Mechanisms of virulence regulation and global distribution of vaccine candidate antigens in the high virulent *Streptococcus pneumoniae* strains

I. Tsvetkova[1], D. Likholetova[1,2], V. Gostev[1,3], S. Belanov[4], E. Nikitina[1], O. Kalinogorskaya[1], M. Volkova[1], A. Mokhov[5], X. Ivanova[1], E. Kalisnikova[1], A. Volodina[6], S. Sidorenko[1,3]

[1]Pediatric Research and Clinical Center for Infectious Diseases, Department of medical microbiology and molecular epidemiology, Saint-Petersburg, Russian Federation,
[2]Saint-Petersburg State Univesity, Biology, Saint-Petersburg, Russian Federation,
[3]North-Western State Medical University, Department of medical microbiology, Saint-Petersburg, Russian Federation,
[4]University of Helsinki, Institute of Biotechnology, DNA Sequencing and Genomics, Helsinki, Finland,
[5] Public Funded Health Facility Mariinskaya Municipal Hospital, Saint-Petersburg, Russian Federation,
[6]Saint-Petersburg scientific research institute of vaccines and serums, Saint-Petersburg, Russian Federation.

Background

Metabolic flexibility is a prerequisite for successful transition of *Streptococcus pneumoniae* (Spn) from colonizing to invasive state. The aim of this study was to investigate the common metabolic trends of invasive strains and propose vaccine candidate antigens.

Methods

The dataset included the 1058 PubMLST Spn (Russian and reference strains). MLST concatenates were used to build maximum likelihood phylogeny. Sequence clusters (SCs) were identified with RhierBAPS. Core genome phylogeny after recombinations filter was performed for subsampling of 495 isolates with Gubbins and RaxML. Core gene variant matrix was derived with GenomeComparator. Initial analysis of the data structure was performed using multiple correspondence analysis (MCA). More than 200 genes were identified that explain the variability of more than 20% of the data (in 20 dimensions). These variables were used for analysis as a secondary matrix. The machine learning algorithms, Random Forest (R package bigrf) and XGBoost in ensemble with Deep Learning, were used to find associations of gene variants with genotypes, and invasiveness. STRING database was used for the acquired genes annotation.

Results

Different algorithms made it possible to identify the following groups in the Spn population: Cluster I and Cluster II (according to the core-genome data, Fig.1), three groups A/B1/B2 and 11 SCs (according to the phylogeny on the concatenates of the six MLST genes without *ddl*, Fig.2.a-c). Distribution of the isolates among SCs_MLST did not correlate well with the phylogenetic tree based on the core genome (Fig.1). However, the SCs and AB1B2 groups were associated with different dominant serotypes and invasiveness (Fig.3a,b). In addition, the MCA-clasterization based on the gene variants of the total reference genome (ATCC700669) described the groups A/B1/B2, SCs_MLST, serotypes and invasive strains (Fig.6a-d), herewith the SCs_MLST were not always genetically homogeneous (Fig.6a) and the A/B1/B2 groups were genetically heterogeneous (Fig.6b). With machine-learning algorithms, the top-lists of the genes were identified for different groups of the isolates. 181 genes were significant for the formation of the A/B1/B2 groups (Z-score values were comparable with Z-score for A/B1/B2). These included the components of the various metabolic pathways (fatty acid biosynthesis, galactose metabolism and others) and virulence factors (ABC-transporters, PTS-system components, metal-transporting P-type ATPase, FpbS, PotD and others). For invasive isolates we found a strong association of the StrH, NanA, Phts, NanB, OppC, PcpA, Wzd (capsule biosynthesis) and several others with clades and invasiveness.

