



Universiteit
Leiden

The Netherlands

Guardians of the gut: harnessing bioinformatics to study the gut microbiome and faecal microbiota transplantation in intestinal disorders

Nooij, S.

Citation

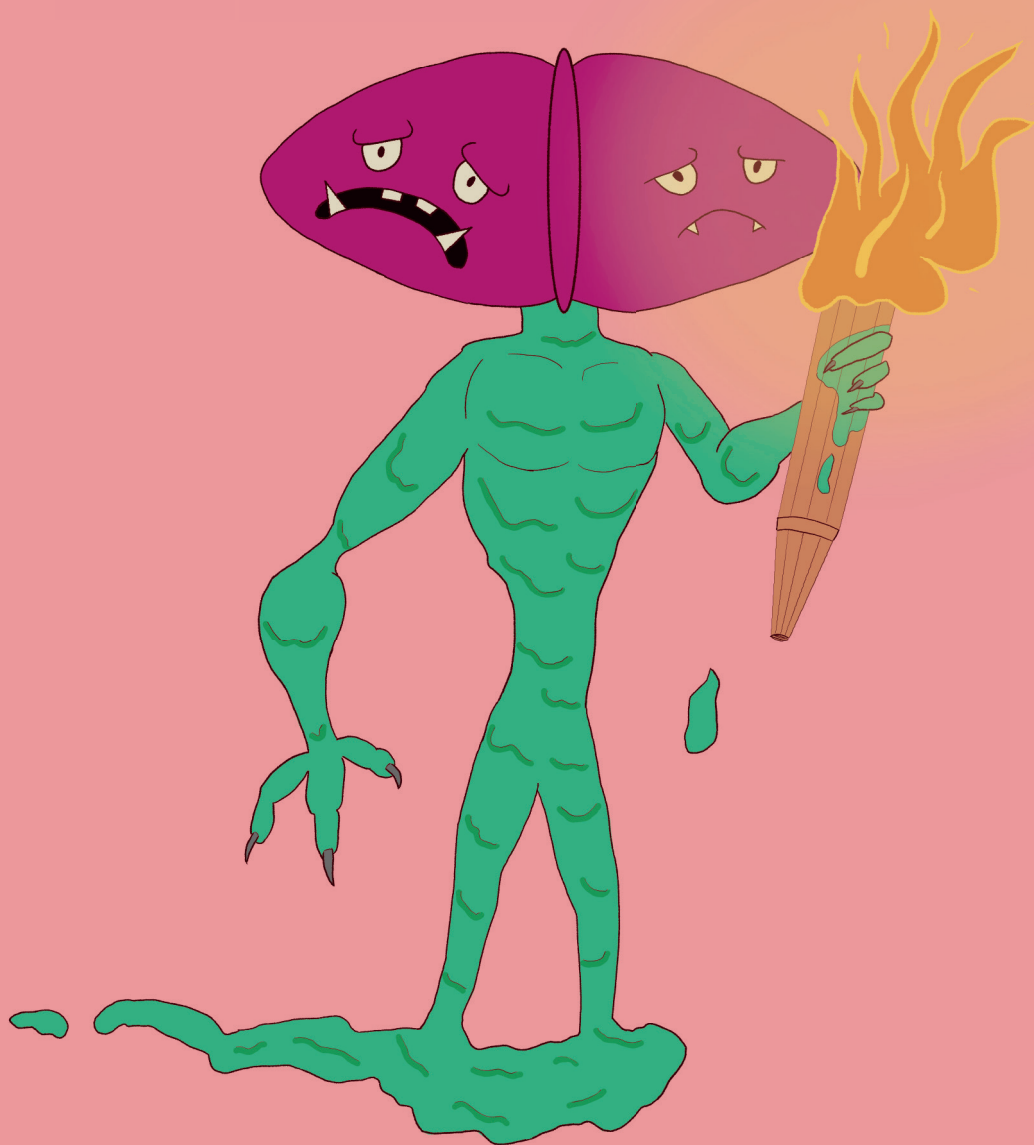
Nooij, S. (2025, October 10). *Guardians of the gut: harnessing bioinformatics to study the gut microbiome and faecal microbiota transplantation in intestinal disorders*. Retrieved from <https://hdl.handle.net/1887/4262800>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4262800>

Note: To cite this publication please use the final published version (if applicable).



Chapter 7.2

Draft and complete genome sequences of 17 *Streptococcus* species

Sam Nooij^a, Ingrid M.J.G. Sanders^a, Lesley Schout^a, Rolf H.A.M. Vossen^b, Susan L. Kloet^b, Wiep Klaas Smits^a, Quinten R. Ducarmon^a

^aLeiden University Center of Infectious Diseases (LUCID), Leiden University Medical Center, Leiden, The Netherlands

^bLeiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands

Microbial Resource Announcements, Accepted 7 July 2025

Abstract

We present seventeen near-complete and complete genomes of *Streptococcus* species obtained from eight mixed cultures of presumed *Ruminococcus gnavus* isolates. The genomes are classified as eight different *Streptococcus* species and three are unclassified (currently have no species representative available in databases). We provide these high-quality genomes to the scientific community for further scrutinization.

