settings.[25] We have observed in our temporal post-vaccination 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 be 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,[26] antibiotic resistance level, and serotype distribution.[27] 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 efficient approach.

Our non-selective 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 the human and pneumococcal population differed from ours,[28] also the heavy selection for meningitis cases could have altered the relative contribution of certain pneumococcal variants.[29] Vice versa, a determinant identified from an artificial distribution of cases (with unnatural pre-odds), may no longer be valid among patient populations presenting to the hospital.

In general, pneumococcal population structures are characterized by linkage disequilibrium, which means that particular groups of sequences (including the capsular sequence variant) co-occur together on a pneumococcal genome. In our GWAS analyses, to prevent identification of sequences that actually represent a magnitude of co-occurring genes, we corrected for this population structure. Also, we assessed whether these so-called sequence clusters as a whole were related to clinical IPD phenotypes, and we found a remarkable concordance to previous serotype-based studies.[30-32]

Our exclusion of strain- and 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.*[33] At the same time, this aspect of our approach may have favoured the detection of bacterial genotypes located on prophage vectors to be associated with clinical phenotypes as found by us and by others previously.[34] Unlike many other genes in clonal populations, the distribution of bacteriophages is not as strictly determined by lineage. 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 without bias from correction for population structure.[35] In any case, if indeed multiple proteins encoded by prophage elements have meaningful interactions with human cells during bloodstream infections, it is more likely that this prophage trait was fixated because of some fitness advantage the bacteriophage or the lysogenic pneumococcus experiences during colonization at the respiratory mucosal surface from where it can actually acquire a viable host.

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 OG_17 *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 OG_2721 being identical, k-mers originating from OG_2721 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 selection criteria. These examples demonstrate the complementarity of the two different GWAS methods employed.

While we identified clinical IPD phenotypes to be associated with pneumococcal genes that were independent of pneumococcal lineage (serotype) and clinical predictors, this does not prove causality. We have not included potential confounders like host genotype, a host factor that mediates susceptibility to meningitis[36] and may simultaneously induce a mucosal environment that welcomes specific pneumococcal variants. On the other hand, 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 a *pblB*-positive *S. pneumoniae* compared to its knock out variant.[22] The predicted function of the protein encoded by the *pblB* gene (e.i. functional annotation based on homology) is a phage tail fiber, and its *S. mitis* orthologue was shown to bind to human platelets as well.[37] Although we have studied pneumococcal blood isolates, it has been shown based on genomic data no adaptation is needed to cross the blood-brain barrier, so DNA sequences from blood isolates seem representative for pneumococci that reach the cerebrospinal fluid and cause meningitis.[38] 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.*[39] In group A *Streptococcus* the presence of *slaA* enhanced the bacterium's potential for epithelial adherence, colonization and invasive disease.[40] Also for *slaA* further studies would be required to elucidate its effects *in vivo* and possibilities to avert these during disease. Furthermore, recent advancements in methods to study functional interactions on pneumococcal genomes [41,42] may help to improve our understanding of why particular sequence clusters are overrepresented in certain phenotypes.

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,[43] outbreak management and combating antimicrobial resistance,[44] may now be extended to individual patient care. Increased appreciation of eliciting microbial variants, could push the tailored adjunctive measures that are heavily searched for in clinical sepsis care.[45]

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. While a systems biology approach may unravel the exact pathophysiology, prompt molecular diagnostics at the emergency department could readily improve risk stratification and alertness for complicated infection in individual patient care.[46]

References

1. Weiser JN. The pneumococcus: why a commensal misbehaves. *J Mol Med (Berl)* 2010; **88**(2): 97-102.
2. 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**(9693): 893-902.
3. Wunderink RG. CAP death: what goes wrong when everything is right? *Lancet Infect Dis* 2015; **15**(9): 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**(6): 451-9.
5. van der Poll T, Opal SM. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *Lancet* 2009; **374**(9700): 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**(3): 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**(9): 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**(6): 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**(10): 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. Cilloniz C, Gabarrus A, Almirall J, et al. Bacteraemia in outpatients with community-acquired pneumonia. *Eur Respir J* 2016; **47**(2): 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**(21): 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**(2): 172-9.
14. McGill F, Heyderman RS, Panagiotou S, Tunkel AR, Solomon T. Acute bacterial meningitis in adults. *Lancet* 2016; **388**(10063): 3036-47.
15. Croucher NJ, Harris SR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. *Science* 2011; **331**(6016): 430-4.
16. Hausdorff WP, Feikin DR, Klugman KP. Epidemiological differences among pneumococcal serotypes. *Lancet Infect Dis* 2005; **5**(2): 83-93.
17. Netherlands Reference Laboratory for Bacterial Meningitis (AMC/RIVM). URL: https://www.amc.nl/web/Research/Overview/Departments/Medical-Microbiology/Medical-Microbiology/Current-research/Reference-Laboratory-for-Bacterial-Meningitis.htm?print=true, accessed 24/07/2017.
18. 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**(8): 1077-85.
19. 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**(3): 559-75.

20. Lees JA, Vehkala M, Valimaki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun* 2016; **7**: 12797.

21. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016; **44**(W1): W16-21.

22. 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**(1).

23. 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**(7): 1119-24.

24. 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.

25. Aston SJ, Rylance J. Community-Acquired Pneumonia in Sub-Saharan Africa. *Semin Respir Crit Care Med* 2016; **37**(6): 855-67.

26. 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**(5): 587-99.

27. Hausdorff WP, Hanage WP. Interim results of an ecological experiment - Conjugate vaccination against the pneumococcus and serotype replacement. *Hum Vaccin Immunother* 2016; **12**(2): 358-74.

28. 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**(9): e44250.

29. 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**(10): 4165-73.

30. Lujan M, Gallego M, Belmonte Y, et al. Influence of pneumococcal serotype group on outcome in adults with bacteraemic pneumonia. *Eur Respir J* 2010; **36**(5): 1073-9.

31. 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**(6): 692-9.

32. 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**(6): 879-910.

33. Brown JM, Hammerschmidt S, Orihuela C. *Streptococcus pneumoniae* Molecular Mechanisms of Host-Pathogen Interactions. *Elsevier inc*. 2015 USA. Chapter 23 Pneumococcal Invasion: Development of Bacteremia and Meningitis; Gratz N, Loh LN, Tuomanen E; 433-451.

34. Kremer PH, Lees JA, Koopmans MM, et al. Benzalkonium tolerance genes and outcome in *Listeria monocytogenes* meningitis. *Clin Microbiol Infect* 2016.

35. Bille E, Ure R, Gray SJ, et al. Association of a bacteriophage with meningococcal disease in young adults. *PLoS One* 2008; **3**(12): e3885.

36. Kloek AT, van Setten J, van der Ende A, et al. Exome Array Analysis of Susceptibility to Pneumococcal Meningitis. *Sci Rep* 2016; **6**: 29351.

37. Bensing BA, Rubens CE, Sullam PM. Genetic loci of *Streptococcus mitis* that mediate binding to human platelets. *Infect Immun* 2001; **69**(3): 1373-80.

38. 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**(1): e000103.

39. 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**(10): 4088-93.

40. 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**(43): 16009-14.

41. 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.

42. 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**(2): e1006508.

43. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol* 2014; **15**(11): 538.

44. 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**(3).

45. Cohen J, Vincent JL, Adhikari NK, et al. Sepsis: a roadmap for future research. *Lancet Infect Dis* 2015; **15**(5): 581-614.

