



PRINT DATE: 2021-12-20 16:05:44 +0100

JOB ID: 3991255f-7f78-4778-8386-39924f41cb08

RESULT PAGE: [https://tygs.dsmz.de/user\\_results/show?guid=3991255f-7f78-4778-8386-39924f41cb08](https://tygs.dsmz.de/user_results/show?guid=3991255f-7f78-4778-8386-39924f41cb08)

## Table 1: Phylogenies

**Publication-ready versions** of both the genome-scale GBDP tree and the 16S rRNA gene sequence tree can be customized and exported either in SVG (vector graphic) or PNG format from within the phylogeny viewers in your TYGS result page. For publications the **SVG format is recommended** because it is lossless, always keeps its high resolution and can also be easily converted to other popular formats such as PDF or EPS. Please follow the link provided above!

## Table 2: Identification

**Note:** the identification of your genomes is not possible in the currently chosen mode because the analysis was restricted to user genomes only.

## Table 3: Pairwise comparisons of user genomes vs. type-strain genomes

The following table contains the pairwise dDDH values between your user genomes and the selected type-strain genomes. The dDDH values are provided along with their confidence intervals (C.I.) for the three different GBDP formulas:

- formula  $d_0$  (a.k.a. GGDC formula 1): length of all HSPs divided by total genome length
- formula  $d_4$  (a.k.a. GGDC formula 2): sum of all identities found in HSPs divided by overall HSP length
- formula  $d_6$  (a.k.a. GGDC formula 3): sum of all identities found in HSPs divided by total genome length

**Note:** Formula  $d_4$  is independent of genome length and is thus robust against the use of incomplete draft genomes. For other reasons for preferring formula  $d_4$ , see the FAQ.

Query	Subject	$d_0$	C.I. $d_0$	$d_4$	C.I. $d_4$	$d_6$	C.I. $d_6$	Diff. G+C Percent
'Enterococcus_faecalis_complete_genome.fasta'	'Listeria_monocytogenes_complete_genome.fasta'	13.1	[10.3 - 16.4]	36.7	[34.3 - 39.2]	13.5	[11.1 - 16.3]	0.43
'Listeria_monocytogenes_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.7	[10.0 - 16.0]	31.6	[29.2 - 34.1]	13.1	[10.8 - 15.9]	5.1
'Bacillus_subtilis_complete_genome.fasta'	'Listeria_monocytogenes_complete_genome.fasta'	12.8	[10.1 - 16.1]	29.6	[27.2 - 32.1]	13.2	[10.9 - 16.0]	5.98
'Bacillus_subtilis_complete_genome.fasta'	'Lactobacillus_fermentum_complete_genome.fasta'	12.7	[10.0 - 16.0]	29.1	[26.7 - 31.6]	13.1	[10.8 - 15.9]	8.36
'Lactobacillus_fermentum_complete_genome.fasta'	'Listeria_monocytogenes_complete_genome.fasta'	12.7	[10.0 - 16.0]	28.9	[26.5 - 31.4]	13.1	[10.8 - 15.9]	14.34
'Enterococcus_faecalis_complete_genome.fasta'	'Lactobacillus_fermentum_complete_genome.fasta'	12.7	[10.0 - 16.0]	28.7	[26.3 - 31.2]	13.1	[10.8 - 15.8]	14.78
'Bacillus_subtilis_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.8	[10.1 - 16.1]	28.6	[26.3 - 31.1]	13.2	[10.9 - 16.0]	11.08

