Comparing GPSCs and clonal complex clusters for *Streptococcus pneumoniae*

Narender Kumar[1], Stephanie W. Lo[1], Kate Mellor[1], John Lees[2], Paulina A. Hawkins[3], Lesley McGee[4], Nicholas J. Chroucher[5], Stephen D. Bentley[1]

[1]Parasites and Microbes Programme, The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[2]EMBL-EBI, Wellcome Genome Campus, Hinxton, UK ,
[3]Rollins School Public Health, Emory University, Atlanta, GA, USA,
[4]Emory Global Health Institute, Emory University, Atlanta, GA, USA,
[5]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, UK

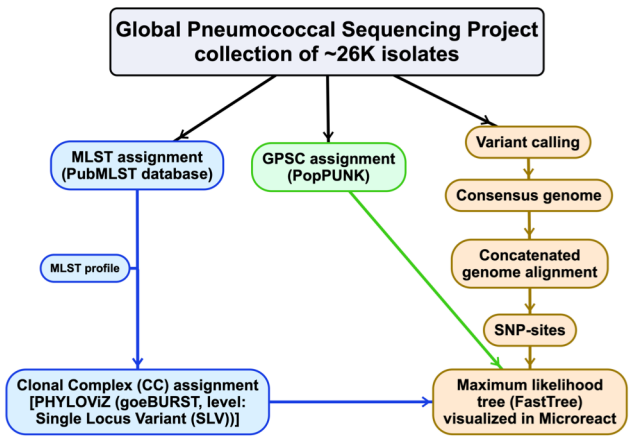
Introduction

Lineages in *Streptococcus pneumoniae* are defined based on clustering of MLSTs into clonal complexes (CCs) dependent on 7 housekeeping genes. This method lacks resolution with some of the genes subject to recombination and the results not easily comparable between studies. Also, the methodology is not very user-friendly and clustering is study specific. Recently, PopPUNK, a new method has been described that sketches the entire genome into k-mers and identifies Global Pneumococcal Sequence Clusters (GPSCs). Here, using a large global collection of pneumococcal isolates, we compare the clustering obtained by the two methods.