46. Peacock S. Health care: Bring microbial sequencing to hospitals. *Nature* 2014; **509**(7502): 557-9.

Acknowledgements

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

Declaration of interests

All authors declare to have no conflicts of interest.

Tables

Table 1 Sequence cluster as determinant of clinical IPD phenotype in the index cohort

Phenotype	SC	β-coefficient	OR	95% CI			p-value
Host specificity							
Age	SC10	-14,1		-21,0	-	-7,3	<0,001
	SC9	-11,7		-18,1	-	-5,3	<0,001
	NA	-7,9		-12,9	-	-2,9	0,002
	SC8	-12,8		-22,9	-	-2,8	0,012
Male	SC1		0,2	0,1	-	0,7	0,013
Charlson comorbidity score	SC10	-1,6		-2,5	-	-0,7	0,001
	SC9	-1,3		-2,1	-	-0,4	0,003
COPD	SC4		<<				
	SC5		0,1	0,0	-	1,1	0,058
Diabetes mellitus	SC1		3,0	1,1	-	8,1	0,025
Cancer	SC4		4,1	1,3	-	12,6	0,014
	SC9		0,2	0,0	-	0,8	0,027
Immunocompromising therapy	SC10		<<				
Cardiovascular disease	SC1		4,0	1,4	-	11,5	0,009
Antibiotics prior to admission	NA		3,4	1,0	-	11,4	0,050
Influenza season	None						
Year of infection	SC7	-1,4		-2,7	-	-0,1	0,037
Presentation							
SIRS	SC3		>>				
	SC6		6,2	0,8	-	46,8	0,076
Cough	SC11		0,3	0,1	-	0,9	0,037
CRP	NA	-71,8		-116,9	-	-26,6	0,002
	SC12	-101,2		-185,1	-	-17,2	0,018
	SC10	73,2		11,0	-	135,4	0,021
Leukocytes	SC5	-5,4		-9,7	-	-1,1	0,015
	SC4	5,9		0,3	-	11,5	0,041
Pneumonia	NA		0,5	0,3	-	0,9	0,015
	SC6		4,6	1,1	-	19,9	0,042
	SC8		>>				
	SC10		7,5	1,0	-	56,9	0,050
PSI risk class	SC10	-1,1		-1,4	-	-0,7	<0,001
	SC9	-0,9		-1,3	-	-0,5	<0,001
Pleural effusion	NA		0,5	0,2	-	0,9	0,022
Empyema	None						
Meningitis	NA		6,0	2,6	-	14,1	<0,001
	SC9		3,5	1,1	-	10,8	0,032
Unknown focus of infection	SC6		<<				
	SC10		<<				
Course							
30-day mortality	SC4		<<				
	SC8		<<				
	SC9		0,2	0,0	-	1,2	0,080
	SC10		0,2	0,0	-	1,4	0,108
Early death	None						

>> / <<: all / none of the cases assigned this sequence cluster displayed the phenotype.

Abbreviations: IPD: invasive pneumococcal disease; SC: sequence cluster; OR: odds ratio; 95%-CI: 95% confidence interval; NA: assembly of strains not assigned to a particular sequence cluster; COPD: chronic obstructive pulmonary disease; SIRS: systemic inflammatory response syndrome; PSI: pneumonia severity index.

Table 2 Optimized prediction models for meningitis and 30-day mortality

Phenotype	**Determinant**	**OR**	**95% CI**	**p-value**
Meningitis	Cough	0.06	0.02 - 0.22	8.6E-6
	OG_2721 *slaA*	10.5	2.61 - 42.30	0.001
	Age	0.97	0.94 - 0.99	0.006
	SC9 (serotype 7F)	3.68	0.96 - 14.05	0.057
30-day mortality	Charlson comorbidity score	1.46	1.24 - 1.72	7.0E-4
	OG_17 *pblB*	3.40	1.53 - 7.59	0.003
	Meningitis	4.61	1.55 - 13.74	0.006
30-day mortality among pneumonia cases	Charlson comorbidity score	1.34	1.06 - 1.68	0.013
	OG_17 *pblB*	3.28	1.18 - 9.11	0.023
	PSI risk class	2.22	1.07 - 4.63	0.033

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

Abbreviations: OR: odds ratio; 95% CI: 95% confidence interval; OG: orthologous gene; SC: sequence cluster; PSI: pneumonia severity index.

Table 3 Detection of orthologous gene sequences in serum from IPD patients

Study ID		**PBCN0382**	**PBCN0389**	**PBCN0420**	**PBCN0480**	**PBCN0442**
Meningitis		yes	no	no	yes	no
30-day mortality		no	yes	yes	yes	no
Serum pneumococcal DNA load (copies/ml)		8E+02	7E+03	3E+03	3E+03	1E+04
OG_2721 *slaA*	Isolate whole genome sequencing [a]	1	0	0	0	0
	Serum OG-PCR [b]	2	0	0	0	0
OG_17 *pblB*	Isolate whole genome sequencing	2	1	1	1	0
	Serum OG-PCR	2	2	1	1	0
OG_675	Isolate whole genome sequencing	2	1	1	1	0
	Serum OG-PCR	2	1	2	2	0
OG_58	Isolate whole genome sequencing	2	1	1	1	0
	Serum OG-PCR	2	2	2	2	2
Prophage sequence	Isolate whole genome sequencing	partial	partial	complete	partial	absent

The results in the matrix represent: [a] number of open reading frames assigned; [b] times target detected in duplicate OG-PCRs.

Abbreviations: IPD: invasive pneumococcal disease; ID: identifier; PBCN: pneumococcal bacteraemia collection Nijmegen; OG: orthologous gene; OG-PCR: orthologous gene polymerase chain reaction.

Figure legends

Figure 1 Cohort characteristics.

The index cohort consisted of 349 patients with a pneumococcal bacteraemia admitted to 2 local hospitals (H), the geographical validation cohort of 482 patients in nationwide IPD surveillance, and the temporal validation cohort of 121 patients admitted to the index hospitals during a later time period (panel A). Main cohort characteristics are presented as mean ± standard deviation, median (interquartile range), or percentage fulfilling the condition (N/N known) (panel B).

*:p<0.05; **:p<0.01; ***:p<0.001.

Abbreviations: PSI: pneumonia severity index; PCV7: serotypes 4, 6B, 9V, 14, 18, and 23F; PCV10: serotypes 1, 5, and 7F; PCV13: serotypes 3, 6A, and 19A; NVT: all other (non vaccine) serotypes.

P-values support differences from the index cohort.

Figure 2 Clinical IPD phenotypes with associated orthologous pneumococcal genes in index cohort.

Rows represent 349 IPD cases and corresponding pneumococcal blood isolates. The tree on the left represents their relative phylogenetic position based on SNPs in the core genome, in which sequence clusters 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 associated OGs that passed Fisher's exact with p < 0.01 and were independent of population structure are displayed. The OGs selected for validation are indicated by an arrow.

Abbreviations: IPD: invasive pneumococcal disease; OG: orthologous gene; SC: sequence cluster.

Figure 3 Geographical and temporal validation of orthologous gene associations.

The prevalence of pneumococcal OGs in the 3 cohorts (panel A). The association between absence (-, empty bar) or presence (+, filled bar) of and OG and the proportion of patients affected by a particular

IPD phenotype in the index and 2 validation cohorts (cancer: panel B; meningitis: panel C; pneumonia: panel D; 30-day mortality: panel E).

