THLBB2021_PUB004: Gut microbiome composition and fatty liver disease -dataset

Cohort: FINRISK 2002 and FINRISK 2007
Data set accession number in THL Biobank: THLBB2021_PUB004

Publication information

Links between gut microbiome composition and fatty liver disease in a large population sample
Matti O Ruuskanen, Fredrik Åberg, Ville Männistö, Aki S Havulinna, Guillaume Méric, Yang Liu, Rohit Loomba, Yoshiki Vázquez-Baeza, Anupriya Tripathi, Liisa M Valsta, Michael Inouye, Pekka Jousilahti, Veikko Salomaa, Mohit Jain, Rob Knight, Leo Lahti, Teemu J Niiranen

Correspondence: matti.ruuskanen@utu.fi, teemu.niiranen@utu.fi, leo.lahti@utu.fi

Published article's journal and reference information

Gut Microbes. 2021 Jan-Dec;13(1):1-22
DOI: https://doi.org/10.1080/19490976.2021.1888673
PMID: 33651661
PMCID: PMC7928040

Dataset availability information

The analysis code written for this study is included with the Supplementary Information. The datasets generated during and analyzed during the current study are not public but are available based on a written application to the THL Biobank as instructed in https://thl.fi/en/web/thl-biobank/for-researchers.

Please note that some of the phenotype variables were derived from national health registers and their use requires a separate data permit from Findata (https://findata.fi/en/data-permits/).

Participant and method description

https://www.tandfonline.com/doi/full/10.1080/19490976.2021.1888673?scroll=top&needAccess=true

Description of datasets

The THLBB2021_PUB004 datasets contain all the data needed to replicate the experiments of the article "Links between gut microbiome composition and fatty liver disease in a large population sample".

The following 9 source files from FINRISK 2002 and 2007 cohorts, two external database files, and the Anaconda environment file, can be used to replicate all analyses contained in the paper with the code included in the SI of the paper (https://www.tandfonline.com/doi/full/10.1080/19490976.2021.1888673?scroll=top&needAccess=true).

The analysis code is looking for these files in their current working directory, which is the root of this folder containing this README file and the phenotype data files.

In addition to these files, the analysis code creates several intermediary files, tables and figures in the working directory.

Some calculations can take a long time (up to a day) to complete the first time the analysis is run, but the intermediary files created can be used subsequently to run the code much faster.

Phenotype data file for FINRISK 2002:
1. "FR_02_phenotype_data.RData"
   - This file contains the R objects "FR02", which contains the data, and "FR02names", which contains the explanations of the following included variables: "Barcode", "BMI", "TRIG", "GGT", "VYOTARO", "MEN", "BL_AGE", "ALKI2_FR02", "PAINO", "SYSTM", "CURR_SMOKE", "PREVAL_DIAB", "PREVAL_CVD", and "ALUE".

Phenotype data file for FINRISK 2007 (used for model validation):
2. "FR07_phenotype_data.Rdata"
   - This file contains the R objects "FR07", which contains the data, and "FR07names", which contains the explanations of the following included variables: "STOOL_ID", "BMI", "TRIG", "GGT", "VYOTARO", "MEN", "BL_AGE", "ALKI2_FR07", "PAINO", "SYSTM", "CURR_SMOKE", "PREVAL_DIAB", "PREVAL_CVD", and "ALUE".

Phyloseq R objects containing GTDB ("gtdb") or NCBI ("ncbi") referenced bacterial ("bac") and archaeal ("arch") counts of the samples in FINRISK 2002 (no specifier) and FINRISK 2007 ("valid") samples:
3. "Liver_analysis_antibiotics/gtdb_bac_data.RDs"
4. "Liver_analysis_antibiotics/gtdb_arch_data.RDs"
5. "Liver_analysis_antibiotics/gtdb_bac_data_valid.RDs"
6. "Liver_analysis_antibiotics/gtdb_arch_data_valid.RDs"
7. "Liver_analysis_antibiotics/ncbi_bac_data.RDs"
8. "Liver_analysis_antibiotics/ncbi_arch_data.RDs"

A table of results produced in the MetaCyC pathway analysis of associated genomes:
9. "Liver_analysis_antibiotics/genome_pathway_inference_results.tsv"

External database files required to run the MetaCyC pathway analyses in the code (can be bypassed, as the result table is included above):
1. "Liver_analysis_antibiotics/assembly_summary_09032020.txt" (NCBI assembly records; downloaded version from 9th of March 2020 included)
2. "Liver_analysis_antibiotics/sp_clusters_r89.tsv" (GTDB species clusters; also available at https://data.gtdb.ecogenomic.org/releases/release89/89.0/sp_clusters_r89.tsv)

Anaconda environment file
1. "conda_environment.yml" (The specific environment which was used to run the analysis code can be created from this file in Anaconda with the command "conda env create -f conda_environment.yml")