International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact

Rebecca A. Gladstone, Wellcome Sanger Institute
Stephanie W. Lo, Wellcome Sanger Institute
John A. Lees, New York University
Nicholas J. Croucher, Imperial College London
Andries J. van Tonder, Wellcome Sanger Institute
Jukka Corander, Wellcome Sanger Institute
Andrew J. Page, Wellcome Sanger Institute
Pekka Marttinen, Helsinki Institute for Information Technology
Leon J. Bentley, Wellcome Sanger Institute
Theresa J. Ochoa, Universidad Peruana Cayetano Heredia

Only first 10 authors above; see publication for full author list.

Journal Title: EBioMedicine
Volume: Volume 43
Publisher: Elsevier: Creative Commons | 2019-05-01, Pages 338-346
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1016/j.ebiom.2019.04.021
Permanent URL: https://pid.emory.edu/ark:/25593/trgb5

Final published version: http://dx.doi.org/10.1016/j.ebiom.2019.04.021

Copyright information:
© 2019 The Authors
This is an Open Access work distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/).

Accessed July 29, 2019 1:04 PM EDT
International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact

Rebecca A. Gladstone a,⁎, Stephanie W. Lo a,1, John A. Lees b, Nicholas J. Croucher c, Andries J. van Tonder a, Jukka Corander d, Andrew J. Page a, Pekka Marttinen e, Leon J. Bentley a, Theresa J. Ochoa f, Pak Leung Ho g, Mignon du Plessis h, Jennifer E. Cornick i, Brenda Kwambana-Adams j,k, Rachel Benisty l, Susan A. Nzenze m,n, Shabir A. Madhi m,n, Paulina A. Hawkins o, Dean B. Everett p, Martin Antonio k,q, Ron Dagan i, Keith P. Klugman o, Anne von Gottberg h, Lesley McGee f, Robert F. Breiman o,s, Stephen D. Bentley a,
The Global Pneumococcal Sequencing Consortium2

⁎ Corresponding author.
E-mail address: rg9@sanger.ac.uk (R.A. Gladstone).
1 These authors contributed equally
2 http://www.pneumogen.net/gps/

Abstract

Background: Pneumococcal conjugate vaccines have reduced the incidence of invasive pneumococcal disease, caused by vaccine serotypes, but non-vaccine-serotypes remain a concern. We used whole genome sequencing to study pneumococcal serotype, antibiotic resistance and invasiveness, in the context of genetic background.

Methods: Our dataset of 13,454 genomes, combined with four published genomic datasets, represented Africa (40%), Asia (25%), Europe (19%), North America (12%), and South America (5%). These 20,027 pneumococcal genomes were clustered into lineages using PopPUNK, and named Global Pneumococcal Sequence Clusters (GPSCs). From our dataset, we additionally derived serotype and sequence type, and predicted antibiotic sensitivity. We then measured invasiveness using odds ratios that relating prevalence in invasive pneumococcal disease to carriage.

Findings: The combined collections (n = 20,027) were clustered into 621 GPSCs. Thirty-five GPSCs observed in our dataset were represented by >100 isolates, and subsequently classed as dominant-GPSCs. In 22/35 (63%) of dominant-GPSCs both non-vaccine serotypes and vaccineserotypes were observed in the years up until, and including, the first year of pneumococcal conjugate vaccine introduction. Penicillin and multidrug resistance were higher (p < .05) in a subset dominant-GPSCs (14/35, 9/35 respectively), and resistance to an increasing number of antibiotic classes was associated with increased recombination (R² = 0.27 p < .0001). In 28/35 dominant-GPSCs, the country of isolation was a significant predictor (p < .05) of its antibiogram (mean misclassification error 0.28, SD ± 0.13).
We detected increased invasiveness of six genetic backgrounds, when compared to other genetic backgrounds expressing the same serotype. Up to 1.6-fold changes in invasiveness odds ratio were observed.

