Construction of a reference genome catalog to decipher shared strains along an agrifood chain with shotgun metagenomic data

Fiona Bottin, Sébastien Theil, Céline Delbès, Panagiotis Sapountzsis, Hélène Chiapello, Pierre Nicolas, Guillaume Kon Kam King, Anne-Laure Abraham, Solène Pety

Construction of a reference genome catalog to decipher shared strains along an agrifood chain with shotgun metagenomic data

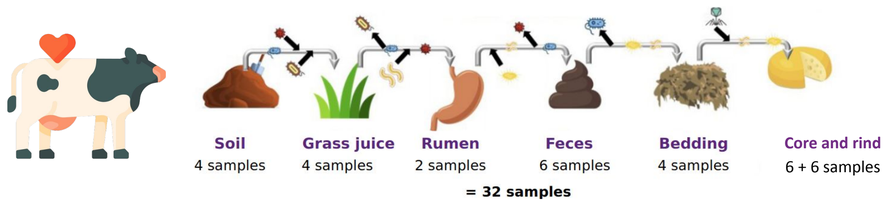
Solène PETY[1], Fiona BOTTIN[2], Sébastien THEIL[2], Céline DELBÈS[2], Panagiotis SAPOUNTZSIS[3], Hélène CHIAPELLO[1], Pierre NICOLAS[1], Guillaume KON KAM KING[*1] and Anne-Laure ABRAHAM[*1]

1. Context

Identifying **fluxes of micro-organisms** between successive compartments of an **agrifood chain** is important to understand and control cheese production. A first step to identify shared micro-organisms is to perform a **taxonomical assignation** at the species level. However, the **strain resolution** is very relevant for the precise analysis of the assembly process of microbiota across habitats. In order to **study strain fluxes**, and take into account **intra-species polymorphism**, we choose an approach based on **mapping metagenomic reads** using the BWA-MEM tool [1] on a **catalog of reference genomes** to **identify shared nucleotidic polymorphism across samples** in our various ecosystems. The use of reference genomes instead of metagenomic assembled genomes allows capturing polymorphisms for more species than only the most abundant, and enables comparison across multiple datasets using a common reference.

Construction of a genome reference database is a key part of our analysis framework and must be **tailored to the ecosystems** under study. We will present the **construction of a dedicated genome catalog** based on the **RefSeq** [2] database with the addition of relevant **genomes from different origins and projects** to complete our database: **metagenomic assembled genomes** (MAGS) from previous experiments, and microbial genomes isolated from cows' rumen and feces, and cheese. In particular, the species in the reference catalog must be different enough to avoid ambiguous mapping of the metagenomic reads. This requires aggregating similar genomes and **choosing a representative** for groups of aggregated genomes.

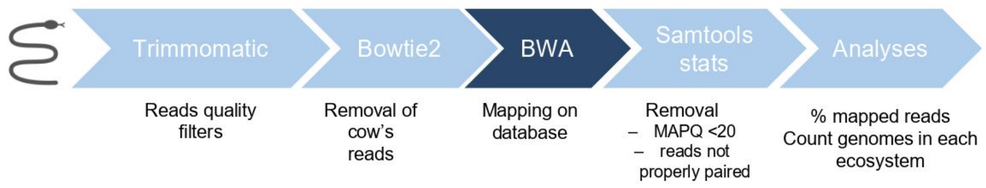
2. MINDS project and objectives



The MINDS project wants to study shared microorganisms and their diversity along 7 ecosystems of an agrifood chain for cows exposed to high or low diversity grassland. **Our work in MINDS project is to identify shared strains between samples. The first step is to build a complete and adapted reference genome database to identify species in samples.**

In order to **test the completeness of genome databases for our ecosystems,** we compare the percentage of mapped reads on each samples using the following 3 reference databases :

- MINDS MAGS database: MAGS assembled from MINDS samples
- Refseq database: procaryotic representative genomes (2019)
- Enriched database: MAGS from MINDS project + MAGS and genomes of 6 other sources



