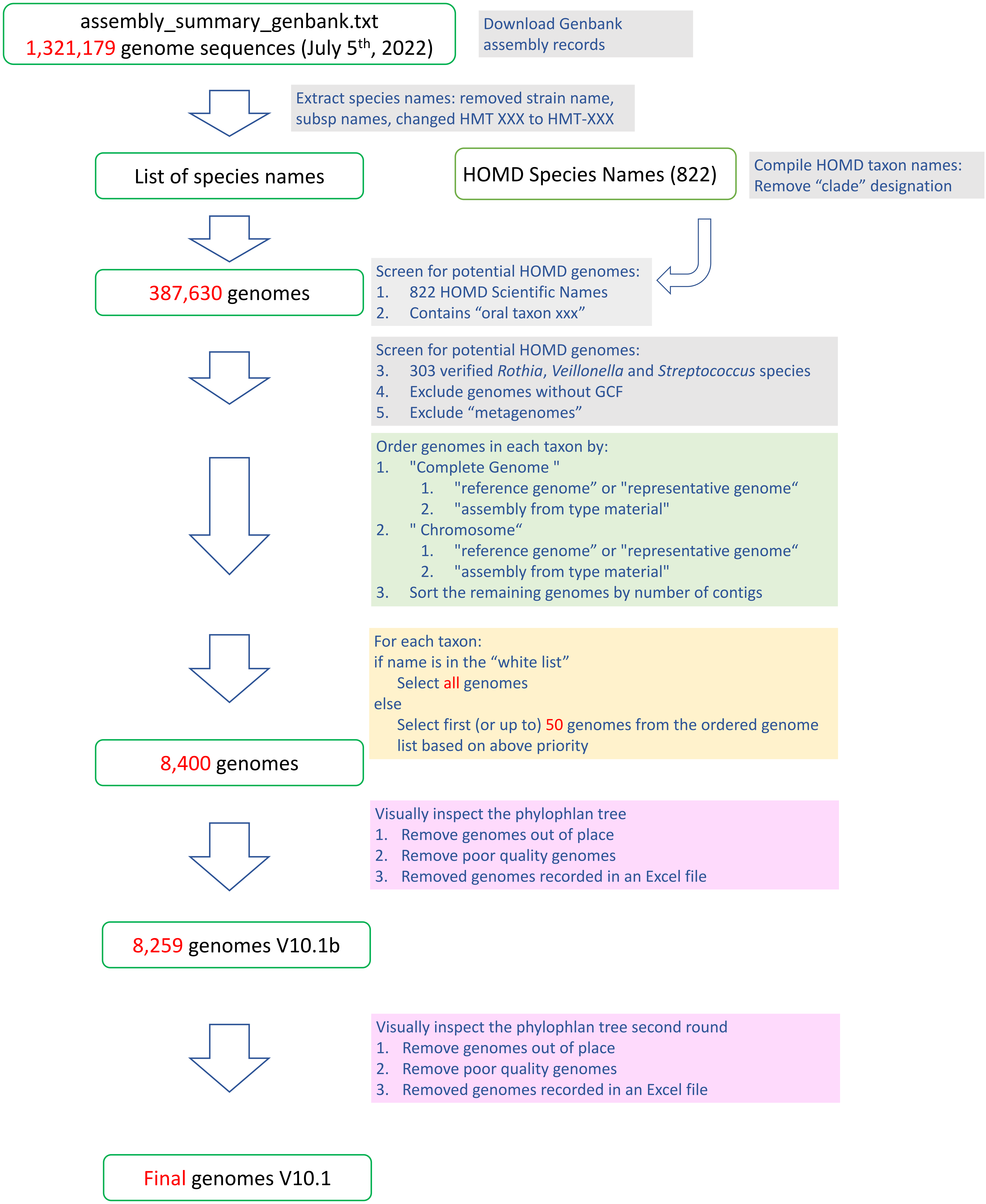

assembly_summary_genbank.txt
1,321,179 genome sequences (July 5th, 2022)

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

Final genomes V10.1