METHODS

Whole genome sequencing and processing

Bacterial DNA was purified utilizing the DNeasy Blood & Tissue kit (Qiagen). Sequencing libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit (NEB). Genomes were sequenced using Illumina MiSeq (2x300 bp) or for isolate LUMC16 Illumina HiSeq2500 (2x150) paired-end sequencing, resulting in 1.0-1.7x10^6 (STSS: 5.7x10^6) read pairs. Sequencing data are deposited in the NCBI SRA and Genome database under accession numbers SAMN10414105–13 and BioProject accession number PRJNA505411.

Raw reads were processed in paired-end mode by fastq-mcf version 1.05. Reads were trimmed with a Phred quality score cut-off of 30 and reads shorter than 50 nucleotides were discarded after quality trimming and adapter clipping. PhiX sequences, added as an internal Illumina sequencing control, were identified by aligning preprocessed reads to the phi-X174 genome (NC_001422.1) using the Bowtie2 version 2.3.2 legacy-build using default parameters. Only unmapped reads were kept and used for downstream analyses.

Processed reads were assembled using SPAdes (version 3.11.1) de novo genome assembler software [1] in careful mode. Contigs with a length lower than 200 bp or average coverage lower than 15 were discarded. Average contig coverage was determined by first mapping processed reads to the assembled contigs utilizing Bowtie2 (version 2.3.2) determining the per base coverage using the SAMtools (version 1.8) [2] and calculating the mean coverage for each contig. All assemblies were annotated utilizing Prokka genome annotation software (version 1.4) [3] using a *Streptococcus agalactiae* database.

Gene comparisons are based on orthologous genes identified by the Proteinortho software (version 5.16b) [4]. Virulence factors were identified by their cognate *S. agalactiae* 2603VR locus tag extracted from the Virulence Factor Database (VFDB) [5].

Whole genome phylogenies were reconstructed by the Genome-to-Genome Distance Calculator (version 2.1) [6]. A distance matrix based on DDH distances was calculated by the recommended DDH formula 2 [7]. Based on the distance matrix a Neighbor Joining tree was constructed and visualized by the APE package (version 5.1) [8] implemented in R.

Processed reads were mapped to the *Streptococcus agalactiae* 2603/VR reference genome (GCF_000007265.1) using Bowtie2 [9]. Variants were called using mpileup part of BCFtools (version 1.8) [10]. Variants were filtered out by a minimum SNP and INDEL distance of 10 bp, variant quality of 30, coverage of 50, mapping quality of 40 and a Z-score of 1.96 [11].

References:

1. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **2012**; 19:455-77.
2. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics **2009**; 25:2078-9.
3. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics **2014**; 30:2068-9.
4. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics **2011**; 12:124.
5. Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res **2005**; 33:D325-8.
6. Meier-Kolthoff JP, Auch AF, Klenk HP, Goker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics **2013**; 14:60.
7. Auch AF, Klenk HP, Goker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci **2010**; 2:142-8.
8. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics **2004**; 20:289-90.
9. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**; 9:357-9.
10. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **2011**; 27:2987-93.
11. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One **2014**; 9:e104984.