Fig.1. Core genome phylogeny

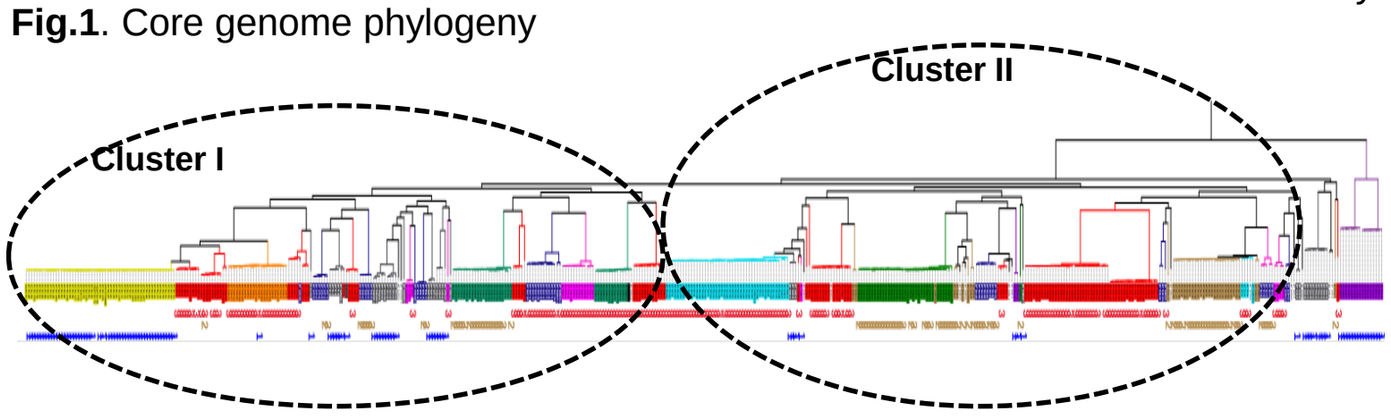
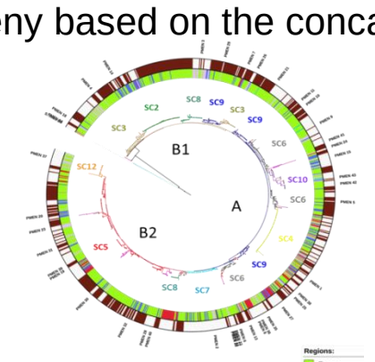
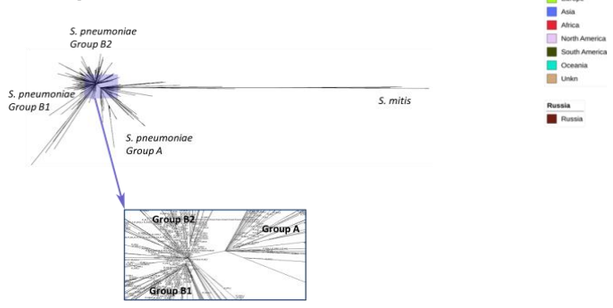


Fig.2. Phylogeny based on the concatenates of the six housekeeping genes

a. RaxML



b. SplitsTree



c. *Spn* global population (14663 STs from PubMLST) (RaxML)

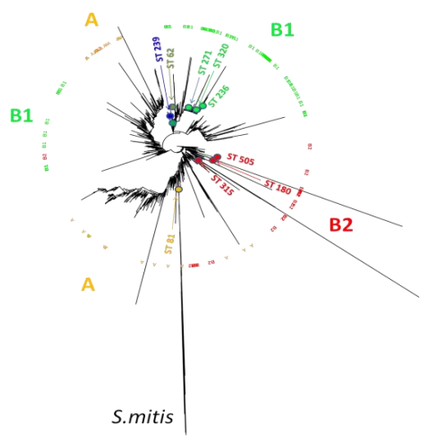


Fig.3a

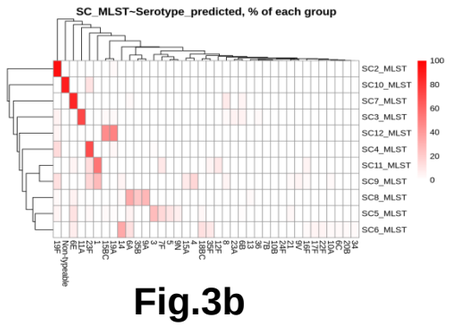


Fig.3b

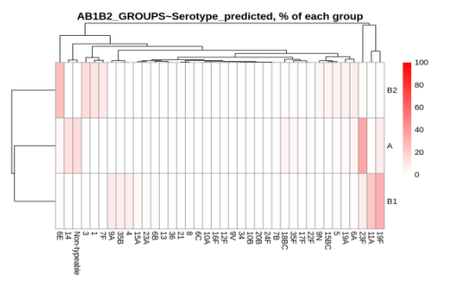


Fig.6a

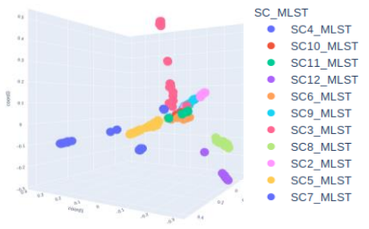


Fig.6b

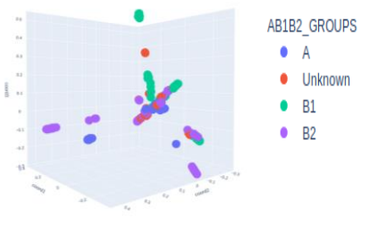


Fig.6c
AB1B2 – whole dataset

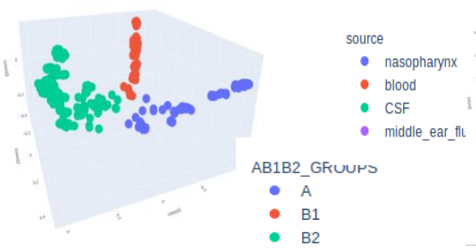
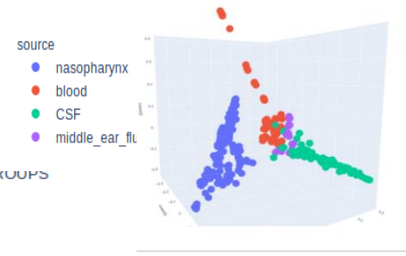


Fig.6d
Source – 228 variables



Conclusion

The population division into the groups A/B1/B2 may arise due to the virulent potential. We achieved the list of vaccine candidates - surface and membrane proteins: StrH, Phts, NanB, OppC, Wzd and several others.