assembly_summary_genbank.txt
1,321,179 genome sequences (July 5th, 2022)

Download Genbank assembly records

Extract species names: removed strain name, subsp names, changed HMT XXX to HMT-XXX

List of species names

HOMD Species Names (822)

Compile HOMD taxon names: Remove “clade” designation

387,630 genomes

Screen for potential HOMD genomes:
1. 822 HOMD Scientific Names
2. Contains “oral taxon xxx”

Screen for potential HOMD genomes:
3. 303 verified *Rothia*, *Veillonella* and *Streptococcus* species
4. Exclude genomes without GCF
5. Exclude “metagenomes”

Order genomes in each taxon by:
1. "Complete Genome "
    1. "reference genome” or "representative genome“
    2. "assembly from type material"
2. " Chromosome“
    1. "reference genome” or "representative genome“
    2. "assembly from type material"
3. Sort the remaining genomes by number of contigs

For each taxon:
if name is in the “white list”
    Select all genomes
else
    Select first (or up to) 50 genomes from the ordered genome list based on above priority

8,400 genomes

Visually inspect the PhyloPhlAn tree
1. Remove genomes out of place
2. Remove poor quality genomes
3. Removed genomes recorded in an Excel file

8,259 genomes V10.1b

Visually inspect the PhyloPhlAn tree second round
1. Remove genomes out of place
2. Remove poor quality genomes
3. Removed genomes recorded in an Excel file

8148 genomes V10.1fa

Added 474 existing genomes not from above name-based search

8622 genomes V10.1 Final

DATABASE

Deposit NCBI and PROKKA annotations to mySQL database:

- 6,933 of 8,622 genomes have NCBI annotation
- All 8,622 genomes have PROKKA annotation

HOMD :: Genome Table

Phylogenetic Trees

Phylogenetic trees with links to taxonomy and genome info pages:
1. PhyloPhlAn conserved protein tree (8,622 genomes, 568 taxa)
2. Ribosomal protein tree (8,622 genomes, 568 taxa)
3. 16S rRNA tree (8,312 genomes, 562 taxa)

https://homd.org/ftp/phylogenetic_trees/genome/

JBrowse

- All 8,622 genomes have PROKKA annotation panels (protein, RNA)
- Only 6,933 have NCBI annotation panels

HOMD JBrowse SEQF1595.2|KI535340.1:634977..952472

FTP

- Provides genomic contigs, annotated protein sequences (PROKKA and NCBI) and nucleotide sequences of proteins.
- Download for all 8,600 genomes of invidual ones.

Index of /ftp/genomes/ (homd.org)

BLAST

- Provide BLAST sequence homology search for all 8,622 genomes or individual search.
- Can be searched again genomic and protein sequences.

SequenceServer: Custom BLAST Server

Viewing and evaluate on the development server:
https://devel.homd.org

Public release on main site:
https://homd.org

Genome V10.1 implementation notes:
https://homd.org/ftp/genomes/V10.1/report.html