3. Completeness of 3 databases

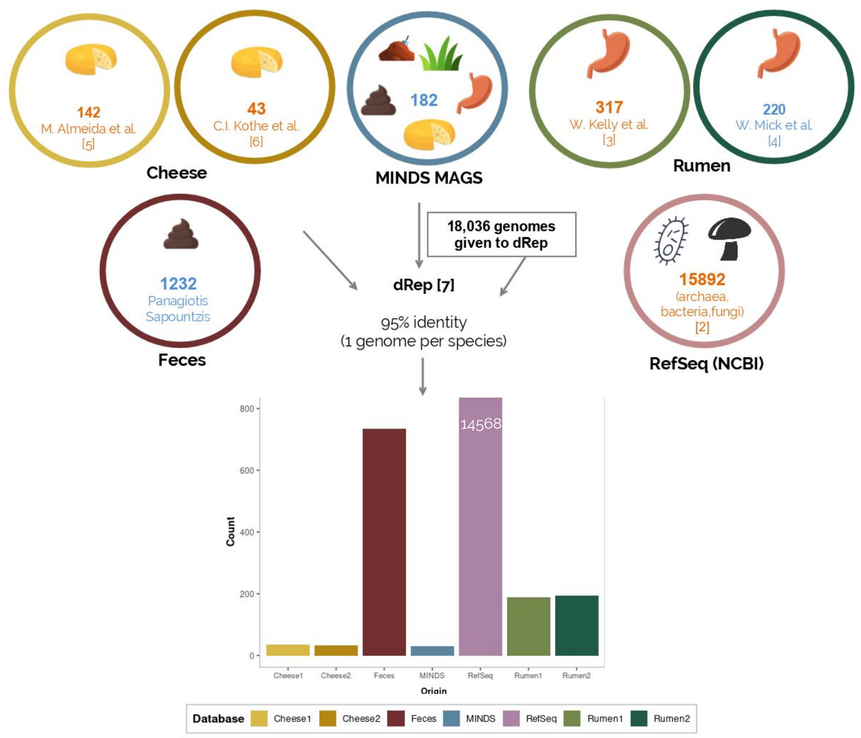
Percentage of reads mapped on MINDS MAGS or Refseq

Fig1. Percentage of reads properly mapped on samples with MINDS MAGS or Refseq.

MAGS (Metagenomic Assembled Genomes) were constructed with **MINDS samples** [cf. Poster T5.14]. We selected assembled genomes with >=50% completeness and <5% contamination for a total of 182 MAGS..
Refseq is a general database containing 14 155 representative prokaryotic genomes (2021). Reads of each sample were mapped separately on MINDS MAGS and Refseq using the snakemake workflow. About 50% of reads are properly mapped on MINDS MAGS for soil, grass, rumen, feces and bedding), and 80% of reads for cheese (core and rind). Only ~40% of reads were mapped on Refseq (except for rind).

Except for cheese, the **percentage of reads mapped is low** and shows that (1) despite being specific of our ecosystems, MINDS MAGS lacks low abundant species and some genomes are incomplete (2) Refseq lacks genomes specific to our ecosystems but contains many good quality genomes. **The idea is to combine several genomes and MAG databases to better cover our ecosystems.**

Building an enriched database



To enrich our database we gathered:
(1) genomes from RefSeq (archeae, bacteria, fungi)
(**2) MINDS MAGS**
(3) MAGS or genomes from other projects with our ecosystems.

Genomes with completeness <50% or contamination >5% were removed. Drep was used to dereplicate genomes with >= 95% identity (selection of one representative genome per species to reduce multi-mapping, same parameters used for UHGG database [9]).
Choice of representative was done with default parameters and takes into account genome quality and similarity with other genomes of the cluster.

Percentage of reads mapped on enriched database

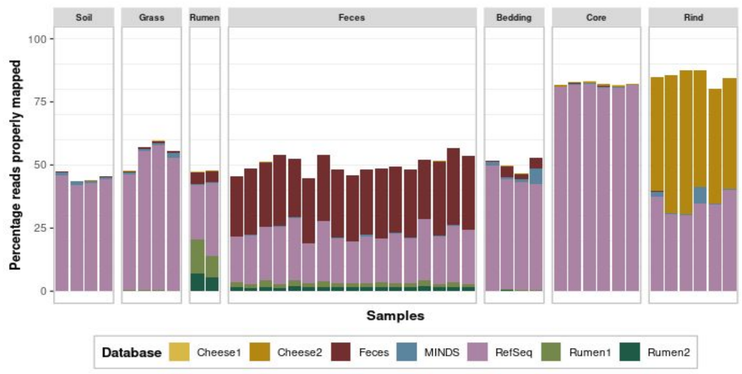


Fig2. Percentage of reads properly mapped on samples with enriched database.

