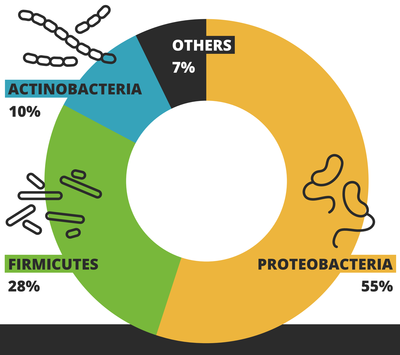
ToRQuEMaDA: Tool for retrieving queried Eubacteria, metadata and dereplicating assemblies

BACKGROUND

The number of available prokaryotic genomes is quickly growing and unevenly distributed taxonomically, which causes problems when trying to assemble high-quality yet broadly sampled genome sets for phylogenomics and comparative genomics.

There are now **over 211,000 prokaryotic genomes** available on NCBI RefSeq (as of March 2021) and these are heavily **biased towards 3 out of 53 recognized phyla**.

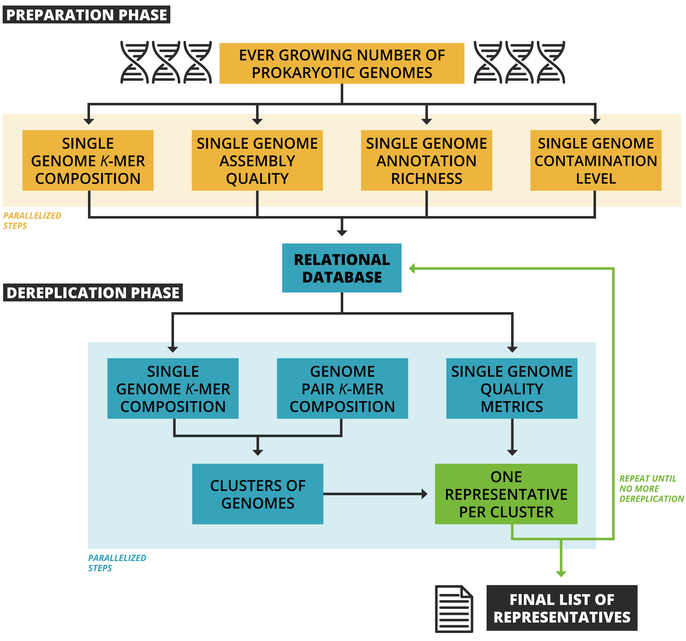


ToRQuEMaDA (TQMD)

TQMD is a tool for **high-performance computing clusters which downloads, stores and produces lists of dereplicated prokaryotic genomes**. It is based on word-based alignment-free methods (*k*-mers), an iterative single-linkage approach and a divide-and-conquer strategy for efficiency and scalability.

ToRQuEMaDA

TQMD has been developed to counter the ever-growing number of prokaryotic genomes and their uneven taxonomic distribution.



PERFORMANCE OF TQMD

We studied TQMD's performance by verifying the influence of its parameters and heuristics on the clustering outcome. We further compared TQMD to two other dereplication tools (dRep and Assembly-Dereplicator).

Our results showed that **TQMD is primarily optimized to dereplicate at higher taxonomic levels** (phylum/class), as opposed to the other dereplication tools, but **also works at lower taxonomic levels** (species/strain) like the other dereplication tools.

ToRQuEMaDA

TQMD is available from source and as a Singularity container at: https://bitbucket.org/phylogeno/tqmd

PeerJ

ToRQuEMaDA: Tool for retrieving queried Eubacteria, metadata and dereplicating assemblies
PeerJ 9:e11348 DOI: 10.7717/peerj.11348
http://peerj.com/articles/11348/