Announcement

In a project focusing on isolating and sequencing genomes of *Ruminococcus gnavus* (reference *R. gnavus* manuscript), we sent presumed pure *R. gnavus* DNA for long-read PacBio circular consensus sequencing (CCS) to the Leiden Genome Technology Center (Leiden University Medical Center, Leiden, Netherlands). The DNA libraries were prepared using the SMRTbell prep kit 3.0 kit and sequenced on the Sequel II platform (Pacific Biosciences of California, Inc., CA, USA). Raw reads were assembled with Flye¹ (version 2.9.2) and resulting contigs were classified using the Contig Annotation Tool² (CAT; version 5.2.3). Contigs were reoriented to start at the *dnaA* for putative chromosomes using dnaapler³ (version 0.3.0). To our surprise, eight of the samples returned complete *Streptococcus* genomes and short (7-55kbp), predicted *R. gnavus* contigs with an estimated depth of coverage of 4X. As we suppose these *Streptococcus* are lab contaminants, we cannot be certain of their exact origin. Nonetheless, we aimed to generate complete genomes and improved assemblies of *Streptococcus* genomes by using a metagenome assembly approach with metaFlye⁴ (version 2.9.2). This yielded up to 194 contigs per sample with a median length of 23kbp. We separated the contigs classified as *Streptococcus* using seqkit⁵ (version 0.16.0) and did a quality control consisting of: 1) a length and GC-content check by QUAST⁶ (version 5.0.2), 2) completeness and contamination estimation by CheckM⁷ (version 1.0.13), CheckM2⁸ (version 1.0.1) and BUSCO⁹ (version 5.4.3; using flag '--auto-lineage-prok'), 3) gene annotation with Bakta¹⁰ (version 1.6.1), and 4) taxonomic classification with GTDB-Tk¹¹ (version 2.3.2) using the Genome Taxonomy database (GTDB) version r207_v2. All tools were run with default parameters unless stated otherwise. Relevant statistics of each genome are listed in table 1. Furthermore, to determine the relatedness between the *Streptococcus* genomes and infer if they might originate from a common source, we used fastANI¹² (version 1.33) to calculate pairwise average nucleotide identity (ANI; figure 1). We found five *Streptococcus equinus* (NCBI taxonomy; GTDB: *Streptococcus sp001556435*) highly similar genomes (ANI $\geq$ 99.9%) from five different samples, four *S. oralis* (GTDB: 2 $\times$ identical *S. oralis_S* and 2 $\times$ *S. oralis_Y*) from three different samples, and six genomes with slightly different classifications that are all close to *S. parasanguinis* (ANI $\geq$ 93.8%) from three different samples (1 $\times$ 3, 1 $\times$ 2, 1 $\times$ 1 genome/sample). We found one genome classified as *S. sp902363395* (GTDB), which is unknown

to NCBI, that remotely resembles (ANI ~86.4%) the *parasanguinis*-like genomes. The final genome was classified as *S. sanguinis* SK49 (GTDB: *S. sanguinis*_C) and is only remotely similar to the other genomes (ANI $\leq$ 80%).

All raw reads and assembled genomes are available through the European Nucleotide Archive under project number [PRJEB76410](https://www.ebi.ac.uk/ena/browser/view/PRJEB76410).