Interpretation: We define GPSCs that can be assigned to any pneumococcal genomic dataset, to aid international comparisons. Existing non-vaccine-serotypes in most GPSCs preclude the removal of these lineages by pneumococcal conjugate vaccines; leaving potential for serotype replacement. A subset of GPSCs have increased resistance, and/or serotype-independent invasiveness.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Background

Pneumococcal conjugate vaccines (PCVs) are highly effective in the prevention of invasive pneumococcal disease, caused by vaccine serotypes [1]. PCVs with seven, 10 or 13 conjugated serotypes are in use in ~150 countries [2]. Further conjugate vaccines are in development including PCV15 (Merck) [3] and a 20-valent formulation (Pfizer) [4], both in phase III clinical trials. Young children are the main carriers of Streptococcus pneumoniae, and immunisation of this age group protects them against invasive pneumococcal disease (IPD), caused by vaccine serotypes. Replacement of vaccine serotypes in carriage and disease by non-vaccine serotypes, termed serotype replacement, has offset some of the disease reductions in vaccinated and unvaccinated age groups [5,6]. The pneumococcal capsular polysaccharides are used to classify the pneumococcus into ~100 serotypes based on antibody binding to specific epitopes. Serotype is considered the primary pneumococcal virulence determinant [7].

Pneumococcal population structure can be characterised using multi-locus sequence typing (MLST), that determines the sequence variation in seven housekeeping genes. At least one MLST gene (ddl) has been linked to a known recombination hotspot in pneumococci [8]. MLST is limited in its ability to infer relationships between all strains [9], as shared ancestry can be masked by recombination and variation that has accumulated over longer timescales. Whole genome sequencing has increased resolution, allowing relationships between strains across the species to be established. Public health bodies are now taking steps to integrate pneumococcal whole genome sequencing into routine microbiology. Public Health England and the Centers for Disease Control and Prevention, have published methods using pneumococcal whole genome sequencing for determining serotype, and predicting antibiotic susceptibility [10,11].

Multiple studies have used pneumococcal genomics to investigate PCV impact [12–15]. They often cluster the population into groups using genomic variation that reflects a shared evolutionary history. To date, these genomic definitions of lineages have been dataset-specific, unlike MLST, hindering their use when comparing studies. The Global Pneumococcal Sequencing project (GPS, http://www.pneumogen.net/gps/), aims to provide an international understanding of pneumococcal population structure and PCV impact. It includes pneumococcal collections from invasive disease, and asymptomatic carriage. Multiple low- and middle-income countries are represented, and where possible, samples collected before and after PCV introductions. We aimed to use genome-wide variation to capture signals of shared descent, and define Global Pneumococcal Sequence Clusters (GPSCs). We used the GPSCs to provide further context on the distribution of serotypes, antibiotic resistance and invasiveness across pneumococcal lineages, which can aid assessments of PCV-impact [16,17].

2. Methods

2.1. Study design

We included 13,454 pneumococcal genomes available from the ongoing GPS project by June 2017. Investigators from each country provided epidemiological information including clinical manifestation, host age group, isolation year and sample source (Supplementary T1). IPD isolates were from a normally sterile site, while carriage isolates were from nasopharyngeal or nasal swabs. Participating laboratories performed antibiotic susceptibility testing where facilities allowed. We interpreted the susceptibility data as SIR (susceptible, intermediate, resistant) using Clinical Laboratory Standard Institute (CLSI) M100-ED28:2018. We applied the meningitis threshold for penicillin on all isolates, to allow assessment and comparison of penicillin resistance.
between GPSCs. Phenotypic antibiotic susceptibility data were available for <50% of isolates. This available data could be used to assess the sensitivity and specificity of genotypic prediction, and the validity of generating new, genome-derived, susceptibility data for several countries. The pre-PCV period was defined as the years when no conjugate vaccine was used and the year of first PCV introduction in each country.

2.2. Whole genome analysis

The following whole genome analyses are expanded in Supplementary Methods. Briefly, isolates were Illumina sequenced and raw data, assembled [18] and deposited in the European Nucleotide Archive (Supplementary T1). We derived MLST sequences types (STs) and serotype from the genome [19,20]. We further screened for the presence of known resistance conferring genes and mutations – for penicillin, tetracycline, erythromycin, chloramphenicol, co-trimoxazole – in the CDC pneumococcal typing pipeline database [11,21,22]. References to serotype and resistance throughout are from genomic inference. Multidrug resistance was defined as predicted resistance to ≥3 antibiotic classes.

