WHITE PAPER

Streptococcus agalactiae – cgMLST schema generation and validation

Fredrik Dyrkell[1], Dimitrios Arnellos[1]

[1]1928 Diagnostics AB – Mölndal (Sweden)

Background

Group B Streptococcus (GBS) can cause serious illness in people, particular in newborns. Therefore it is of value to monitor and perform surveillance at hospitals for HAI and prevent outbreaks.

Using 1928 Diagnostics (1928) core-genome MLST schema generator, we created a putative cgMLST schema for *Streptococcus agalactiae* and evaluated its performance in terms of ability to detect core genes (i.e. the definition of a core gene is that it is present in the majority of strains – in our case > 95% of reference genomes) as well as having high resolution to be able to detect possible outbreaks and related samples.

Method

All the complete genomes of species *Streptococcus agalactiae* available from NCBI RefSeq were retrieved on the 21th of April 2023 (n=102) and used as reference genomes. Genes present in > 95% of reference genomes were retained as core genes. *Streptococcus agalactiae* isolate SA111 (NZ_LT545678.1) was used as the seed reference, providing the base of the core genome schema. The resulting schema yielded 1258 core genes representing 60.0% (1258 out of total 2095) of the coding genes from the seed reference.

The core genome schema performance was evaluated using four different, publicly available datasets.

BIOPROJECT	# SAMPLES	ARTICLE
PRJNA345233	45	–
PRJEB26578+PRJEB34494	15	[1]
PRJEB18093	801	[1]
	861	

About 1928 Diagnostics AB

The 1928 Bioinformatics Platform supports microbiologists and infection control professionals who want to take full advantage of DNA-sequencing for transmission analysis and taxonomic classification of bacteria and fungi.

1928 Bioinformatics Platform
– where complex data turns into actionable results

Results

Out of the 861 samples analysed, nine failed the quality control, and were removed from further analysis.

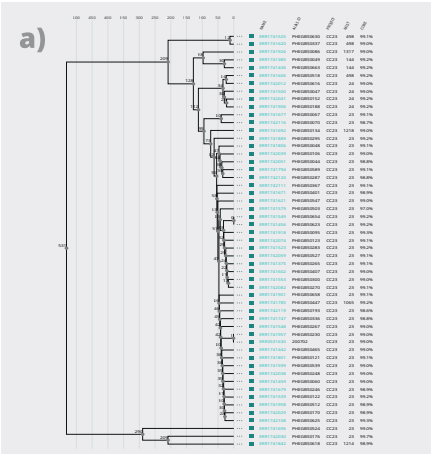
- Three failed due to sequence depth under the requirements of > 30x
- Six samples failed due to detection of multiple core gene alleles indicating strain intra species contamination

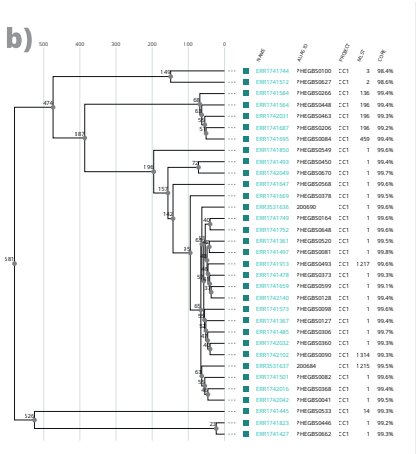
MLST results showed a diversity of 113 different sequence types (not counting novel profiles, of which there were 16).

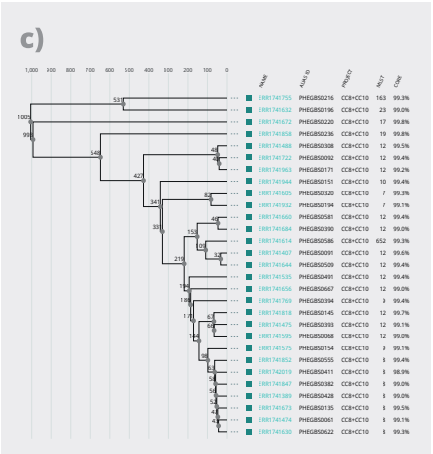
The average and median core genes found in the 852 samples were 99.2% and 99.36% respectively. Only two samples had < 95% of core genes found. Both of these samples were well-represented in the datasets with core genes > 99% (32 out 34 for ST459, 20 of 22 for ST28) indicating uneven sequence depth being the reason for poor performance.

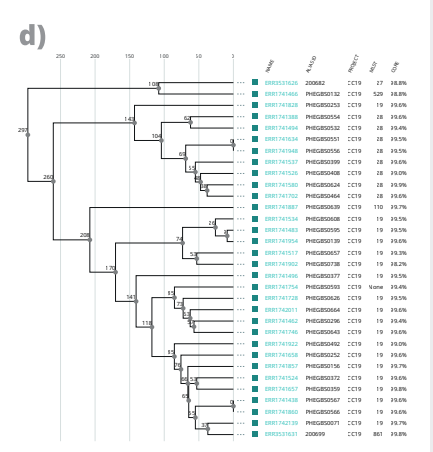
SAMPLE ACCESSION	CORE-GENE %	SEQUENCE TYPE
SRR4414147	84.10%	ST459
ERR1741716	94.28%	ST28

