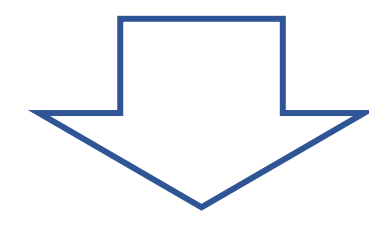


assembly\_summary\_genbank.txt  
1,321,179 genome sequences (July 5<sup>th</sup>, 2022)

Download Genbank  
assembly records

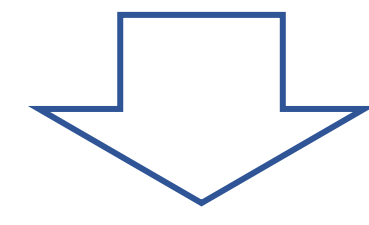


Extract species names: removed strain name,  
subsp names, changed HMT XXX to HMT-XXX

List of species names

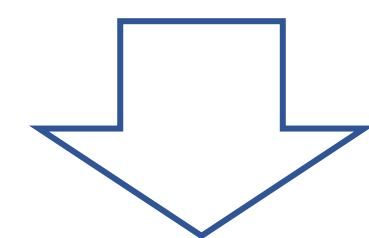
HOMD Species Names (822)

Compile HOMD taxon names:  
Remove "clade" designation

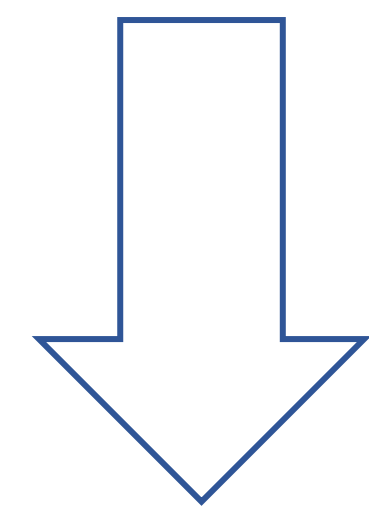


387,630 genomes

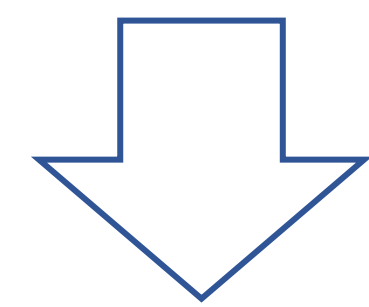
Screen for potential HOMD genomes:  
1. 822 HOMD Scientific Names  
2. Contains "oral taxon xxx"



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

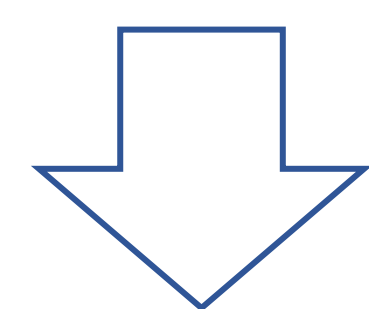


Order genomes in each taxon by:  
1. "Complete Genome "  
1. "reference genome" or "representative genome"  
2. "assembly from type material"  
2. "Chromosome"  
1. "reference genome" or "representative genome"  
2. "assembly from type material"  
3. Sort the remaining genomes by number of contigs



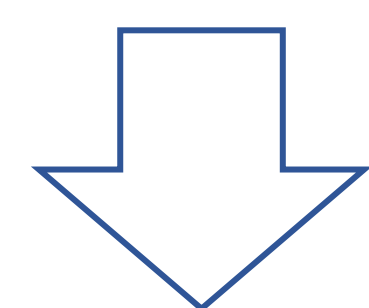
8,400 genomes

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



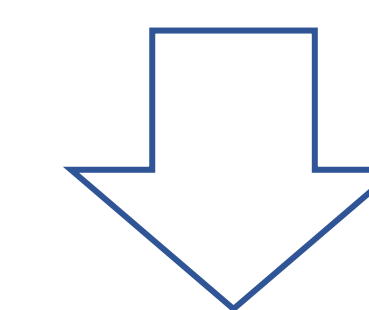
8,259 genomes V10.1b

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



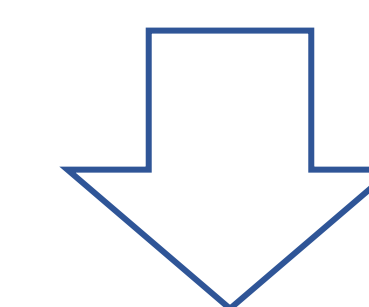
8148 genomes V10.1fa

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



Added **474** existing genomes not from above name-based search

8622 genomes V10.1 Final



Tasks to be done for final public release of V10.1:

1. Deposit both NCBI and PROKKA annotations into HOME Genome Database (Andy)
2. Compile phylogenetic trees for:
  1. Conversed protein tree (Phylophlan) (George)
  2. 16S rRNA extracted from all genomes (George)
  3. Ribosomal protein tree (George)
3. Compile BLAST data files for BLAST search (George)
4. Compile NCBI and PROKKA download files on FTP (George)
5. Render Jbrowse viewing for all genomes (George)