*:p<0.1; **:p<0.01; ***:p<0.001.

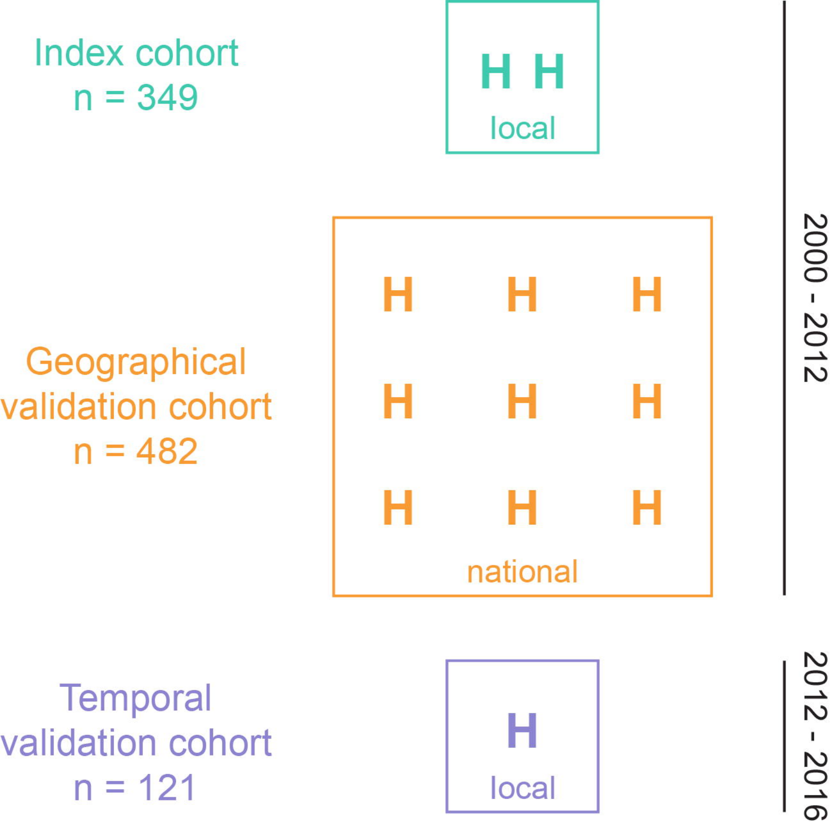
Abbreviations: OG: orthologous gene; IPD: invasive pneumococcal disease.

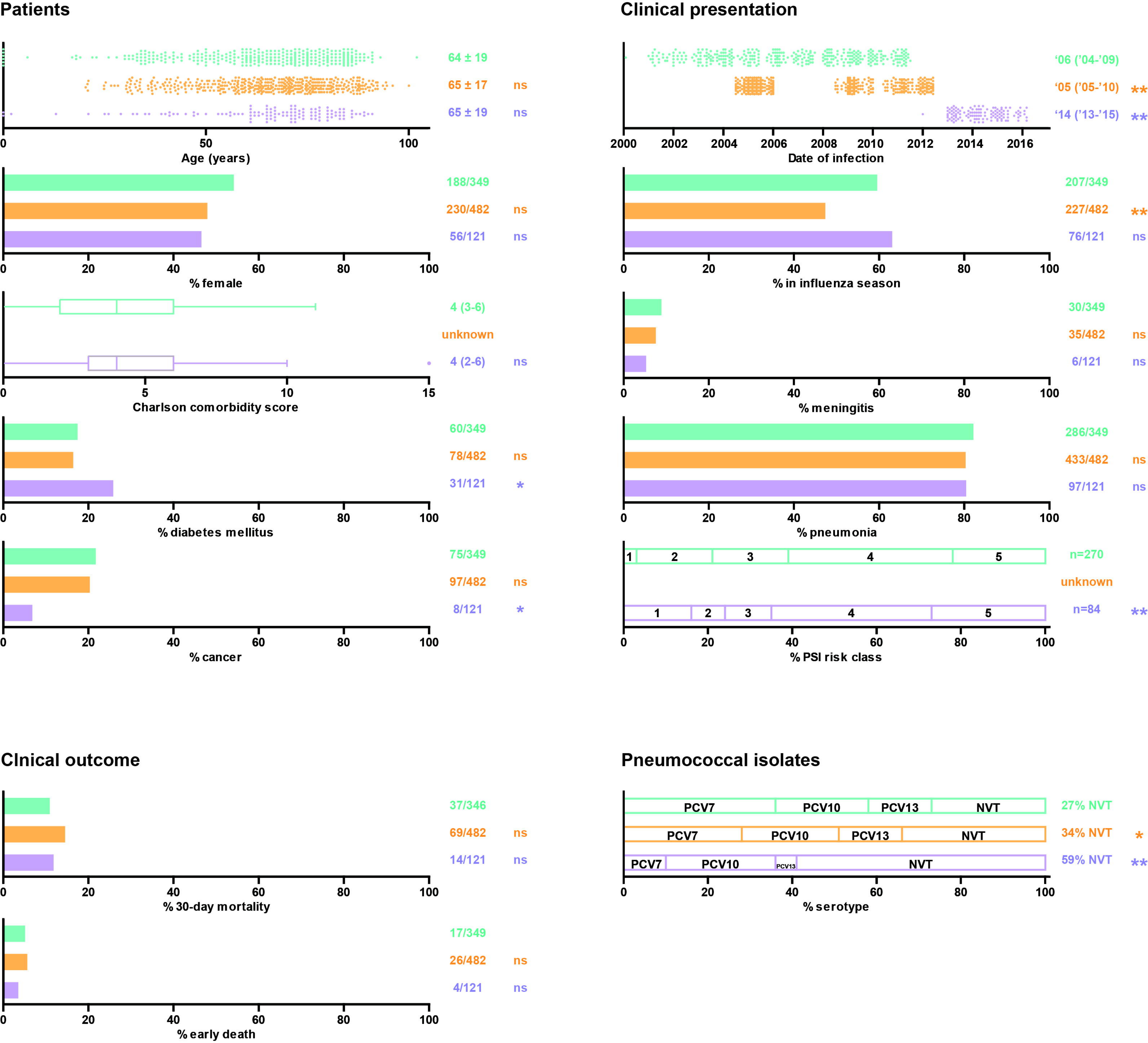
P-values indicate differences in the proportion of patients affected by a particular IPD phenotype.