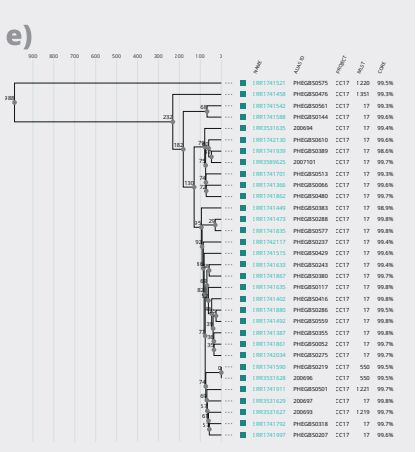
The phylogenetic relatedness was evaluated within five groups of clonal complexes and compared to the results in [1].
The groups were: a) CC23. b) CC1. c) CC8/CC10. d) CC19. e) CC17











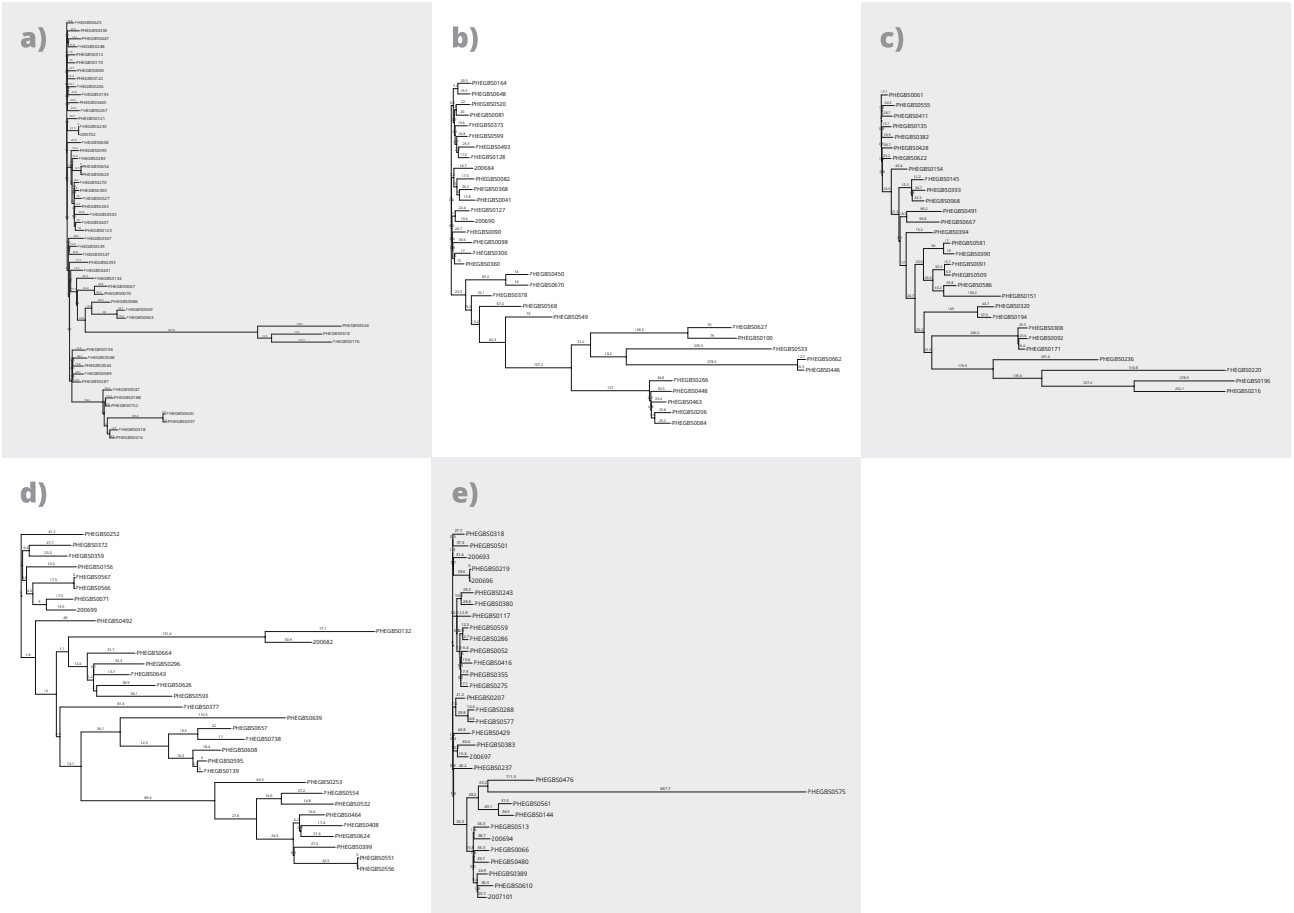
Phylogenetic trees generated using UPGMA clustering.
Each tree shows one of the 5 groups of clonal complexes.

Conclusions

The core scheme performed well and is robust within all the phylogenetic clades being tested. Comparing the phylogenetic trees generated from pairwise comparisons of the cgMLST profiles show that the core schema is able to capture the same clusters as the SNP analysis in the original article at a similar or slightly lower resolution. Since core gene alleles can capture one or more SNP this slightly lower resolution is expected.

Supplementary figure

Neighbour-joining tree generated from distance matrices of pairwise comparison of cgMLST profiles – for easy comparison with original article using the same clustering method.



References

1. Khan UB, Jauneikaite E, Andrews R, Chalker VJ, Spiller OB. **Identifying large-scale recombination and capsular switching events in *Streptococcus agalactiae* strains causing disease in adults in the UK between 2014 and 2015**. Microbial Genomics. Microbiology Society; 2022. doi:10.1099/mgen.0.000783