Query	Subject	$d_0$	C.I. $d_0$	$d_4$	C.I. $d_4$	$d_6$	C.I. $d_6$	Diff. G+C Percent
'Lactobacillus_fermentum_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.7	[10.0 - 16.0]	27.0	[24.6 - 29.4]	13.1	[10.8 - 15.9]	19.44
'Bacillus_subtilis_complete_genome.fasta'	'Enterococcus_faecalis_complete_genome.fasta'	12.8	[10.1 - 16.0]	26.5	[24.1 - 29.0]	13.2	[10.8 - 15.9]	6.42
'Escherichia_coli_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	31.8	[28.4 - 35.3]	24.2	[21.8 - 26.6]	28.9	[26.0 - 32.0]	1.61
'Enterococcus_faecalis_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.8	[10.1 - 16.0]	23.9	[21.6 - 26.4]	13.2	[10.8 - 15.9]	4.66
'Escherichia_coli_complete_genome.fasta'	'Pseudomonas_aeruginosa_complete_genome.fasta'	12.6	[10.0 - 15.9]	21.9	[19.7 - 24.4]	13.0	[10.7 - 15.8]	15.59
'Pseudomonas_aeruginosa_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	12.6	[10.0 - 15.9]	21.6	[19.4 - 24.1]	13.0	[10.7 - 15.8]	13.99
'Bacillus_subtilis_complete_genome.fasta'	'Pseudomonas_aeruginosa_complete_genome.fasta'	12.6	[9.9 - 15.9]	19.0	[16.8 - 21.4]	13.0	[10.7 - 15.7]	22.23
'Salmonella_enterica_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.6	[9.9 - 15.9]	18.5	[16.3 - 20.9]	13.0	[10.7 - 15.8]	19.33
'Pseudomonas_aeruginosa_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.6	[9.9 - 15.8]	18.5	[16.3 - 20.8]	13.0	[10.6 - 15.7]	33.31
'Escherichia_coli_complete_genome.fasta'	'Staphylococcus_aureus_complete_genome.fasta'	12.6	[9.9 - 15.9]	18.4	[16.2 - 20.8]	13.0	[10.7 - 15.8]	17.72
'Listeria_monocytogenes_complete_genome.fasta'	'Pseudomonas_aeruginosa_complete_genome.fasta'	12.6	[9.9 - 15.8]	18.1	[16.0 - 20.5]	13.0	[10.6 - 15.7]	28.22
'Bacillus_subtilis_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	12.6	[10.0 - 15.9]	18.0	[15.9 - 20.3]	13.0	[10.7 - 15.8]	8.25
'Enterococcus_faecalis_complete_genome.fasta'	'Pseudomonas_aeruginosa_complete_genome.fasta'	12.6	[9.9 - 15.8]	17.7	[15.5 - 20.0]	13.0	[10.6 - 15.7]	28.65
'Bacillus_subtilis_complete_genome.fasta'	'Escherichia_coli_complete_genome.fasta'	12.6	[10.0 - 15.9]	17.6	[15.4 - 19.9]	13.0	[10.7 - 15.8]	6.64
'Listeria_monocytogenes_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	12.6	[9.9 - 15.9]	17.4	[15.3 - 19.8]	13.0	[10.7 - 15.8]	14.23
'Enterococcus_faecalis_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	12.6	[9.9 - 15.9]	17.4	[15.2 - 19.7]	13.0	[10.7 - 15.7]	14.66
'Enterococcus_faecalis_complete_genome.fasta'	'Escherichia_coli_complete_genome.fasta'	12.6	[9.9 - 15.9]	17.1	[15.0 - 19.4]	13.0	[10.7 - 15.7]	13.06
'Lactobacillus_fermentum_complete_genome.fasta'	'Pseudomonas_aeruginosa_complete_genome.fasta'	12.6	[9.9 - 15.8]	17.1	[15.0 - 19.4]	13.0	[10.6 - 15.7]	13.87
'Escherichia_coli_complete_genome.fasta'	'Listeria_monocytogenes_complete_genome.fasta'	12.6	[10.0 - 15.9]	17.0	[14.9 - 19.3]	13.0	[10.7 - 15.8]	12.62
'Lactobacillus_fermentum_complete_genome.fasta'	'Salmonella_enterica_complete_genome.fasta'	12.6	[10.0 - 15.9]	16.6	[14.5 - 18.9]	13.0	[10.7 - 15.8]	0.11
'Escherichia_coli_complete_genome.fasta'	'Lactobacillus_fermentum_complete_genome.fasta'	12.6	[10.0 - 15.9]	16.4	[14.3 - 18.7]	13.0	[10.7 - 15.8]	1.72

Table 4: Strains in your dataset

Joint dataset of automatically determined closest type strains (if this mode was chosen), manually selected type strains (if selected accordingly) and the provided user strains, if provided (marked in **yellow**).

Strain	Authority	Other deposits	Synonyms	Base pairs	Percent G+C	No. proteins	Goldstamp	Bioproject accession	Biosample accession	Assembly accession	IMG OID
Bacillus_subtilis_complete_genome.fasta				4045677	43.9	3998					
Enterococcus_faecalis_complete_genome.fasta				2845392	37.5	2665					
Escherichia_coli_complete_genome.fasta				4765434	50.6	4443					
Lactobacillus_fermentum_complete_genome.fasta				1905333	52.3	1869					
Listeria_monocytogenes_complete_genome.fasta				2992342	38.0	2906					
Pseudomonas_aeruginosa_complete_genome.fasta				6792330	66.2	6281					
Salmonella_enterica_complete_genome.fasta				4759746	52.2	4552					
Staphylococcus_aureus_complete_genome.fasta				2718780	32.9	2468					

