**DEFINING THE LINEAGES OF THE PNEUMOCOCCUS TO UNDERSTAND DISEASE AND VACCINE IMPACT: A MULTI-CONTINENT GENOME ANALYSIS**


1 Infection genomics, Wellcome Sanger Institute, UK. 2 Faculty of Medicine, School of Public Health, Imperial College London, UK. 3 Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, Johannesburg, South Africa. 4 Malawi-Liverpool-Wellcome Trust, Malawi. 5 Vaccines and Immunity Theme, MRC Unit, The Gambia. 6 Medical Research Council: Respiratory and Meningoccal Pathogens Research Unit, University of the Witwatersrand, South Africa. 7 Rollins School Public Health, Emory University, USA. 8 Queens Research Institute, University of Edinburgh, UK. 9 Centers for Disease Control and Prevention, Atlanta, USA. 10 The Emory Global Health Institute, Atlanta, USA. 11 http://www.pneumogen.net/gps/

**Background**
Understanding pneumococcal conjugate vaccine (PCV) impact, particularly non vaccine serotype (NVT) replacement, requires the context of the genetic backgrounds in which serotypes are expressed. Genotyping methods, such as MLST, are incomplete representations of the population structure. We provide a genome based definition of pneumococcal lineages to understand pneumococcal disease and vaccination.

**Methods**
12,239 genomes from disease (60%) and carriage (40%) representing 29 countries (Fig 1) were used to delineate the species into 356 Global Pneumococcal Sequence Clusters (GPSCs), clustering used genome-wide kmer sequence similarity with mash [Croucher et al BioRxv]. Fisher’s exact with FDR adjustment for multiple testing was used to identify GPSCs with a significantly higher proportion of resistant isolates.

**Comparison with MLST**
Up to 111 STs could be observed within a single GPSC that represents shared evolutionary history. Whilst only one occurrence of ST convergence in unrelated isolates was observed for ST6011, incorrect inference of relatedness was a major issue at the level of clonal complex (CC). For 14/374 CCs, isolates were found spread across the breadth of the phylogenetic tree (Fig 3). MLST genes were found to be in the top 16% of genes by recombination frequency.

**Antibiotic resistance**
Penicillin resistance had a significantly higher prevalence (p<0.0001) in 19 GPSCs, 15 of which were observed to have high prevalence (>200 isolates). Seven of those 15 GPSCs were also enriched (p<0.01) for multi-drug resistance (<3 classes). 14/15 of these GPSCs express VTs and NVTs, in 2 and 3 GPSCs NVTs have an equal or higher prevalence of penicillin resistance or MDR respectively.

**Vaccine escape**
Only a fifth of the dominant GPSCs exclusively express vaccine types (VTs), these GPSCs will likely be eradicated by PCV use, but account for less than a quarter of the disease cases. The majority of GPSCs express multiple serotypes. Importantly NVT variants exist alongside VTs in most GPSCs.

**Conclusions**
The substantial diversity of the pneumococcus provides the opportunity to rapidly adapt to selection pressures. The existence of NVTs in the majority of GPSCs negates the need for contemporaneous capsule switch for vaccine escape. Resistance is a key shared feature of prevalent GPSCs, it may contribute to the success of common GPSCs and influence the progression of replacement.