To define GPSCs and improve global representation, the GPS dataset (n = 13,454) was supplemented with published datasets from the Netherlands (n = 2803), Thailand (n = 2663), USA (n = 616) and UK (n = 491) (Supplementary T2). We used PopPUNK to group isolates from this combined dataset (n = 20,027) into lineages, which clusters them using core and accessory distances (Fig. S1) [23]. We coined these lineages Global Pneumococcal Sequence Clusters (GPSCs), and created a reference database – available at https://www.pneumogen.net/gps/assigningGPSCs.html – that can be used with PopPUNK to assign the GPSCs to new data. HierBAPS was used as a clustering comparator [24]. It was run on a SNP alignment generated by mapping GPS isolates to ATCC 700669. Recombination was quantified among core genes using FastGEAR on a representative set of STs (Supplementary T4) [25,26]. Pairwise SNP distances were calculated for a core gene alignment of the GPS dataset generated via Roary [27], and for recombination-free alignments per dominant-GPSC, using the Pairsnp-r R package.

2.3. Statistical analysis

Estimates of the number of GPSCs in the true population were modelled using the R Vegan package [28], Simpson’s Diversity index 1-D (SDI) reports no diversity (zero) to unlimited diversity (one). We tested the predictive value of dominant-GPSCs, for antibiogram or serotype. We then tested the predictive value of country, for antibiogram or serotype, within each dominant-GPSC. We compared multinomial logit regressions to null models using a likelihood ratio test. Input data was restricted to the un-perturbed pre-PCV population. Corrections for multiple testing (n ≥10) used the Benjamini-Hochberg false discovery rate of 5%. Pneumococcal heritability (h^2) of invasiveness was calculated with a regression of all genomic variation using Pyseer (n = 1193 core genes, for recombination frequency (Supplementary T7). Distance matrices were calculated with a regression of all genomic variation using Pyseer [25,26]. The following whole genome analyses are expanded in Supplementary Methods. Brie

3. Findings

Our GPS dataset included 13,454 isolates representing 30 countries, and 5 continents: Africa (13 countries, 59% of collection), Asia (8, 18%), South America (2, 8%), Europe (4, 3%) and North America (3, 12%). GPS key countries (n > 1000) included South Africa (n = 4615, 63% IPD), The Gambia (n = 1637, 24% IPD), Malawi (n = 1304, 43% IPD), Israel (n = 1143, 100% IPD) and USA (n = 1584, 100% IPD). Fifty-four percent of the collection were isolated from IPD, 96% of the collection were isolated between 2000 and 2017 and 74% were from children aged ≤5 years old (Table 1, Fig. S2).

3.1. GPS diversity

Genome-wide variation in our dataset combined with published collections (total n = 20,027) clustered isolates into 621 GPSCs (Fig. S3). Our 621 GPSCs represent over 61% of the 1012.7 GPSCs (SE ± 76) estimated to be in the true population. However, most GPSCs (407 of 621, 66%) were rare lineages with <10 isolates, together representing 1043 of 20,027 (5%) of the combined collection. Within the GPS dataset, we observed 538 GPSCs. 35 GPSCs had >100 isolates in the GPS dataset and were classified as dominant-GPSCs. Together they represented 8356 of 13,454 (62%) of the GPS dataset, and 5979 of 6805 (89%) of the GPS dataset disease isolates (Fig. S4). Sampling multiple countries detected significantly more GPSCs (p < 0.001) than equivalent sampling from a single location (Fig. S5).

We defined MLST Clonal Complexes (CC) as STs with single locus variant (SLV) differences, within the GPS dataset. GPSCs often encompassed related CCs, with a mean number of 2.6 CCs per dominant-GPSC (SD ± 1.5, excluding singleton STs). GPSCs identify a shared history not captured by CC designations. CC217, CC615, CC3581 and 2 singleton STs were clustered into GPSC2, a grouping of CCs which is widely recognised as a clonal lineage that expresses serotype 1. CCs captured phylogenetic sub-structure well in dominant-GPSCs with more than 1 CC (n = 25, mean Consistency Index 0.9864 SD ± 0.04, mean Retention Index 0.9942 SD ± 0.02). Our clustering additionally revealed shared descent of CC53, CC1012, CC62 and CC100 within GPS3, that shared 0–5 MLST alleles. HierBAPS supported the clustering of 28/35 (80%) dominant-GPSCs, including GPSC3. The species-wide, core genome, pairwise SNP distances between GPSCs and within GPSCs, were predominantly non-overlapping (Fig. 1). The mean pairwise SNP distances from recombination free alignments, were broadly comparable between dominant-GPSCs, though eight GPSCs had SNPs distances >500 for a subset of their isolates (Fig. 2). HierBAPS supported half of those clusters with >500 SNP distances, but GPSC18, GPSC23, GPSC37 and GPSC41 were split into two subclusters. Conversely GPSC1, the clonal serotype one lineage GPSC2 and GPSC16 were split by HierBAPS into two sub-clusters even when the maximum SNP distances were >500.