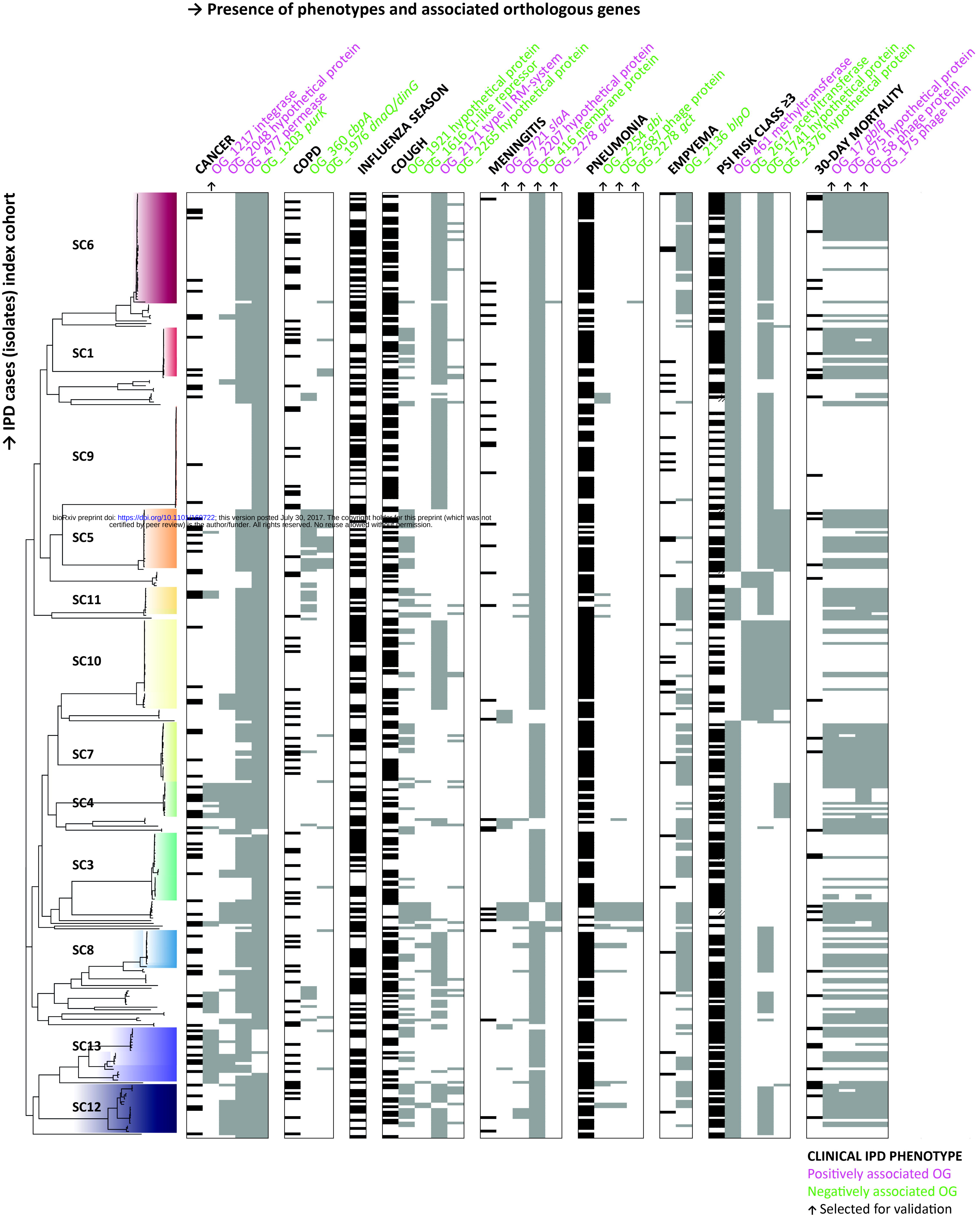
Figure 4 Co-occurrence of mortality-related orthologous genes on single prophage.

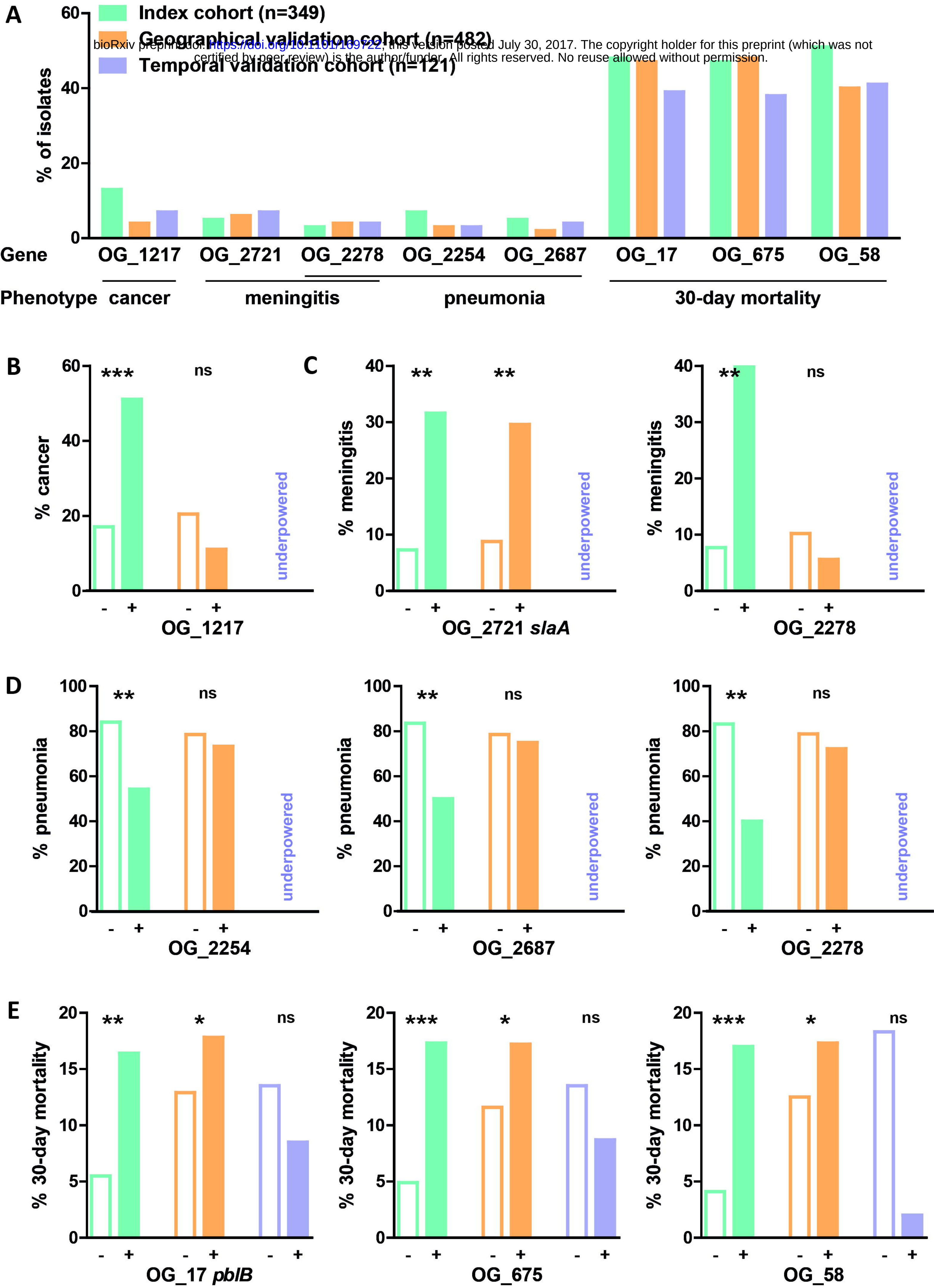
A. The level of co-occurrence of 3 pneumococcal OGs is expressed in phi coefficient (in bold) for the index, geographical, and temporal cohort (from left to right). B. Gene card depicting an example of the co-localisation of 4 OGs associated with 30-day mortality on a single prophage in pneumococcal isolate PBCN0420.