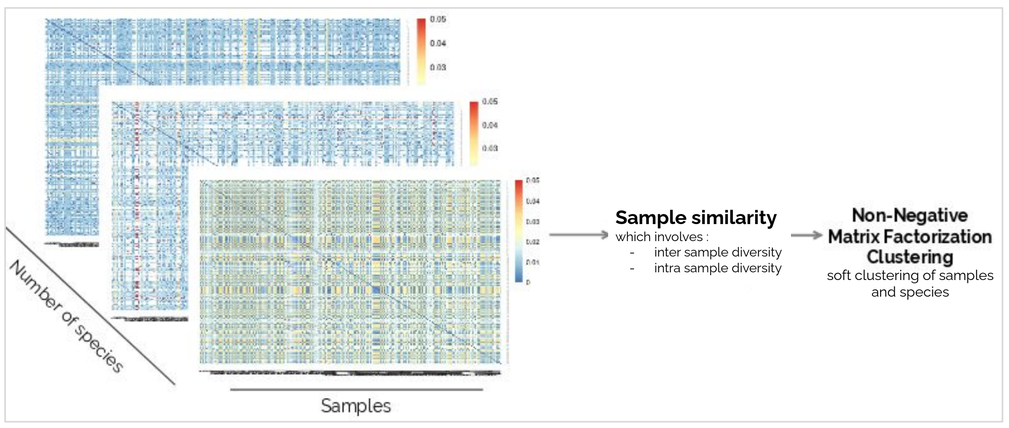
- Adding **several genomes sources allow to map more reads than** using **only Refseq.**
- Surprisingly, the **percentage of mapped reads is a little lower than obtained with MINDS MAGS only**, and, **MINDS MAGS genomes contribute weakly to these percentages**.
- Few representatives of MINDS MAGS were kept with default parameters, indicating that these genomes were either of lower quality, either less similar to other genomes of their clusters.
- Abundant species were assembled into MAGS, and are probably the same present in other genome sources. We probably miss a large number of low abundance species (lacking both in MINDS MAGS and in other genome sources).
- Mapping improving on cheese core is due to the addition of eukaryotes and fungi in Refseq.

Although some MIND MAGS could have a lower quality, they correspond to strains present in the samples. A better result could be obtained by **choosing preferentially a MINDS MAGS representative**, when available and if genome quality is good enough.

4.Conclusion & future work

After improvement, the enriched database will be used to map metagenomic reads on each representative genome, and identify **shared polymorphisms between samples**. For the most abundant species, we will attempt strain reconstruction using methods like DESMAN [9]. This is unrealistic for less abundant species, so we will also compute, for each species, diversity and similarity indices inspired from Nei's standard genetic distance [10] and adapted to metagenomic datasets.

Similarity between samples for each species will be jointly analysed using nonnegative matrix and tensor factorisation techniques.



Contact : solene.pety@inrae.fr

1 Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
2 Université Clermont-Auvergne, INRAE, VetAgro Sup, UMR545 Fromage, 15000, Aurillac, France
3 Université Clermont Auvergne, INRAE, UMR 454 MEDIS, Clermont-Ferrand, France

(*) co-last authors

Acknowledgments

We are grateful to :
the **Holoflux INRAE Metaprogram** for funding this project,

to the **MGX-Montpellier GenomiX platform** for the sequencing and,

to the **INRAE MIGALE bioinformatics facility** (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources.

Bibliography

1. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60, 2009.
2. Nuala A O'Leary, Mathew W Wright and *al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44(D1):D733-45, 2016.
3. Kelly Bill, Attwood G. T., Cookson A. L., et al. The Hungate 1000. A catalogue of reference genomes from the rumen microbiome. USDOE *Joint Genome Institute (JGI)*, Berkeley, CA (United States), 2011.
4. Watson, Mick. et al. Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism. [dataset]. *The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh*, 2017
5. Almeida, M., Hébert, A., Abraham, AL. *et al.* Construction of a dairy microbial genome catalog opens new perspectives for the metagenomic analysis of dairy fermented products. *BMC Genomics* 15, 1101 (2014)
6. Kothe, C.I., Bolotin A., Kraïem, B., Dridi B., Renault, P. Unraveling the world of halophilic and halotolerant bacteria in cheese by combining cultural, genomic and metagenomic approaches, *International Journal of Food Microbiology* 358, 2021
7. Olm, Matthew R et al. "dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication." *The ISME journal* vol. 11,12 (2017): 2864-2868. doi:10.1038/ismej.2017.126
8. Almeida, A., Nayfach, S., Boland, M. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 39, 105–114 (2021)
9. Christopher Quince, Tom O. Delmont, Sébastien Raguideau, *and al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol* 18-181, 2017.
10. Masatoshi Nei. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89, 583 –590, 1978.