## Methods, Results and References

The genome sequence data were uploaded to the Type (Strain) Genome Server (TYGS), a free bioinformatics platform available under <https://tygs.dsmz.de>, for a whole genome-based taxonomic analysis [1]. The analysis also made use of recently introduced methodological updates and features [2]. Information on nomenclature, synonymy and associated taxonomic literature was provided by TYGS's sister database, the List of Prokaryotic names with Standing in Nomenclature (LPSN, available at <https://lpsn.dsmz.de>) [2]. The results were provided by the TYGS on 2021-12-20. The TYGS analysis was subdivided into the following steps:

### Determination of closely related type strains

Determination of closest type strain genomes was done in two complementary ways: First, all user genomes were compared against all type strain genomes available in the TYGS database via the MASH algorithm, a fast approximation of intergenomic relatedness [3], and, the ten type strains with the smallest MASH distances chosen per user genome. Second, an additional set of ten closely related type strains was determined via the 16S rDNA gene sequences. These were extracted from the user genomes using RNAmmer [4] and each sequence was subsequently BLASTed [5] against the 16S rDNA gene sequence of each of the currently 15857 type strains available in the TYGS database. This was used as a proxy to find the best 50 matching type strains (according to the bitscore) for each user genome and to subsequently calculate precise distances using the Genome BLAST Distance Phylogeny approach (GBDP) under the algorithm 'coverage' and distance formula  $d_5$  [6]. These distances were finally used to determine the 10 closest type strain genomes for each of the user genomes.

### Pairwise comparison of genome sequences

For the phylogenomic inference, all pairwise comparisons among the set of genomes were conducted using GBDP and accurate intergenomic distances inferred under the algorithm 'trimming' and distance formula  $d_5$  [6]. 100 distance replicates were calculated each. Digital DDH values and confidence intervals were calculated using the recommended settings of the GGDC 3.0 [2,6].

### Phylogenetic inference

The resulting intergenomic distances were used to infer a balanced minimum evolution tree with branch support via FASTME 2.1.6.1 including SPR postprocessing [7]. Branch support was inferred from 100 pseudo-bootstrap replicates each. The trees were rooted at the midpoint [8] and visualized with PhyD3 [9].

### Type-based species and subspecies clustering

The type-based species clustering using a 70% dDDH radius around each of the 0 type strains was done as previously described [1]. The resulting groups are shown in Table 1 and 4. Subspecies clustering was done using a 79% dDDH threshold as previously introduced [10].

## Results

### Type-based species and subspecies clustering

The resulting species and subspecies clusters are listed in Table 4, whereas the taxonomic identification of the query strains is found in Table 1. Briefly, the clustering yielded species clusters and the provided query strains were assigned to of these. Moreover, user strains were located in of subspecies clusters.

### Figure caption SSU tree

**Figure 1.** Tree inferred with FastME 2.1.6.1 [7] from GBDP distances calculated from 16S rDNA gene sequences. The branch lengths are scaled in terms of GBDP distance formula  $d_5$ . The numbers above branches are GBDP pseudo-bootstrap support values > 60 % from 100 replications, with an average branch support of 92.2 %. The tree was rooted at the midpoint [8].

### Figure caption genome tree

**Figure 2.** Tree inferred with FastME 2.1.6.1 [7] from GBDP distances calculated from genome sequences. The branch lengths are scaled in terms of GBDP distance formula  $d_5$ . The numbers above branches are GBDP pseudo-bootstrap support values > 60 % from 100 replications, with an average branch support of 64.8 %. The tree was rooted at the midpoint [8].

## References

- [1] Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun.* 2019;10: 2182. DOI: 10.1038/s41467-019-10210-3
- [2] Meier-Kolthoff JP, Sardà Carbasse J, Peinado-Olarte RL, Göker M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acid Res.* 2021. Epub ahead of print. DOI: 10.1093/nar/gkab902
- [3] Ondov BD, Treangen TJ, Melsted P, et al. Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17: 1–14. DOI: 10.1186/s13059-016-0997-x
- [4] Lagesen K, Hallin P. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* Oxford Univ Press; 2007;35: 3100–3108. DOI: 10.1093/nar/gkm160
- [5] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10: 421. DOI: 10.1186/1471-2105-10-421
- [6] Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics.* 2013;14: 60. DOI: 10.1186/1471-2105-14-60
- [7] Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol Biol Evol.* 2015;32: 2798–2800. DOI: 10.1093/molbev/msv150
- [8] Farris JS. Estimating phylogenetic trees from distance matrices. *Am Nat.* 1972;106: 645–667.
- [9] Kreft L, Botzki A, Coppens F, Vandepoele K, Van Bel M. PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics.* 2017;33: 2946–2947. DOI: 10.1093/bioinformatics/btx324
- [10] Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, et al. Complete genome sequence of DSM 30083<sup>T</sup>, the type strain (U5/41<sup>T</sup>) of *Escherichia coli*, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci.* 2014;9: 2. DOI: 10.1186/1944-3277-9-2