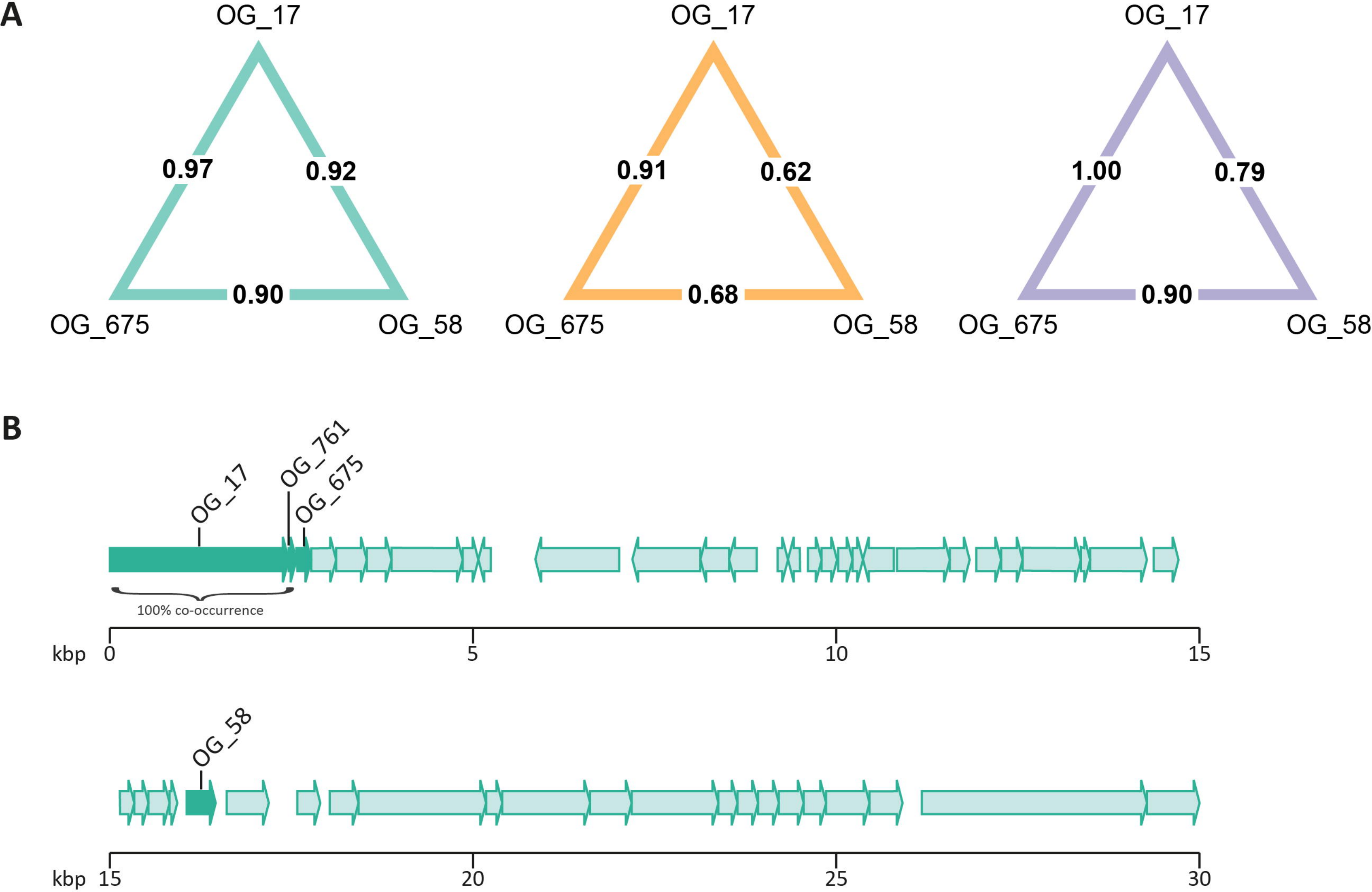
Abbreviations: OG: orthologous gene; PBCN: pneumococcal bacteraemia collection Nijmegen; kbp: kilobase pair.


Index cohort
n = 349
H H
local
2000 - 2012
Geographical validation cohort
n = 482
H H H
H H H
H H H
national
Temporal validation cohort
n = 121
H
local
2012 - 2016


Patients
64 ± 19
65 ± 17
ns
65 ± 19
ns
0
50
100
Age (years)
188/349
230/482
ns
56/121
ns
0
20
40
60
80
100
% female
4 (3-6)
unknown
4 (2-6)
ns
0
5
10
15
Charlson comorbidity score
60/349
78/482
ns
31/121
*
0
20
40
60
80
100
% diabetes mellitus
75/349
97/482
ns
8/121
*
0
20
40
60
80
100
% cancer
Clinical presentation
'06 ('04-'09)
'05 ('05-'10)
**
'14 ('13-'15)
**
2000
2002
2004
2006
2008
2010
2012
2014
2016
Date of infection
207/349
227/482
**
76/121
ns
0
20
40
60
80
100
% in influenza season
30/349
35/482
ns
6/121
ns
0
20
40
60
80
100
% meningitis
286/349
433/482
ns
97/121
ns
0
20
40
60
80
100
% pneumonia
1
2
3
4
5
n=270
unknown
1
2
3
4
5
n=84
**
0
20
40
60
80
100
% PSI risk class
Clnical outcome
37/346
69/482
ns
14/121
ns
0
20
40
60
80
100
% 30-day mortality
17/349
26/482
ns
4/121
ns
0
20
40
60
80
100
% early death
Pneumococcal isolates
PCV7
PCV10
PCV13
NVT
27% NVT
PCV7
PCV10
PCV13
NVT
34% NVT
*
PCV7
PCV10
PCV13
NVT
59% NVT
**
0
20
40
60
80
100
% serotype


→ Presence of phenotypes and associated orthologous genes
→ IPD cases (isolates) index cohort
CANCER
OG_1217 integrase
OG_2048 hypothetical protein
OG_477 permease
OG_1203 purK
COPD
OG_360 cbpA
OG_1976 dnaQ/dinG
INFLUENZA SEASON
COUGH
OG_1921 hypothetical protein
OG_1616 CI-like repressor
OG_2171 type II RM-system
OG_2265 hypothetical protein
MENINGITIS
OG_2721 slaA
OG_2207 hypothetical protein
OG_416 membrane protein
OG_2278 gct
PNEUMONIA
OG_2254 abi
OG_2687 phage protein
OG_2278 gct
EMPYEMA
OG_2136 blpO
PSI RISK CLASS ≥3
OG_461 methyltransferase
OG_2617 acetyltransferase
OG_1741 hypothetical protein
OG_2376 hypothetical protein
30-DAY MORTALITY
OG_17 pblB
OG_675 hypothetical protein
OG_58 phage protein
OG_175 phage holin
SC6
SC1
SC9
SC5
SC11
SC10
SC7
SC4
SC3
SC8
SC13
SC12
CLINICAL IPD PHENOTYPE
Positively associated OG
Negatively associated OG
↑ Selected for validation


A
Index cohort (n=349)
Geographical validation cohort (n=482)
Temporal validation cohort (n=121)
% of isolates
0
20
40
60
Gene
OG_1217
OG_2721
OG_2278
OG_2254
OG_2687
OG_17
OG_675
OG_58
Phenotype
cancer
meningitis
pneumonia
30-day mortality
B
% cancer
***
ns
underpowered
- +
- +
OG_1217
C
% meningitis
**
**
underpowered
OG_2721 slaA
% meningitis
**
ns
underpowered
OG_2278
D
% pneumonia
**
ns
underpowered
OG_2254
% pneumonia
**
ns
underpowered
OG_2687
% pneumonia
**
ns
underpowered
OG_2278
E
% 30-day mortality
**
*
ns
- +
- +
- +
OG_17 pblB
% 30-day mortality
***
*
ns
OG_675
% 30-day mortality
***
*
ns
OG_58