MLST genes gki, gdh, recP, spi and ddl ranked in the top 6–22% of the 1193 core genes, for recombination frequency (Supplementary T7). Disruption of vertical inheritance may result in isolates that are missed by CC: within the dominant-GPSCs 370 of 8356 isolates (4.4%), belonging to 165 STs, were not assigned to a CC. Conversely recombination can result in convergent MLST profiles in disparate isolates, and CC designation using large collections are more vulnerable to spurious connections. Sixteen CCs spanned multiple related GPSCs and 24 isolates were assigned to 6 CCs highly discordant with their core genome phylogenetic relationship (Fig. S6).

Geographical diversity varied considerably per GPSC (Fig. 3A), though all dominant-2-GPSCs were observed in more than one country.
Eight of 35 (23%) had high geographical diversity (SDI>0.70) representing even distribution across 5 continents (Supplementary T8). Conversely, seven of 35 (20%) dominant-GPSCs were observed only in Africa. Genomic inference of serotype was reliable with 97% (95% CI 97.15–97.75%) serotype co-occurrence with available phenotypic data (n = 10,466, Supplementary T9). Of 35 dominant-GPSCs, 30 (86%) were observed with more than one serotype. Dominant-GPSC was a significant predictor of serotype with a mean misclassification error of 0.18, although serotype diversity within dominant-GPSCs varied considerably (p < 0.0001, Fig. 3B). Country was a significant predictor of serotype within 18 of 35 dominant-GPSCs (51%) with a mean misclassification error of 0.19 (SD ± 0.14, p < 0.05). For example, in GPSC3, serotype 8 was expressed in 113 of 148 (76%) IPD isolates from South Africa and all were CC53, conversely, serotype 33F was expressed in 49 of 87 IPD isolates from the USA and 44 of 49 (90%) from South Africa and all were CC53, conversely, serotype 33F was expressed in 49 of 87 IPD isolates from the USA and 44 of 49 (90%) from South Africa and all were CC53, conversely, serotype 33F was expressed in 49 of 87 IPD isolates from the USA and 44 of 49 (90%) from South Africa and all were CC53.

Two of the 35 dominant-GPSC (6%, GPSC14, GPSC37) included isolates exclusively expressing PCV7-serotypes pre-PCV and therefore were completely covered by PCV7. However, GPSC14 and 37 accounted for only 169 of 4221 (4%) of pre-PCV disease isolates. The number of dominant-GPSCs expressing only VTs pre-PCV, increased to 6 GPSCs for PCV10 (885 of 4221 [21%] of pre-PCV IPD isolates), and 11 GPSCs for PCV13 (1234 of 4221 [29% pre-PCV IPD]; Fig. 4A). The experimental 15-valent vaccine offered no additional dominant-GPSCs expressing only VTs. The putative 20-valent formulation meant that a further four dominant-GPSCs were expressing only VTs pre-PCV (1592 of 37% of pre-PCV IPD isolates).

