# Protocol for whole shotgun metagenomics pipeline for the study of stool human digestive microbiota

**Authors**: Meslier V.[1*], Ren Y. [1*], Geiger M.[1*], Gilles M. [1], Famechon A. [1], David A. [1], Morabito C. [1], Quinquis B. [1], Almeida M. [1]

[1]Université Paris Saclay, INRAE MetaGenoPolis, 78350 Jouy-en-Josas, France

*DNA extraction and high throughput sequencing.* Frozen faecal materials were aliquoted to ≤ 1000µL with or without the addition of a liquefaction liquid (OMNIgene Liquefaction Reagent OM-LQR, DNAGenotek) to facilitate sample recovery. DNA extraction was performed following the procedure previously described in dx.doi.org/10.17504/protocols.io.dm6gpjm11gzp/v1, with the following modifications. The samples were transferred into a deep-well plate containing 400 µL of 0.1 mm glass beads (not in suspension) and centrifuged 3,486 ×g for twenty minutes prior to discarding the supernatant and adding 250 µL guanidium thiocyanate, 40 µL N-lauroyl sarcosine (10 % solution) and 500 µL N-lauroyl sarcosine (5 % solution in PBS 1X) to the samples. Subsequently, the sample plate was incubated at 70°C in a thermomixer for one hour, with stirring at 1,400 rpm. Following centrifugation of the plate at 3,486 ×g for a period of five minutes, the lysate was collected in a new plate. The pellet of the previous plate was then washed with 500 µL of TENP (50 mM Tris-HCL 20 mM EDTA 10 mM NaCl, saturated with PVPP). The plate was then vortexed and centrifuged at 3,486 ×g for five minutes, after which the recovered lysate was pooled with the previous one. Finally, the final lysate was centrifuged for a 10 minutes' period at 3,486 ×g, after which 800 µL were collected in a new plate. This plate was employed for purification with magnetic beads on the QIASymphony. The utilised protocol has been designed for MGP with the QIAGEN DSP Virus/Pathogen kit. DNA was quantified using Qubit Fluorometric Quantitation (ThermoFisher Scientific, Waltham, US) and qualified using DNA size profiling on a Fragment Analyzer (Agilent Technologies, Santa Clara, US). Either 500ng or 1µg of high molecular weight DNA (>10 kbp) was used to build the library. Shearing of DNA into fragments of approximately 150 bp was performed using an ultrasonicator (Covaris, Woburn, US) and DNA fragment library construction was performed using the Ion Plus Fragment Library and Ion Xpress Barcode Adapters Kits (ThermoFisher Scientific, Waltham, US). Purified and amplified DNA fragment libraries were sequenced using the Ion Proton Sequencer (ThermoFisher Scientific, Waltham, US), with a minimum of 20 million high-quality 150 bp reads generated per library (1).

*Read Mapping.* We performed QC check to remove any low-quality sequences using Alientrimmer software v2.0 (2) with the following parameters: "-k 10 -l 45 -m 5 -p 40 -q 20", and potential host-related reads with Bowtie2 v2.5.1 (3) and using the human reference genome Homo sapiens T2T-CHM13v2.0 (accession GCF_009914755.1) (4). High-quality reads were mapped onto the 10.4 million gut human gene reference catalogue v1 (https://doi.org/10.15454/FLANUP) (5) and the 8.4 million human oral microbial catalogue v1 (6) using the METEOR software v1 (7). Read mapping was performed in a two-step procedure, using an identity threshold of 95% to the reference gene catalogues. First, unique mapped reads were attributed to their corresponding genes. Second, shared reads were weighted according to the ratio of unique mapping counts. A downsizing procedure was performed to normalize gene counts between samples by randomly selecting a subset of reads depending on the sequencing depth (18 million reads for an average 24 million reads depth sequencing). The gene abundance table was then normalized using the FPKM strategy and analysed using MetaOMineR (*momr*) R package v1.31 (https://forgemia.inra.fr/metagenopolis/momr) (8). Finally, we performed a final QC validation on the MSP species composition using the CroCoDeEL methodology (https://github.com/metagenopolis/CroCoDeEL), combined with a spearman correlations threshold of >0.65, to remove any low quality samples, and yielding the final cohort to 464 individuals.