A
OG_17
0.97
0.92
0.90
OG_675
OG_58
OG_17
0.91
0.62
0.68
OG_675
OG_58
OG_17
1.00
0.79
0.90
OG_675
OG_58
B
OG_17
OG_761
OG_675
100% co-occurrence
kbp
0
5
10
15
OG_58
kbp
15
20
25
30

Supplementary methods

Supplementary methods 1 Details in handling of clinical variables

Data collection

For all 3 cohorts, clinical data were collected from medical charts and registered with patient identifiers in a secured source file. In a separate working file, labeled clinical data were stored together with the non-identifying study code assigned to each IPD case included. The following clinical data were only collected for patients admitted to the index hospitals: Charlson comorbidity score, COPD, cardiovascular disease, antibiotics prior to admission, SIRS, blood CRP and leukocyte count, PSI risk class, and time to death. Otherwise, clinical data were handled in an equal manner across the 3 cohorts.

Variables

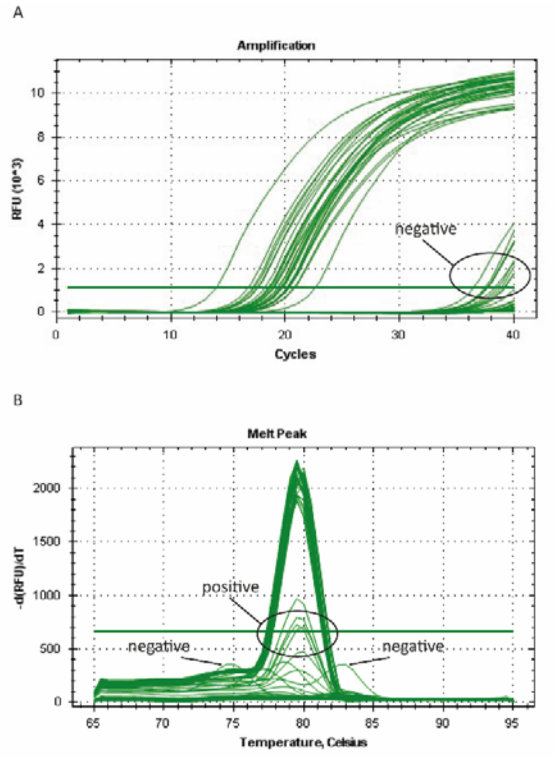
The following clinical variables were defined, processed or classified in a particular way. Included invasive pneumococcal disease cases: *S. pneumoniae* isolated from culture of cerebrospinal fluid or blood. Charlson comorbidity score: calculated for cases ≥18 years old. Immunocompromising therapy: actual use of systemic corticosteroids or chemotherapy. 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. Date of infection: date of blood culture collection. Influenza season: defined annually from the first to the last week with >5 reported influenza cases in the Netherlands as reported by WHO FluNet (http://apps.who.int/flumart/Default?ReportNo=12). Systemic inflammatory response syndrome: percentage of immature neutrophils not accounted for. Cough: as reported or observed at admission.

Pneumonia, empyema, and meningitis were not mutually exclusive. Pneumonia, meningitis, and unknown focus of infection: as reported by the attending physician. Pleural effusion: as reported by the attending radiologist. Empyema: *S. pneumoniae* isolated from pleural fluid culture. CRP, leukocytes, and other chemistry results included in clinical algorithms: at day of admission, except for bilirubin and albumin during hospital stay. Pneumonia severity index score: calculated for cases ≥18 years old who suffered from pneumonia, formatted into PSI risk class and stratified to PSI risk class 1-2 versus 3-5. 30-day mortality: in hospital death within 30 days from admission. Early death: in hospital death within 48 hours from admission.

Missing data

Missing data were not replaced. Data were considered to be missing if the corresponding section was not present in the medical chart. 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.

Supplementary methods 2 Interpretation of real time PCR amplification and melting curves



Supplementary methods 3 Primer characteristics

Clinical phenotype	OG	Name primer pair	Sequence forward primer (5'-3')	Sequence reverse primer (5'-3')	Amplicon size (bp)	MT target (°C)	MT side-product (°C)
Cancer	**1217**	OG_1217	TGTGGATGGAAGAACTTCCC	CCCTTATACCGAGAAATGG	506	74.5	76.5
Meningitis	**2721**	OG_2721	CGTAACGGAGAATTGTTGAAGAG	CGATTGGAGCTATAGGATGTTG	443	78	
	2207	OG_2207	Sequence not compatible with design real time primer pair				
	416	OG_416	Inadequate distinction between positive and negative samples				
	2278	OG_2278_set1_F1	GAAACGTGCTAAACAGCTAGG	GGGTGATGATTGGAAAGGTA	233	77	
		OG_2278_set1_F2	GAGAGCAAAGCAATTAGGAG	GGGTGATGATTGGAAAGGTA	228	77.5	
		OG_2278_set1_F3	GGTGACTATCTAGTGGTAG	GGGTGATGATTGGAAAGGTA	214	78	
Pneumonia	**2254**	OG_2254_set1	CTAAAGCAGCTAAGTACCTGC	CCACCAGAAACCTTGATATC	593	79-80.5	
		OG_2254_set2	GCATATCCAACACCATACGC	GCTTCCACACAATACGCTCA	551	77	
	2687	OG_2687	GAACATCTTCATGAACGC	CCCTAATTTCTATAGAAGACGC	125	77	
	2278	OG_2278_set1_F1	GAAACGTGCTAAACAGCTAGG	GGGTGATGATTGGAAAGGTA	233	77	
		OG_2278_set1_F2	GAGAGCAAAGCAATTAGGAG	GGGTGATGATTGGAAAGGTA	228	77.5	
		OG_2278_set1_F3	GGTGACTATCTAGTGGTAG	GGGTGATGATTGGAAAGGTA	214	78	
30-day	**17**	OG_17	TACAGCTGTGAAAGCCTTGG	CCTGAGAATCCAGATGGCTATC	161	80	84

mortality	**675**	OG_675	CGTTGCAAGAATGTAAGCGATGA	CAGAGGGCAATCCTGACT	208	80.5-82
	58	OG_58	GCTTGACGGCTACGAGG	CGGCTGGGTGTTTGATTG	174-195	78.5-82

To detect all sequence variants of OG_2278 3 different forward primers were used in separate reactions.

Abbreviations: OG: orthologous gene; bp: base pairs; MT: melting temperature.