The positive predictive values for genetic determinants of resistance to penicillin, tetracycline, erythromycin, chloramphenicol, and co-trimoxazole were all >90% (95% CI 90–98.7%, Supplementary T10). Resistance to at least one antibiotic class was predicted for 8241 of 13,454 isolates in the GPS dataset (61%; Fig. S7). The percentage of isolates predicted to be resistant per class was not uniform across dominant-GPSCs (Fig. 5). The predicted resistance profile of an isolate could be predicted by which dominant-GPSC it belonged to half of the time, (p < 0.0001, misclassification error 0.49). Generally, higher recombination ratios (r/theta, r/m), were associated with a higher mean number of classes of predicted antibiotic resistance (r/theta R² = 0.27, p < 0.0001, r/m R² = 0.22, p < 0.0001, Supplementary T8 and T11). GPSC1 had an average above r/m (8.3) and r/theta (0.14) and a predominant predicted MDR antibiogram of penicillin, co-trimoxazole, erythromycin and tetracycline resistance, but susceptibility to chloramphenicol, occurring in 388 of 504 isolates (77%). Although this MDR antibiogram was the most common pattern in 17 of 19 countries represented in GPSC1, country was a predictor of the antibiogram for 28 of the 35 dominant-GPSCs (80%; p < 0.05, mean misclassification error 0.28, SD ± 0.13). Predicted penicillin resistant isolates accounted for a higher proportion of isolates within 14/35 dominant-GPSCs than expected given proportion of predicted penicillin resistance in the rest of the GPS dataset (p < 0.05, 63–100%, Supplementary T12). Predicted multidrug resistant isolates, accounted for higher proportion of isolates within 9/35 dominant-GPSCs than expected (p < 0.05, 45–77%). Eight of these were GPSCs also found to have a higher proportion of isolates penicillin resistant isolates (Supplementary T13). Prior to PCV introductions, penicillin resistance was predicted to occurred in 2133 of 4975 (43%) of the isolates expressing PCV13 VTs and in only 256 of 2135 (12%) of the NVT expressing strains (p < 0.0001).

In 9 of the 22 (40%) GPSCs expressing both VT and NVT, the NVT component had a significantly lower proportion of predicted resistant isolates than their VT counterparts (Supplementary T14). Seven intermediate-GPSCs expressing only NVTs had >90% of isolates predicted resistant to penicillin (GPSC55, 89 of 90, [99%]; GPSC59, 37/37 of [100%]; GPSC81, 35 of 38 [92%]; GPSC132, 17 of 17, 100%); GPSC136, 19 of 21 [90%]; GPSC168, 15 of 15 [100%]; GPSC200, 11 of 11 [100%]).

In the South African heritability dataset, serotype explained a third of the strain variation in clinical manifestation (carriage or disease, pseudo-R² 0.32). Total pneumococcal genetic variation (including the cps locus which encodes the CPS) was a better explanation (R² 0.57), explaining over half of the variation in clinical manifestation, leaving some invasiveness explained by genes outside the cps locus.

The 95% CI, for invasiveness ORs did not overlap between at least one pair of genotypes (GPS n = 96, ST n = 112) within serotypes 6A, 14, 16F, 19F, 23B and 23F (Table 2, Fig. S8). Only within GPSC14 was a genotype with increased invasiveness found to be significant at the ST level but not the GPS level, as ST6729 and ST2059, both found within GPSC14, significantly differed in invasiveness (Table 2, Supplementary T15 and T16). We detected significant heterogeneity in invasiveness for serotype 38-GPSC38/ST393 (Q = 3.877 df 1, p < .05), between South Africa (OR 0.67, 95% CI 0.24–1.88) and the USA (OR 6.83, 95% CI 0.86–54.20) in the meta-analysis estimate of the OR. Despite the small sample numbers, the two countries significantly differed in the proportion of serotype 38-GPSC38 from IPD (p = .008). A conservative comparison using the upper CI of the least invasive genotype and the lower CI of the most invasive genotype had 1.05 to 1.6-fold changes in OR (Table 2). The influence of genotype on invasiveness can be of a similar magnitude to some serotypes. For example a 1.6-fold change was observed between the upper CI of the less invasive serotype 35A (0.12 [0.016–0.938], p = .043) and the lower CI of the more invasive serotype 18C (3.237[1.514–6.921], p = .0024) determined in this dataset (Supplementary T17).
4. Discussion

We present the distribution of key themes in pneumococcal disease control, such as serotype, antibiotic resistance and invasiveness, in a large international collection. We used genome-wide variation to define Global Pneumococcal Sequence Clusters (GPSCs), to produce a dataset independent genomic definition of lineages. Increasing knowledge of the spread of traits across the pneumococcal population and geographical regions, gives greater context for assessing the impact of PCV introduction.

