Title: 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 Bentley[1]

[1]Parasites and Microbes Programme, The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge,UK
[2]EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, 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

Background: Lineages in *Streptococcus pneumoniae* are defined based on clustering of STs termed 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. Recently, PopPUNK, an alternative method that sketches the entire genome into k-mers and identifies Global Pneumococcal Sequence Clusters (GPSCs) in *S. pneumoniae,* has been described. Here, using a large collection of global isolates, we evaluate the concordance between the methods.

Methods: We used PopPUNK to designate each of the 26, 577 GPS collection isolates to a GPSC. Clonal complexes (CC) were identified using goeBURST clustering (Phylovizv2.0) based on the MLST profiles of the isolates in the collection. The CCs identified were then compared to GPSC assignments to determine the concordance between the two clustering methods.

Results: The 26,577 isolates in the GPS collection clustered into 830 GPSCs with 57 GPSCs represented by >100 isolates comprising 70% (18,701/26,577) of the collection. The goeBURST clustering identified 1,667 CCs, and 47 CCs were represented by >100 isolates and comprised 61% (16,297/26,577) of the collection. The observed concordance between the two clustering methods was 97% (1,612/1,667); in fact, there were only 48 CCs with isolates belonging to more than one GPSC.

Conclusions: 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.