Supplementary tables

Supplementary table 1 Individual k-mers associated with clinical IPD phenotype in index cohort p<e-6

Phenotype	K-mer	OG	Core OG	Annotation	Direction	p-value
Pneumonia	X10_9	856	no	gp19	neg	7E-07
	X10_9	163	no	zinc metalloproteinase	neg	7E-07
Meningitis	X11_41	2093	no	transmembrane protein	pos	6E-07
	X11_32	340	yes	DNA polymerase III subunits gamma and tau	neg	7E-07
	X11_86	10	no	ABC transporter ATP-binding protein/permease	pos	9E-07
Immunocompromising therapy	X3_1	35	no	endo-beta-N-acetylglucosaminidase	neg	3E-07
	X3_1	212	no	ABC transporter ATP-binding protein/permease	neg	3E-07
	X3_1	175	no	phage holin	neg	3E-07
	X3_1	1078	yes	phosphatase	neg	3E-07
Diabetes mellitus	X5_1	1415	yes	6-phosphogluconolactonase	neg	8E-08
	X5_5	521	yes	ABC transporter permease	neg	8E-08
Antibiotics prior to admission	X8_1	870	no	UDP-glucuronate 5'-epimerase	neg	5E-09
	X8_1	255	yes	excinuclease ABC subunit A	neg	5E-09
	X8_1	1277	yes	flavoprotein	neg	5E-09
	X8_10	323	yes	alpha-amylase	neg	5E-09
	X8_2	335	no	metallo-beta-lactamase superfamily protein	neg	5E-09
	X8_2	169	no	hypothetical protein	neg	5E-09
	X8_3	1701	no	BlpM	neg	5E-09
	X8_4	216	yes	DNA polymerase I - 3'-5' exonuclease and polymeras	neg	5E-09
	X8_5	967	yes	hypothetical protein	neg	5E-09
	X8_5	610	no	serine/threonine protein kinase	neg	5E-09
	X8_5	243	no	HAD superfamily hydrolase	neg	5E-09
	X8_7	546	no	single-stranded DNA-binding protein	neg	5E-09
	X8_7	3012	no	hypothetical protein	neg	5E-09
	X8_14	192	yes	single-stranded DNA-specific exonuclease	pos	5E-09
	X8_14	1533	no	UDP-N-acetylglucosamine-2-epimerase	pos	5E-09
	X8_15	6	no	cell-division ATP-binding protein FtsE	pos	5E-09
	X8_16	1266	yes	GIY-YIG domain-containing protein	pos	5E-09
	X8_17	321	yes	cation transporter E1-E2 family ATPase	pos	5E-09
	X8_18	2968	no	GNAT family acetyltransferase	pos	5E-09
	X8_18	221	no	tyrosine-protein phosphatase CpsB	pos	5E-09
	X8_19	932	yes	xanthine phosphoribosyltransferase	pos	5E-09
	X8_19	670	no	lactoylglutathione lyase	pos	5E-09
	X8_20	1359	no	dTDP-4-dehydrorhamnose 3	pos	5E-09
	X8_21	26	no	oligopeptide binding lipoprotein	pos	5E-09
	X8_22	12	no	IS1239 transposase	pos	5E-09
	X8_22	1232	no	branched-chain aa ABC transporter permease	pos	5E-09
	X8_23	347	yes	Gfo/Idh/MocA family oxidoreductase	pos	5E-09
	X8_23	10	no	ABC transporter ATP-binding protein/permease	pos	5E-09
	X8_26	21	no	cell envelope integrity inner membrane protein Tol	pos	5E-09
	X8_29	669	yes	dihydrofolate reductase	pos	5E-09

Supplementary table 2 OG-clustered k-mers associated with clinical IPD phenotype in index cohort

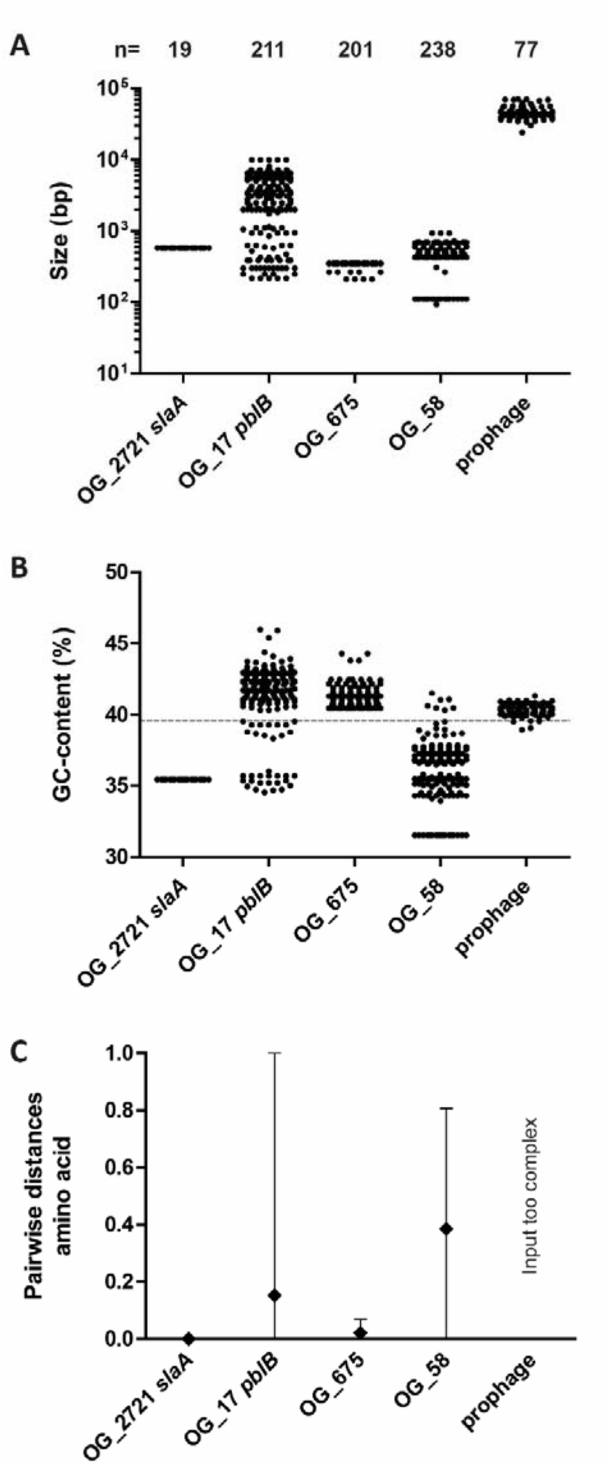
Phenotype	OG	Annotation	# k-mers	Direction	avg p-value
Pneumonia	5	choline binding protein A	2	neg	4E-06
	856	gp19	2	neg	4E-06
Meningitis	9	zinc metalloprotease	27	pos	5E-06
	2725	O-Antigen ligase	5	pos	8E-06
	1613	phage protein	4	pos	4E-06
	1	transposase, IS1193	4	pos	7E-06
	876	cof family protein	3	both	
	2721	phospholipase A2 SlaA	3	pos	5E-06
	1216	SNF2 family protein	3	pos	6E-06
	8	competence factor transporting ATP-binding protein	3	pos	6E-06
	1076	CvpA family protein	3	neg	2E-06
	5	choline binding protein A	2	both	
	43	endo-alpha-N-acetylgalactosaminidase	2	both	
	225	ABC zinc transporter	2	both	
	301	ROK family protein	2	both	
	412	DNA ligase	2	both	
	423	Esterase	2	both	
	1452	serine/threonine metallophosphatase	2	both	
	2093	transmembrane protein	2	pos	2E-06
	1374	TraG/TraD family protein, Type IV secretory pathway	2	pos	3E-06
	271	ABC-type transport system involved in multi-copper	2	pos	4E-06
	313	caax amino protease family protein	2	pos	4E-06
	22	glycosyl transferase family protein	2	pos	4E-06
	2053	hypothetical protein	2	pos	5E-06
	1799	capsular polysaccharide synthesis protein	2	pos	6E-06
	1095	glycoside hydrolase family protein	2	pos	7E-06
	41	PTS system lactose-specific transporter subunit II	2	pos	7E-06
	24	choline-binding protein F	2	pos	8E-06
	2947	Restriction endonuclease BpuJI - N terminal	2	pos	8E-06
	1616	CI-like repressor	2	pos	9E-06
	47	ABC transporter ATP-binding protein	2	neg	2E-06
	2429	AAA ATPase	2	neg	2E-06
	555	D-ala ligase	2	neg	4E-06
	568	SpeK	2	neg	4E-06
	470	lantibiotic modifying enzyme	2	neg	8E-06
	519	ABC transporter-sugar transport, N-acetylneuramina	2	neg	8E-06
	747	ABC transporter substrate-binding protein	2	neg	8E-06
	1392	lantibiotic modifying enzyme, serine/threonine prot	2	neg	8E-06
Early death	1531	FtsK/SpoIIIE protein	5	pos	4E-06
	318	MurD D-glutamic acid adding enzyme	4	both	4E-06
	1938	oligosaccharide repeat unit polymerase Wzy	3	pos	4E-06
	252	Integrase	2	both	
	163	G5 domain family	2	pos	4E-06
	840	flippase	2	pos	4E-06
	1441	SAP domain-containing protein	2	pos	4E-06
	33	beta-galactosidase	2	neg	4E-06
	43	endo-alpha-N-acetylgalactosaminidase	2	neg	4E-06
	941	orotate phosphoribosyltransferase	2	neg	4E-06
	1597	competence associated protein	2	neg	4E-06
Diabetes mellitus	521	ABC transporter permease	2	neg	6E-07
Antibiotics prior to admission	1277	flavoprotein	4	both	5E-09
	6	cell-division ATP-binding protein FtsE	3	both	
	21	cell envelope integrity inner membrane protein Tol	3	pos	5E-09
	243	HAD superfamily hydrolase	3	both	5E-09