At an international level, we have shown that pneumococcal non-vaccine serotypes exist alongside vaccine serotypes, within dominant GPSCs that account for the majority of the pneumococcal population. The existence of non-vaccine serotype variants negates reliance on contemporaneous capsule switch events for “vaccine escape” of a GPSC. Given that the pneumococcus has multiple lineages that are globally disseminated there is potential for non-vaccine types established in one geographical region to spread globally, or be present but undetected in other countries [32]. Indeed, previous carriage studies have observed the importation of lineages not previously observed in that...
We showed that antibiotic resistance was enriched in a subset of GPSCs, many of which were dominant and globally disseminated. We observed that both GPSC, and country within GPSCs were significant predictors of the antibiotic resistance pattern of an isolate. GPSCs with an increased propensity for resistance, whether associated with increased capacity for recombination, duration of carriage [34], can spread to other locations. Loss of resistance, in the absence of selection, has been reported for lineages in countries with lower antibiotic prescription rates; multiple independent losses of resistance to chloramphenicol, tetracycline and erythromycin were observed for Pneumococcal Molecular Epidemiology Network (PMEN)2 in Iceland [35]. However, over a decade after a reduction in antibiotic consumption, the majority of Icelandic PMEN2 remained resistant, albeit at a reduced prevalence, and as such lineages remain a risk to high usage settings. Antibiotic resistance is lower in non-vaccine serotypes, but this prevalence varies substantially by GPSC. Some notable GPSCs expressing only non-vaccine serotypes do have high levels of penicillin resistance, and within GPSCs that express both non-vaccine and vaccine serotypes, the non-vaccine serotypes occasionally have similar or higher resistance profiles to their vaccine serotype counterparts.

Preservation of gene frequencies in the population through negative frequency-dependent selection has been shown to exist in pneumococci, and can be used to predict serotype replacement in carriage [16,36]. This suggests that the gene content of a GPSCs influences whether it will undergo replacement or expansion after vaccine perturbation. Only genomic data combined with a robust clustering method...
has the power to model such complex dynamics. Non-vaccine and vaccine serotypes within the same GPSC, will share similar gene complements and ecological phenotypes, including resistance. Non-vaccine variants may therefore have increased potential to replace their vaccine-type counterparts compared to other GPSCs expressing non-vaccine serotypes [13,37]. The GPSCs involved in any replacement will determine the extent to which reductions, not only in total disease, but in antibiotic resistance, could be partially offset by non-vaccine serotypes. This has been observed with pneumococcal serotype replacement after routine use of the 7-valent conjugate vaccine by multidrug resistant 19A in the US within GPSC1 (CC320), which slowed reductions in total disease, antibiotic resistance, and subsequently, the cost effectiveness of PCV7 [38].

Serotype is a potent virulence determinant, however other virulence factors exist in the genome outside of the cps locus. Genotypes have previously been implicated in invasiveness in a number of small studies using MLST/pulsed-field gel electrophoresis definitions, some of which are complicated by age-related differences in invasiveness [30,39–41]. With a substantial collection of pneumococcal genomes, we have used heritability analysis and difference in invasiveness ORs to demonstrate that genome variation beyond serotype contributes to invasiveness in children <7 years old. Measurable differences in invasiveness ORs between individual GPSCs and STs were comparable to a change in serotype by genotype (GPSC or ST) significantly differed within a serotype.

<table>
<thead>
<tr>
<th>Serotype</th>
<th>Least invasive (predominant CC)</th>
<th>OR [95% CI]</th>
<th>Most invasive (predominant CC)</th>
<th>OR [95% CI]</th>
<th>Conservative OR fold change</th>
<th>Fisher’s p-value (country)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6A</td>
<td>GPSC5 (CC172)</td>
<td>0.34 [0.12–1.01]</td>
<td>GPSC41 (CC1094)</td>
<td>2.96 [1.61–5.45]</td>
<td>1.6</td>
<td>0.0004 (ZA)</td>
</tr>
<tr>
<td>14</td>
<td>GPSC31 (CC31)</td>
<td>0.62 [0.22–1.73]</td>
<td>GPSC38 (CC15)</td>
<td>12.45 [2.82–54.98]</td>
<td>1.6</td>
<td>0.0005 (ZA)</td>
</tr>
<tr>
<td>16F</td>
<td>GPSC33 (CC4088)</td>
<td>0.14 [0.06–0.36]</td>
<td>GPSC46 (CC30)</td>
<td>2.62 [0.44–15.73]</td>
<td>1.2</td>
<td>0.0099 (ZA)</td>
</tr>
<tr>
<td>19F</td>
<td>GPSC21 (CC347)</td>
<td>0.32 [0.19–0.54]</td>
<td>GPSC1 (CC230)</td>
<td>1.49 [0.78–2.88]</td>
<td>1.4</td>
<td>0.0006 (ZA)</td>
</tr>
<tr>
<td>23B</td>
<td>GPSC7 (CC439)</td>
<td>0.14 [0.05–0.42]</td>
<td>GPSC5 (CC172)</td>
<td>3.81 [0.44–32.79]</td>
<td>1.05</td>
<td>0.005 (USA)</td>
</tr>
<tr>
<td>23F</td>
<td>GPSC14 ST26279 (CC6279)</td>
<td>0.81 [0.43–1.54]</td>
<td>GPSC14 ST2059 (CC6279)</td>
<td>5.31 [1.71–16.54]</td>
<td>1.1</td>
<td>0.004 (ZA)</td>
</tr>
</tbody>
</table>

