BakRep

A searchable large-scale web repository for bacterial genomes, characterizations and metadata

bakrep.computational.bio

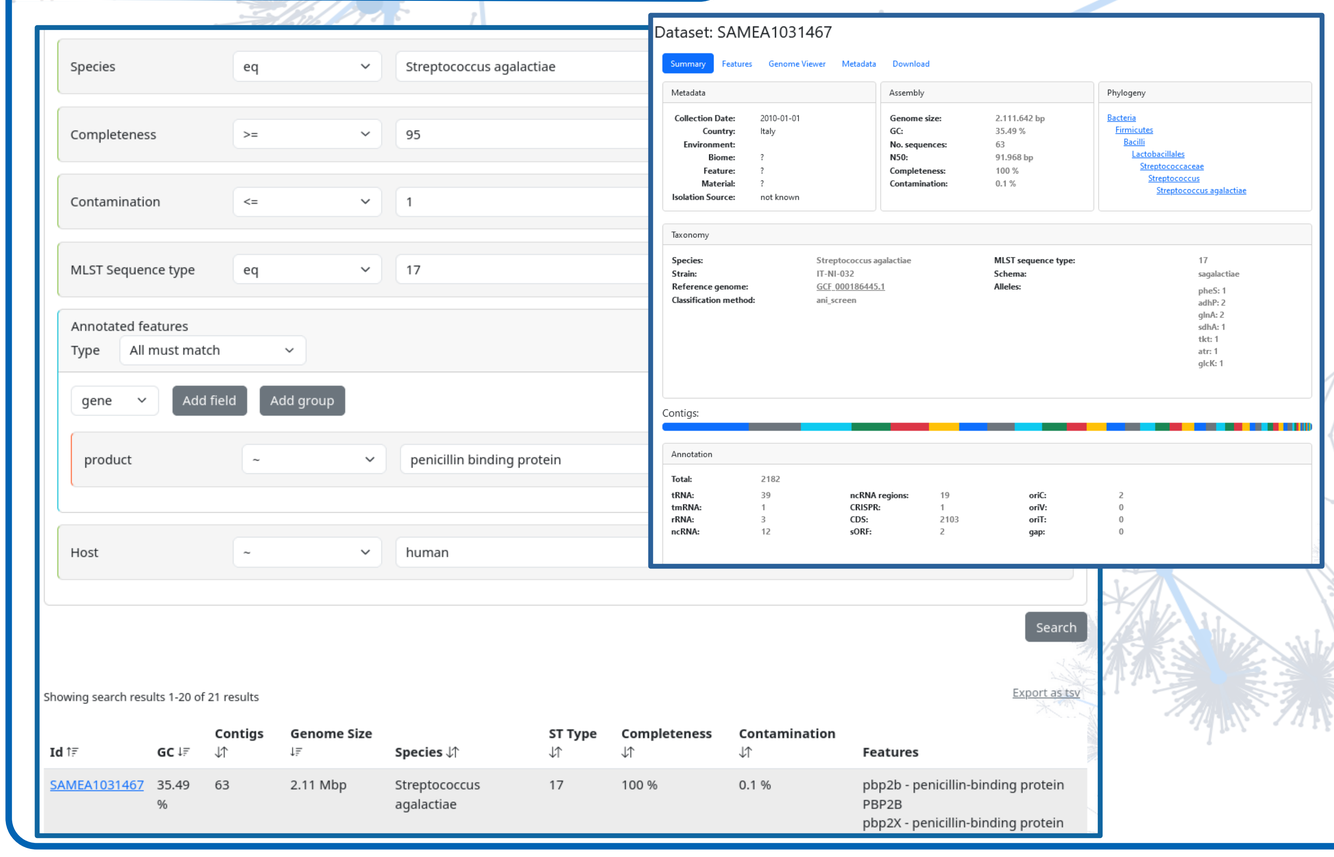
Linda Fenske[1], Lukas Jelonek[1], Alexander Goesmann[1], Oliver Schwengers[1]

[1]Justus-Liebig University Giessen, Bioinformatics and Systems Biology, Giessen, Germany