	335	metallo-beta-lactamase superfamily protein	3	both	5E-09
	669	dihydrofolate reductase	3	pos	5E-09
	192	single-stranded DNA-specific exonuclease	2	pos	5E-09
	546	single-stranded DNA-binding protein	2	neg	5E-09
	744	galactose mutarotase-like enzyme, protein LacX	2	neg	8E-06

54 **Supplementary figures**

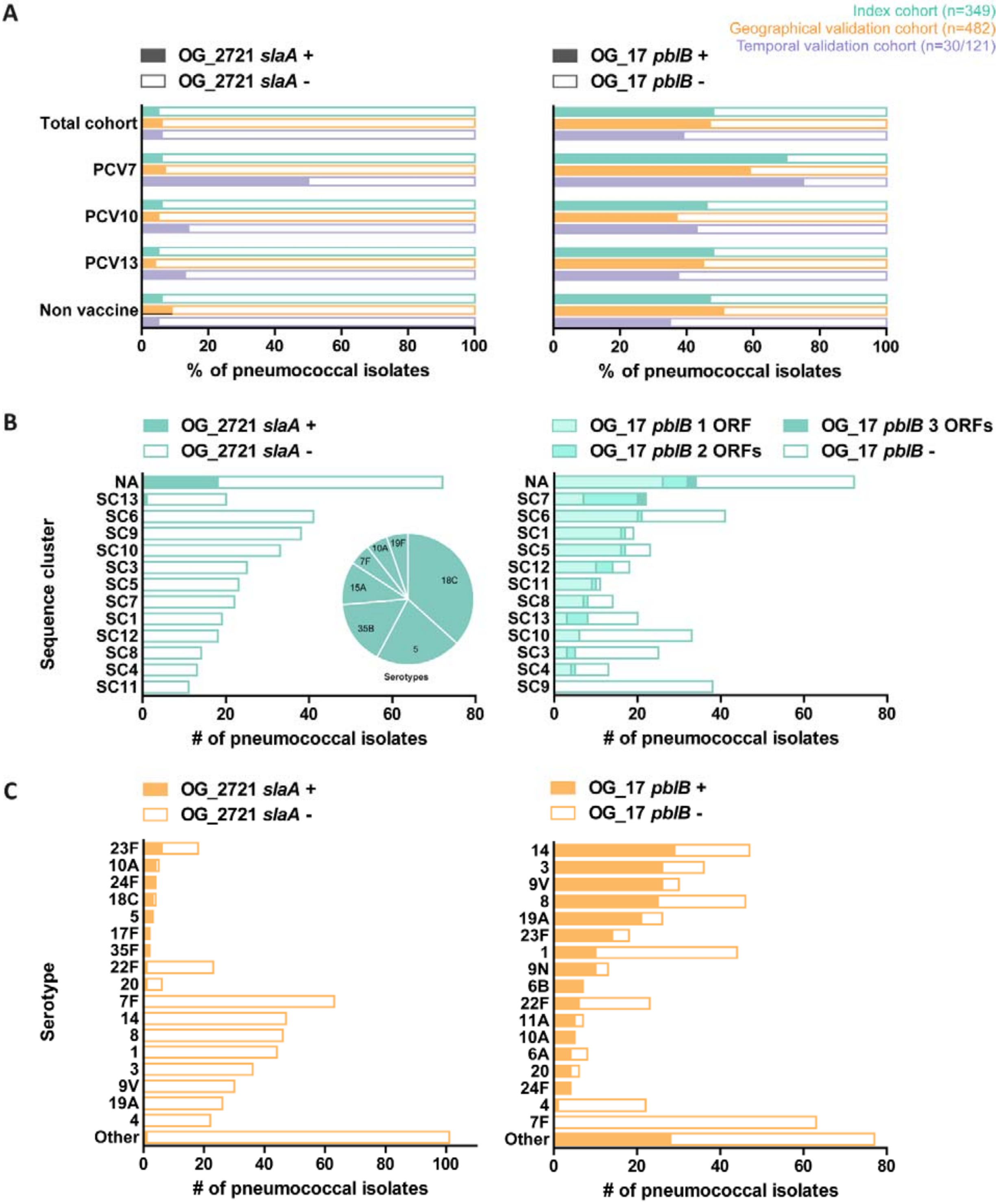
55

56 Supplementary figure 1 Sequence variation within confirmed orthologous genes



57

Supplementary figure 2 Distribution of 2 confirmed orthologous genes in the pneumococcal population


A
Index cohort (n=349)
Geographical validation cohort (n=482)
Temporal validation cohort (n=30/121)
OG_2721 slaA +
OG_2721 slaA -
OG_17 pblB +
OG_17 pblB -
Total cohort
PCV7
PCV10
PCV13
Non vaccine
% of pneumococcal isolates
B
OG_2721 slaA +
OG_2721 slaA -
OG_17 pblB 1 ORF
OG_17 pblB 2 ORFs
OG_17 pblB 3 ORFs
OG_17 pblB -
Sequence cluster
NA
SC13
SC6
SC9
SC10
SC3
SC5
SC7
SC1
SC12
SC8
SC4
SC11
19F
10A
7F
15A
18C
35B
Serotypes
NA
SC7
SC6
SC1
SC5
SC12
SC11
SC8
SC13
SC10
SC3
SC4
SC9
# of pneumococcal isolates
C
OG_2721 slaA +
OG_2721 slaA -
OG_17 pblB +
OG_17 pblB -
Serotype
23F
10A
24F
18C
17F
35F
22F
7F
9V
19A
Other
9V
19A
23F
9N
6B
22F
11A
10A
6A
24F
7F
Other
# of pneumococcal isolates