GPSC (Global Pneumococcal Sequence Cluster) OR (Odds ratio), ST (Sequence Type), CC (Clonal Complex), ZA (South Africa), CI (confidence interval), The conservative fold change in OR was calculated by dividing the lower CI of the most invasive genotype by the upper CI of the least invasive genotype, within each serotype.
a reproducible clustering nomenclature. Our definition of the pneumococcal lineages (GPSCs) on an internationally sampled population, can be used to assign GPSCs to any collection of pneumococcal genomes using our database of GPSC reference genomes and PopPUNK [23]. The GPSC database can be updated when novel GPSCs are assigned in future collections, enabling stable international comparisons of pneumococcal population structure [15]. The GPSCs are broadly back-compatible with MLST as the vast majority of STs were found exclusively within a GPSC. To that end, we provide a ST-GPSC conversion table with noted exceptions, to facilitate cross referencing of non-genomic datasets (Supplementary T18). We have used these GPSC designations, genome-derived serotype and antibiotic resistance to facilitate an in depth assessment of the lineages causing invasive disease in young children in the post-PCV13 era [17], and to explore the mechanisms driving the progression of serotype replacement.

Understanding the underlying genetic variation and characteristics of GPSCs that influence resistance, invasiveness and pneumococcal population dynamics in a global context is highly informative. Such information can be used for modelling vaccine replacement, predicting vaccine impact and rational vaccine design. Our high-resolution genomic approach for defining pneumococcal lineages across different collections, in a manner that reflects pneumococcal biology, increases the evidence required to build a global strategy for continued control of pneumococcal disease.

Acknowledgements

We would like to thank all members of the GPS consortium for their collaborative spirit and determination during the monumental task of sampling, extracting and sequencing this dataset, and all contributions to experimental design and input into this manuscript. We also would like to thank members of teams 284 and 81 at the Wellcome Sanger Institute (WSI) for their advice and critique and the pathogen informatics team at the WSI for the pipelines and expertise that made genomic analysis at this scale possible. We would like to thank Diederik van de Beek and Phillip Kremer for their critical review and clinical perspective on the manuscript.

Funding

This study was co-funded by the Bill and Melinda Gates Foundation, the Wellcome Sanger Institute, and the US Centers for Disease Control and Prevention.

Role of funding sources

This study was co-funded by the Bill and Melinda Gates Foundation (grant code OPP1034556), the Wellcome Sanger Institute (core Wellcome grants 098051 and 206194) and the US Centers for Disease Control and Prevention. The funding sources had no role in isolate selection in each country participating in GPS. Appropriate approvals for use of isolates was obtained from each institution contributing isolates. No tissue material or other biological material was obtained from humans. All information regarding these isolates was anonymised.

Declaration of interests

Dr. Gladstone reports PhD studentship from Pfizer, outside the submitted work; Dr. Lees reports grants from Pfizer, outside the submitted work; Dr. Madhi reports grants from BMGF, during the conduct of the study; grants and personal fees from BMGF, grants from Pfizer, grants from GSK, grants from Sanofi, grants from BIOVAC, outside the submitted work; Dr. Dagan reports grants and personal fees from Pfizer, during the conduct of the study; grants and personal fees from MSD, personal fees from MeMed, outside the submitted work; Dr. von Gottberg reports grants and other from Pfizer, during the conduct of the study; grants and other from Sanofi, outside the submitted work; Dr. Bentley reports personal fees from Pfizer, personal fees from Merck, outside the submitted work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2019.04.021.

References