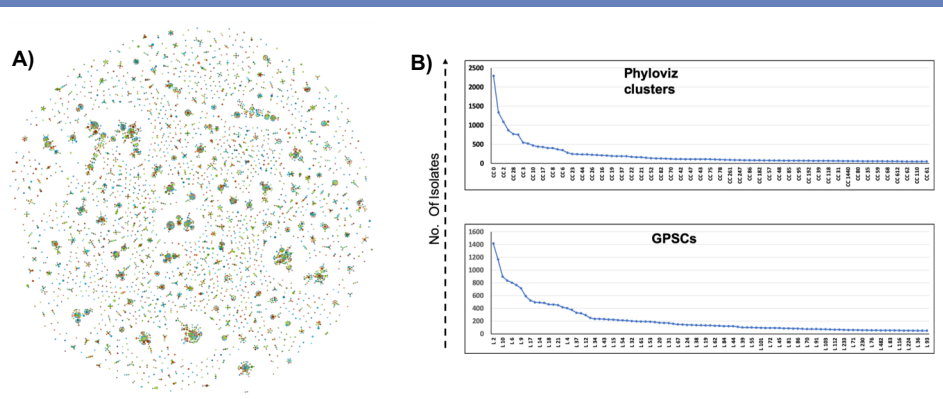
Aim

Comparing the genome based GPSCs against the MLST based CC assignment

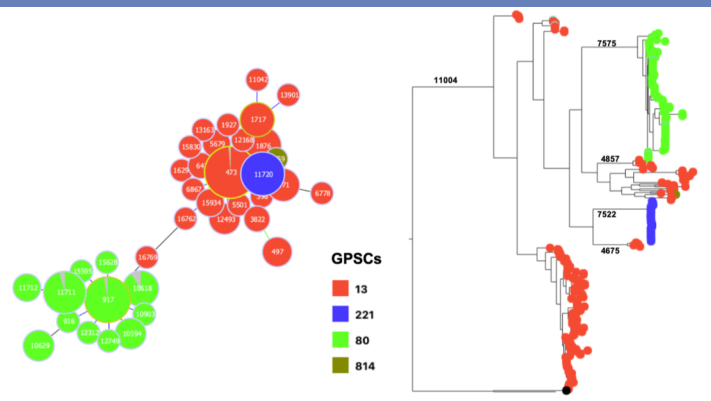
Methods



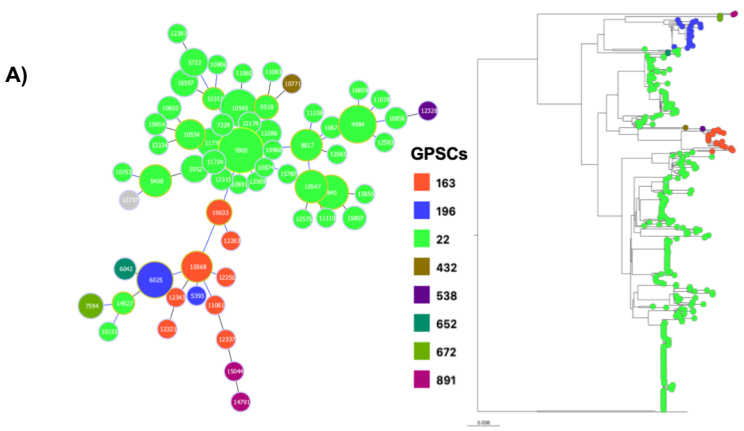
Results

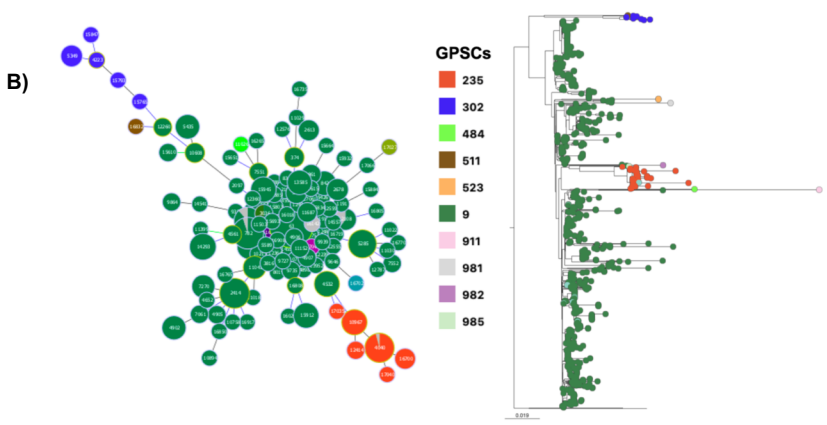


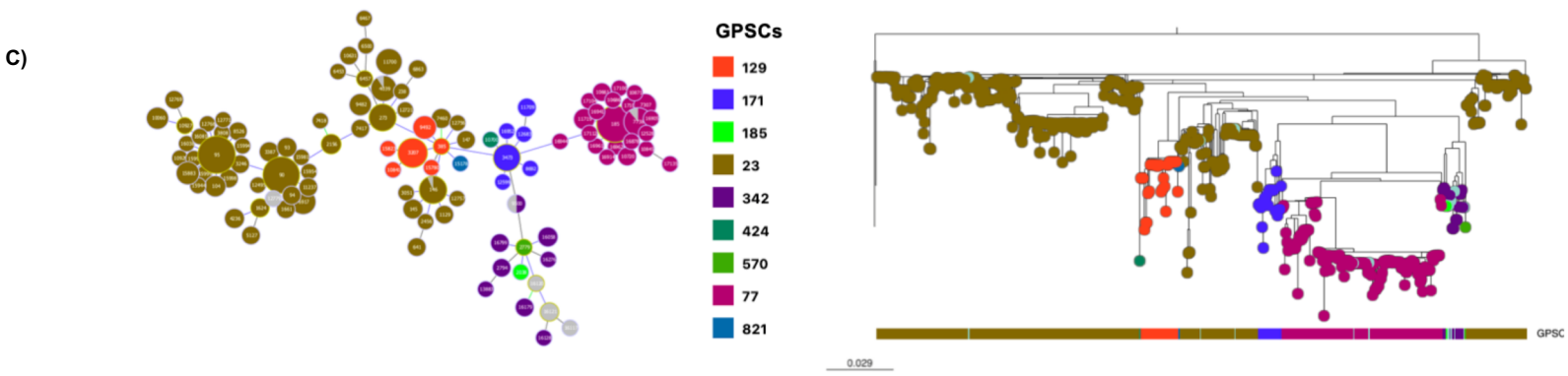
1. High diversity in GPS collection: The 26K isolates were clustered into 1657 clusters of which 810 (49%) were represented a single isolate. Using PopPUNK we identified 830 GPSCs of which 272 (32%) had a single isolate. Of the 1657 clusters there were 1608 (97%) with isolated belonging to a single GPSCs, showing a high concordance between the methods. 49 clusters had more than 1 GPSCs identified and were investigated further.



2. Cluster 14 (CC473) with 4 GPSCs: The figure above show that different GPSCs (coloured circles) are placed in distinct clusters within the phylogenetic tree. This is further supported by the single long branches reflects high genetic distance (number of SNPs) between the GPSCs.







3. A) Cluster 5 with 8 GPSCs; B) cluster 1 with 10 GPSCs; C) cluster 3 with 9 GPSCs: In the above three figures it can be again observed that the GPSCs form separate clusters separated by single long branches supportive of the higher genetic distances from other isolates. Therefore, GPSCs assignments provide a higher resolution when compared to CCs.

Conclusion

The sequence-based method for defining GPSC lineages presents a robust platform to study population structure of *S. pneumoniae* and enables comparisons of populations across regions to track the global spread of this pathogen.

Acknowledgements

This study was co-funded by the Bill & 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. We would like to thank all members of the Global Pneumococcal Sequencing Consortium for their collaborative spirit and determination, for the monumental task of sampling and extracting data. We also acknowledge the Wellcome Sanger Institute Sequencing Facility and Pathogen Informatics team for their technical supports.