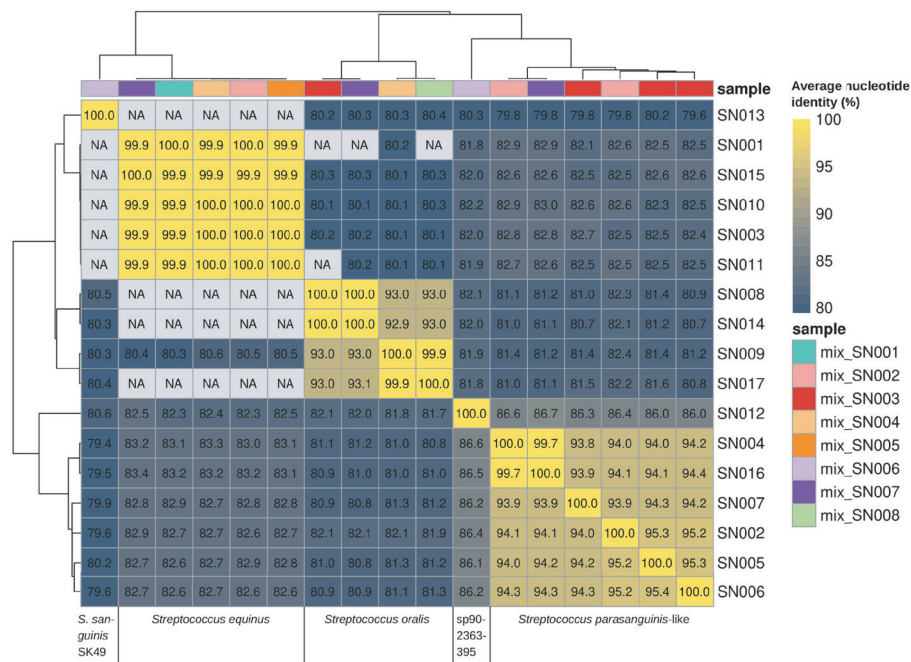


Figure 1. Whole-genome comparison of 17 *Streptococcus* genomes. Whole-genome sequences of 17 *Streptococcus* bacteria were compared using Average Nucleotide Identity (ANI). Values in cells indicate percentage ANI. Genomes are annotated with the sample they derive from.

Acknowledgements

This study was sponsored by the Leiden University Fund / Dr. F.F. Hofman Fonds, www.luf.nl.

References

1. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37, 540-546. 10.1038/s41587-019-0072-8.
2. von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 20, 217. 10.1186/s13059-019-1817-x.
3. Bouras, G., Grigson, S., Papudeshi, B., Mallawaarachchi V., Roach, M. J. (2023). Dnaapler: A tool to reorient circular microbial genomes <https://github.com/gbouras13/dnaapler>.
4. Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., and Pevzner, P.A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 17, 1103-1110. 10.1038/s41592-020-00971-x.
5. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* 11, e0163962. 10.1371/journal.pone.0163962.
6. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072-1075. 10.1093/bioinformatics/btt086.
7. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25, 1043-1055. 10.1101/gr.186072.114.
8. Chklovski, A. (2024). CheckM2. <https://github.com/chklovski/CheckM2>.
9. Manni, M., Berkeley, M.R., Seppey, M., Simao, F.A., and Zdobnov, E.M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* 38, 4647-4654. 10.1093/molbev/msab199.
10. Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., and Goesmann, A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* 7. 10.1099/mgen.0.000685.
11. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38, 5315-5316. 10.1093/bioinformatics/btac672.
12. Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9, 5114. 10.1038/s41467-018-07641-9.

Table 1. Streptococcus genome assembly characteristics and quality measures. Initial assemblies were made with Flye. In case of multiple linear contigs, a secondary assembly was made using metaFlye. The most complete assembly was selected from either approach.

Genome	Length (bp)	Predicted genes	GC%	Circular?	Species	CheckM completeness	CheckM contamination	Assembly method	Depth of coverage
SN001	2140162	1914	40.12	Y	<i>Streptococcus equinus</i>	99.69	0.15	metaFlye	17
SN002	2223664	2093	41.52	N	<i>Streptococcus</i> sp. HMSC072G04	100	0.09	Flye	13
SN003	2141365	1913	40.11	Y	<i>Streptococcus equinus</i>	99.69	0.15	Flye	21
SN004	2160498	2021	42.06	N	<i>Streptococcus</i> sp.	100	0.17	metaFlye	18
SN005	2125957	2016	41.72	Y	<i>Streptococcus parasanguinis</i>	99.66	0.11	Flye	120
SN006	2081198	1934	42.11	Y	<i>Streptococcus parasanguinis</i>	100	0.07	Flye	120
SN007	2118644	2017	41.95	Y	<i>Streptococcus</i> sp.	100	0.23	Flye	158
SN008	2068017	1940	41.06	Y	<i>Streptococcus oralis</i>	99.87	0.04	Flye	25
SN009	2117116	2018	40.83	Y	<i>Streptococcus oralis</i>	99.87	0.06	metaFlye	24
SN010	2147719	1922	40.12	Y	<i>Streptococcus equinus</i>	99.69	0.15	metaFlye	25
SN011	2141236	1914	40.11	N	<i>Streptococcus equinus</i>	99.69	0.15	metaFlye	14
SN012	2025938	1881	42.21	Y	<i>Streptococcus</i> sp.	100	0.17	metaFlye	21
SN013	2316732	2233	43	N	<i>Streptococcus sanguinis</i>	100	0	metaFlye	20
SN014	2068026	1939	41.06	Y	<i>Streptococcus oralis</i>	99.87	0.04	metaFlye	117
SN015	2141124	1914	40.12	Y	<i>Streptococcus equinus</i>	99.69	0.15	metaFlye	87
SN016	2166044	2028	42.02	Y	<i>Streptococcus</i> sp.	100	0.17	Flye	153
SN017	2119531	2017	40.83	Y	<i>Streptococcus oralis</i>	99.87	0.6	metaFlye	48