Grant Agreement Number: 686271

AD-gut: Alzheimer Disease - gut connection

D2.10 Database construction

RIA - Research and Innovation action

H2020-NMP-12-2015 - Biomaterials for treatment and prevention of Alzheimer's disease

Project start date: 01.04.2016 Duration: 48 months
Deliverable D2.10, due date: M24, 31.03.2018
Actual submission date: 23.03.2018
Dissemination level: CO
Workpackage: WP2 – Metagenomics
Workpackage leader: KU Leuven/KUL
Contact: Prof. Johan Hofkens

Deliverable Description

This deliverable is a custom database of human gut microbiome sequencing data. The database can be accessed at: http://raeslab.org/companion/gutsequencedb/index.html

Results

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).

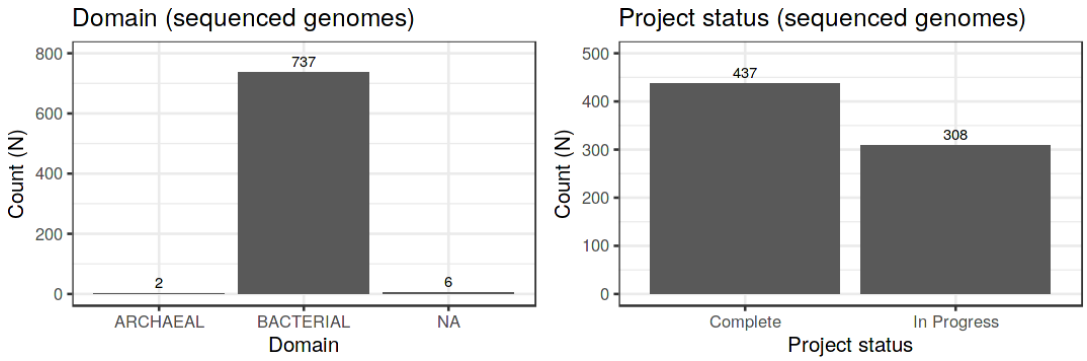

Fig. 1 Summary of the domain and project status for the sequenced genomes in the database.

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

- Metadata for 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.

Data retrieval and database construction The reference sequences for annotated bacterial genomes were downloaded from HMP reference genomes database (https://www.hmpdacc.org/hmp/catalog). We filtered the results to include only bacterial genomes from the human gastrointestinal tract (https://www.hmpdacc.org/hmp/catalog/grid.php?dataset=genomic&hmp_isolation_body_site=gastrointestinal_tract) (accessed March 18, 2018), yielding 823 genome sequences. This data collection was stored in FASTA (ASM) and Genbank formats.

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 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

Assessment of impact for AD-gut project

This database is used to facilitate efficient informatics for the sample analysis and allow matching of the DNA patterns identified in experimental samples in order to identify and quantify the microbial species that are present. The main impact to the AD-gut project is two-fold: 1) the database is used to construct *in silico* microbiome populations for experimental verification of the analysis pipeline; and 2) it is used for experimental analysis of species composition in new samples, and to assess the diagnostic potential.

Conclusion

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