Abstract

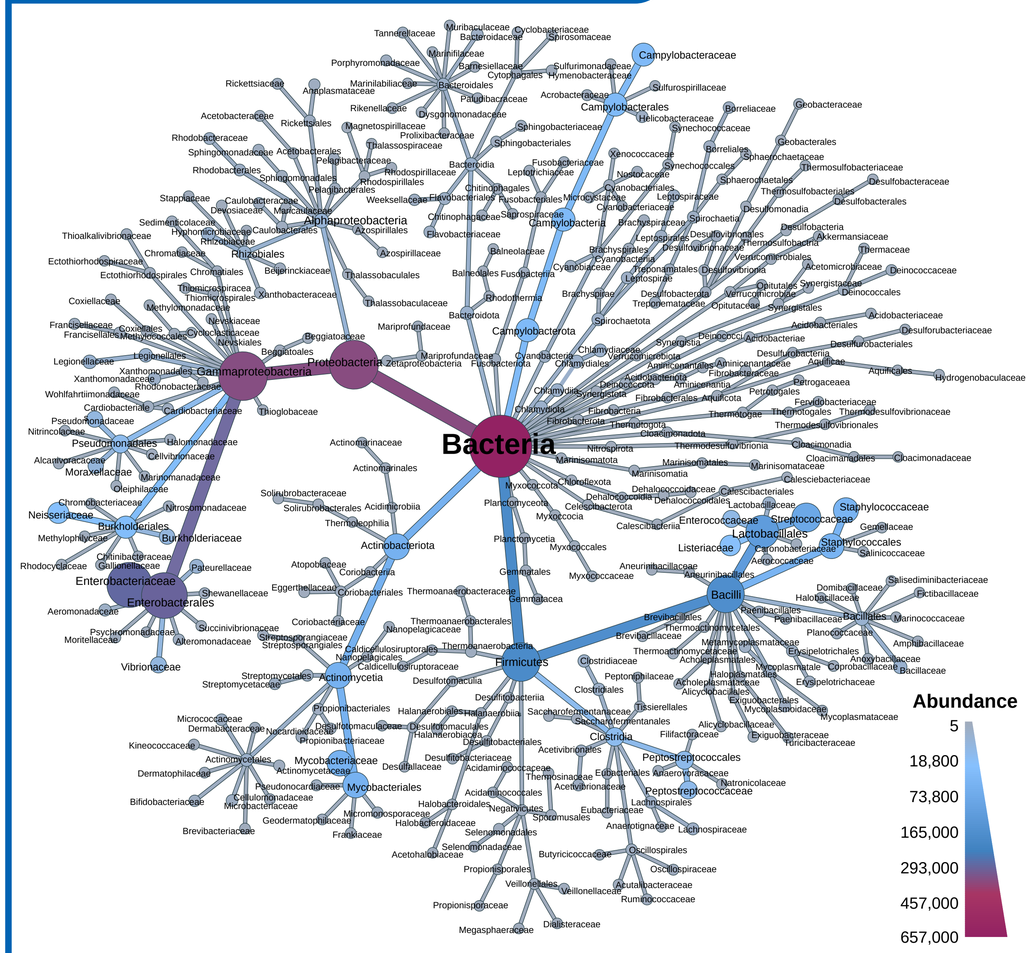
Bacteria are fascinating research objects in many disciplines for countless reasons, and whole-genome sequencing has become the paramount methodology to advance our microbiological understanding. Meanwhile, access to cost-effective sequencing platforms has accelerated bacterial **whole-genome sequencing** to unprecedented levels, introducing new challenges in terms of data accessibility, computational demands, heterogeneity of analysis workflows, and thus, ultimately its scientific usability. To that end, ***Blackwell et al.*** released a uniformly processed set of **661,405 bacterial genome assemblies** obtained from the **European Nucleotide Archive** (ENA) as of **November 2018** [1]. Building on these accomplishments, we conducted further genome-based analyses like **taxonomic classification**, **MLST subtyping**, and **annotation** of all genomes. Here, we present **BakRep**, a **searchable large-scale web repository** of these genomes enriched with consistent genome characterizations and original metadata. The platform provides a flexible search engine combining **taxonomic, genomic, and metadata information**, as well as interactive elements to visualize genomic features. Furthermore, all results can be downloaded for offline analyses via an accompanying command line tool. The web repository is accessible via **https://bakrep.computational.bio**.