*MSP microbial species determination.* The 10.4 million gut gene and the 8.4 million oral gene catalogues were previously organized into 1990 and 853 MSP species (6,9–11), that correspond to clusters of co-abundant genes used as proxies for microbial species, and containing core and accessory genes. We removed duplicated species obtained from the catalogues, yielding to a total of 2055 MSPs, further filtered at a 10% occurrence threshold for a final MSPs species count at 627. Taxonomical annotation was assigned using the Genome Taxonomy Database GDTB R07-RS207 (12), using an in-house pipeline as described below. First, all genes were aligned on public databases (ncbi, wgs (13)) using Blast (14). MSP were annotated with the lowest taxonomical rank (from species to superkingdom) that brought consensus in at least 50% of its genes. To avoid misleading annotations due to error in databases, for each gene the 20 first hits were considered. MSP definition and taxonomy are available from Data INRAe (https://doi.org/10.15454/WQ4UTV and https://doi.org/10.15454/FLANUP). Relative abundance of a given MSP was computed as the mean abundance of its 100 'marker' genes (that is, the genes that correlate the most altogether). If less than 10% of 'marker' genes were seen in a sample, the abundance of the MSP was set to 0. MSP richness was computed as the number of detected MSP species in a particular sample, before proceeding to occurrence filtering.

*Microbial functional potentials determination.* To assess the functional potential of the gut microbiota at the module level, we used the METEOR software, which includes several functional database as described in Thirion et al. (15). Three databases were used to predict gene functions: Kyoto Encyclopedia of Genes and Genomes (KEGG) (16), eggNOG database (version 3.0) (17) and TIGRFAM (version 15.0) (18). First, genes of the catalogues were annotated using KEGG107 database using Diamond (19) and further clustered into functional pathway modules according to KEGG (Kyoto Encyclopaedia of Genes and Genomes) Orthology (KO) groups, Gut Metabolic Modules (GMM) (20), and Gut Brain Modules (GBM) (21), this later using the annotations from the three KEGG, enggNOG and TIGRFAm datasases. Second, KEGG, GMM and GBM modules were reconstructed in each MSP using their reaction pathways based on their detected annotated KO genes. GMM and GBM functional modules were selected because they are specific to gut bacterial and gut-brain axis functions. For each pair of MSP/individual, the completeness of any given functional modules was calculated by considering the set of genes detected in the MSP of each individual and the MSP completeness in each individual. For a given MSP in a specific individual, completeness of the modules was corrected by the abundance of the MSP. After correction, functional modules in each MSP/ individual were considered as complete if at least 90% of the involved reactions were detected. Abundance of functional modules in each sample was computed as the sum of the MSP abundances containing the complete functional module.

## References

1. Meslier V, Quinquis B, Da Silva K, Plaza Oñate F, Pons N, Roume H, et al. Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data. 2022;9(1):694.

2. Criscuolo A, Brisse S. AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics. 2013;102(5-6):500-6.

3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9(4):357-9.

4. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. avr 2022;376(6588):44-53.

5. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. Genome biology. juill 2017;18(1):142.

6. Le Chatelier E, Almeida M, Plaza Oñate F, Pons N, Gauthier F, Ghozlane A, et al. A catalog of genes and species of the human oral microbiota. [Internet]. 2021. Disponible sur: https://data.inrae.fr/citation?persistentId=doi:10.15454/WQ4UTV

7. Amine Ghozlane, Florence Thirion, Florian Plaza Oñate et al. Accurate profiling of microbial communities for shotgun metagenomic sequencing with Meteor2, 05 March 2025, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-6122276/v1]

8. Le Chatelier Emmanuelle, Prifti Eddi. Mining Metaomics Data In R. Retrived from https://forgemia.inra.fr/metagenopolis/momr.

9. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology. 2014;32(8):822-8.

10. Plaza Oñate F, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, et al. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. Bioinformatics. 1 mai 2019;35(9):1544-52.

11. Plaza Oñate F, Le Chatelier E. Metagenomic Species Pan-genomes (MSPs) of the human gastrointestinal microbiota. Portail Data INRAE Recherche Data Gouv. 2020;

12. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil PA, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research. 7 janv 2022;50(D1):D785-94.

13. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 8 janv 2019;47(D1):D23-8.

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. oct 1990;215(3):403-10.

15. Thirion F, Speyer H, Hansen TH, Nielsen T, Fan Y, Le Chatelier E, et al. Alteration of Gut Microbiome in Patients With Schizophrenia Indicates Links Between Bacterial Tyrosine Biosynthesis and Cognitive Dysfunction. Biological psychiatry global open science. avr 2023;3(2):283-91.

16. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 1 janv 2000;28(1):27-30.

17. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 4 janv 2016;44(D1):D286-93.

18. Haft DH. The TIGRFAMs database of protein families. Nucleic Acids Research. 1 janv 2003;31(1):371-3.

19. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. janv 2015;12(1):59-60.

20. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, et al. Species–function relationships shape ecological properties of the human gut microbiome. Nat Microbiol. 13 juin 2016;1(8):16088.

21. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential of the human gut microbiota in quality of life and depression. Nat Microbiol. 4 févr 2019;4(4):623-32.