Gut sequence database

This document describes the custom database on human gut microbiome sequencing data

Access the database at: http://raeslab.org/companion/gutsequencedb/index.html

Gut genome sequences

The database contains sequenced and fully annotated reference genomes for the human gut bacteria from public sources. Currently this includes 823 genome sequences for microbial (mostly bacterial) organisms from the human gastrointestinal tract (Fig. 1).

Data access The following data sets on sequenced genomes are provided for download (from Human Microbiome Jumpstart Reference Strains Consortium):

- Metadata of the included genome sequences
- Genome sequence database in FASTA format
- Genome sequence database in Genbank format

Data description Metadata for 745 of the genome sequences (90.5%) is summarized in the Metadata file, including information of organism name, domain, gene count, the current sequencing quality, various IDs (HMP/GOLD/NCBI/Genbank/IMG/HDMD), sequencing center, and other information. The data set contains 737 bacterial genomes and 2 archaeal genomes. For further 6 genomes the domain information was not listed in the original data source but could be manually classified into the bacterial domain in all cases. The data collection includes both quality draft sequences (308; 41%), and completed genomes (437; 59%). For further details of genome annotation, see the Genbank file.

Details of data retrieval and database construction The reference sequences for annotated bacterial genomes were downloaded from HMP reference genomes database[1]. We filtered the results to include only bacterial genomes from the human gastrointestinal tract[2] (accessed March 18, 2018), yielding 823 genome sequences. This data collection was stored in FASTA (ASM) and Genbank formats.

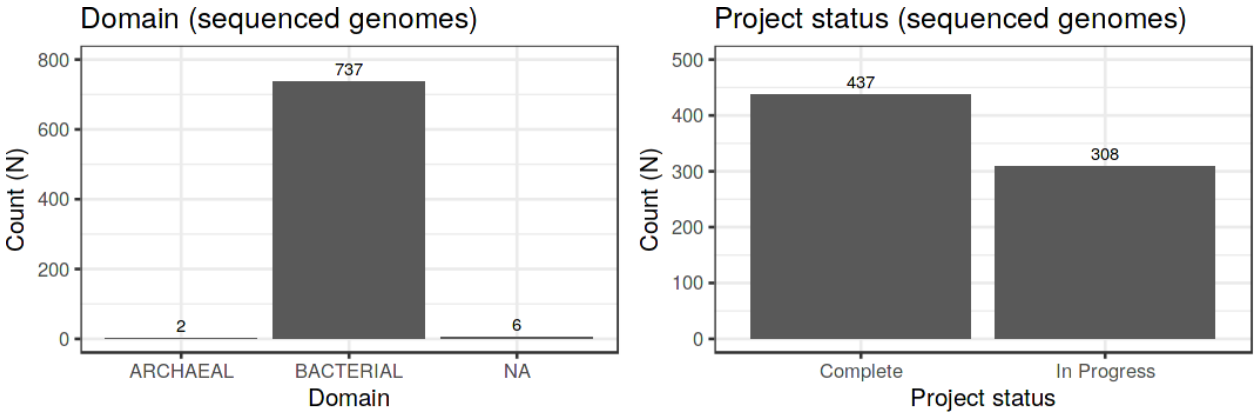

Fig. 1: Summary of the domain and project status for the sequenced genomes in the database.

Gut gene catalog

The gut microbiome gene catalog was retrieved from the Integrated reference catalog of the human gut microbiome (http://meta.genomics.cn/meta/dataTools - Li et al., 2014). This represents the state-of-the-art collection of gut microbiome gene sequences. The sequence data is based on 1267 intestinal samples, and

[1]https://www.hmpdacc.org/hmp/catalog/
[2]https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic&hmp_isolation_body_site=gastrointestinal_tract

includes 9,879,896 Open Reading Frames (ORFs). 21.3% of these sequences have been assigned to Phylum level taxonomic annotations. See the original publication for further technical details.

The following gut gene catalog files are available for download:

- Gene (nucleotide) sequences in FASTA format for the integrated non-redundant gut gene catalog

Database maintenance and updates

This custom database will be updated when new sequenced and fully annotated reference genomes from the human gut microbiota become available. Contact the project coordinator for further details (leo.lahti@iki.fi).

References

Gut genome sequence databases

Human Microbiome Jumpstart Reference Strains Consortium. A catalog of reference genomes from the human microbiome. Science 21 May 2010: 328(5981), pp. 994-97. URL: https://dx.doi.org/10.1126/science.1183605

Human Microbiome Project (HMP) Reference Genomes (https://www.ncbi.nlm.nih.gov/bioproject/28331)

Gut microbial gene catalogs

Li J *et al.* An integrated catalog of reference genes in the human gut microbiome. Nature Biotechnology volume 32, pages 834–841 (2014). URL: https://dx.doi.org/10.1038/nbt.2942

Qin J *et al.* A human gut microbial gene catalog established by metagenomic sequencing. Nature. 2010 Mar 4; 464(7285): 59–65. URL: https://dx.doi.org/10.1038%2Fnature08821