Workflow and web repository



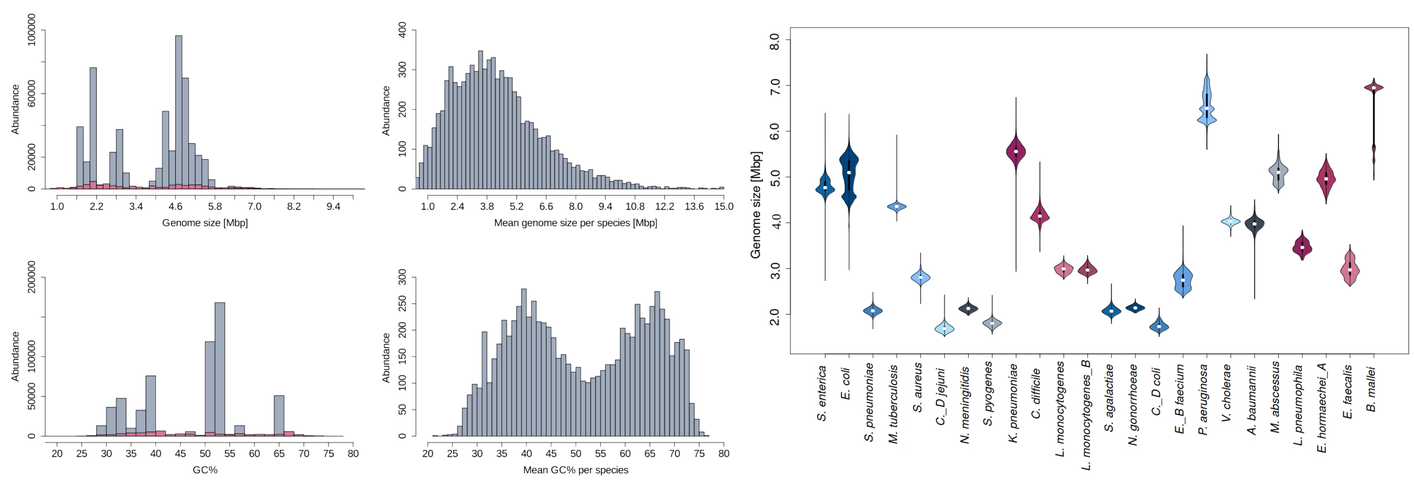
- **661,405 assemblies** and associated **metadata** from the previous study
- Taxonomic classification with **GTDB-Tk** [2]
- Contamination and completeness estimated with **CheckM2** [3]
- Assembly statistics collected with **assembly-scan**
- Multilocus sequence typing conducted using **mlst** [4]
- Annotation using **Bakta** [5]
- Data was stored in a public **S3 bucket**
- **Extensive** yet **adaptable** search engine by values like:
  - **Metadata**
  - **Genomic information:** genome size; GC content; no. of contigs
  - **Genomic features:** MLST; annotated features
  - **Taxonomic information:** taxonomic ranks

Repository statistics



Overview of the taxonomic composition at the family level
Nodes and branches are colored and sized by aggregated genome abundances at each taxonomic rank.

- **648,567 / 661,405 genomes** successfully characterized
- A total of **6.15 TB** of genomic information
- **24 most prevalent species** constitute for **90 %** of all genomes
- 66 phyla; 132 classes; 345 orders; 722 families; 2,466 genera; 8,207 species
- Genome size range: **100,943 bp - 20,285,777 bp; GC%: 23.6 % - 76.5 %**
- Examination of **key assembly metrics** for several species: genome size; #contigs; N50



Distribution of genomic metrics in the repository
The genome size (top left) and the GC content (bottom left) are displayed. In comparison, the mean genome size (top right) and the mean GC content (bottom right) per species are shown. The magenta-highlighted region illustrated the distribution excluding the most abundant species.

Distribution of the genome size for the most abundant species
White points indicate medians; bold black bars represent interquartile ranges; thin black lines represent outliers. Genomes were filtered for 95 % completeness and less than 1 % contamination.

Conclusion

The **BakRep project** conducts comprehensive and **standardized characterizations** of one of the largest collections of bacterial genomes comprising assembly metrics, robust taxonomic classifications, MLST, genome annotations, and original metadata. We envision BakRep as a **high-quality open resource** for microbial researchers worldwide.

[1] Blackwell G.A. *et al.* (2021); DOI: 10.1371/journal.pbio.3001421
[2] Chaumeil P.A. *et al.* (2019); DOI: 10.1093/bioinformatics/btz848
[3] Chklovski A. *et al.* (2023); DOI: 10.1038/s41592-023-01940-w
[4] https://github.com/tseemann/mlst
[5] Schwengers O. *et al.* (2021); DOI: 10.1099/mgen.0.000685

Linda Fenske (M. Sc.)
linda.fenske@computational.bio.uni-giessen.de