A quantitative metagenomics analysis of the French cheese ecosystems

Anne-Laure ABRAHAM[1], Nicolas PONS[1], Sean KENNEDY[1], Antoine HERMET[2], Emmanuelle LE CHATELIER[1], Mathieu ALMEIDA[1], Jean-Michel BATTO[1], Benoît QUINQUIS[1], Nathalie GALLERON[1] and Pierre RENAULT[1]

[1] INSTITUT MICALIS, UMR1319 INRA, Domaine de Vilvert, 78352, Jouy-en-Josas, Cedex, France
[2] LUBEM - EA3882, ESMISAB, Technopôle de Brest Iroise, 29280 Plouzané, France
anne-laure.abraham@jouy.inra.fr, pierre.renault@jouy.inra.fr

The cheese ecosystem

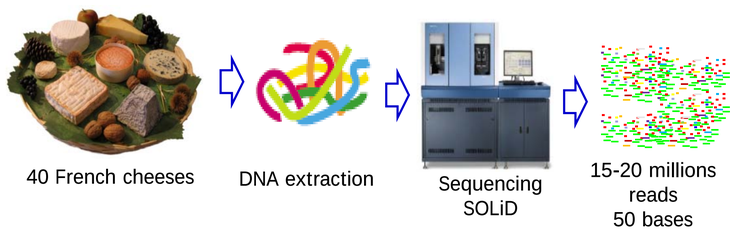
The manufacturing process of cheese, as for most fermented food, involves a complex flora, composed of bacteria but also yeast and filamentous fungi. The wide range of final products found on the dairy market is representative of the diversity of natural starters and ripening cultures used by dairy industries or coming from the food chain, from milk to the factory. However the cheese ecosystem is not completely understood. The natural starters are not constructed from pure strains and the knowledge of their exact composition remains incomplete, moreover number of species present in ripening cultures and in the food chain are little studied. Classical microbiological analysis or genetic methods (qPCR, MLST ...) can be used to identify and quantify to some extent these species. However these techniques are expensive and time consuming to provide a representative view of the flora.

Objectives:
Developing a metagenomic tool to achieve a rapid and accurate view of the cheese ecosystem in order to:

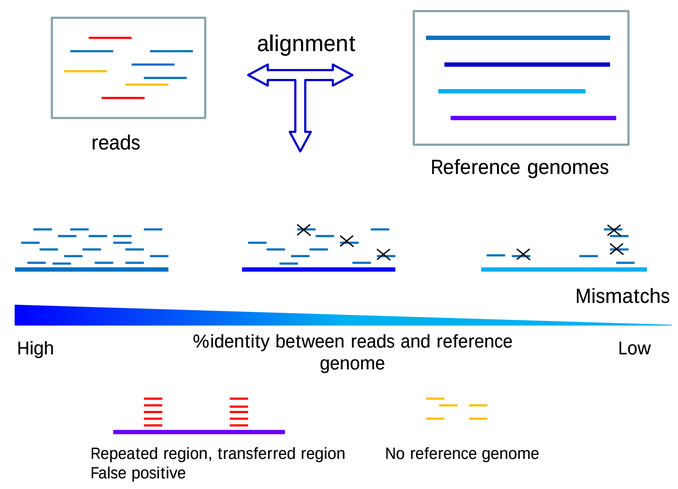
- Better understand the diversity of cheese flora by proposing an exhaustive and quantitative catalog of the present species.
- Provide a tool that gives a rapid and accurate vision of the species present in cheese in different production sites and over time in order to maintain a constant quality of the cheese product.

Metagenomics: a rapid and precise view of the ecosystem

The metagenomics technique allows the sequencing of the species contained in a sample without isolating the species and culturing them.



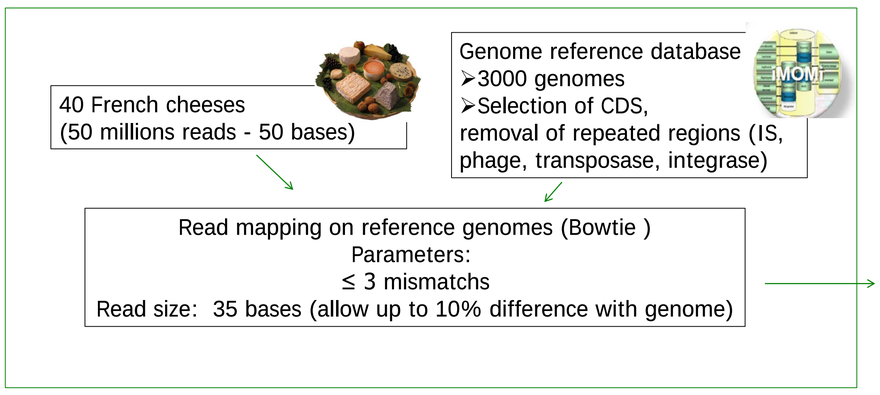
Challenges in species identification



Which species are present?
How many reads are necessary to identify a species?
Can we identify not only the species but also the strain?

Overview of the method

Mapping of reads



Observed coverage

Reads count:
% of CDS covered with ≥ 1 read
% nucleotides covered by ≥ 1read

Expected coverage

The expected coverage can be computed from the reads number, the genome size and the read length

$$C = 1 - \exp^{-\frac{ReadLength \times ReadNumber}{GenomeSize}}$$

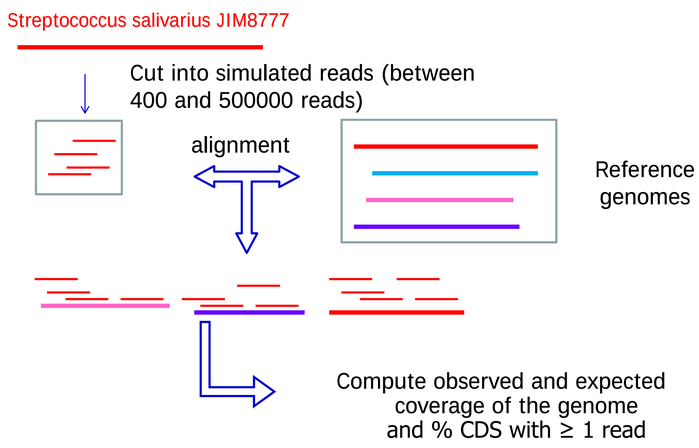
The expected number of CDS with ≥ 1 read can be computed from the reads number and the number of CDS in the genome

$$CDS = 1 - \exp^{-\frac{ReadNumber}{CDSNumber}}$$

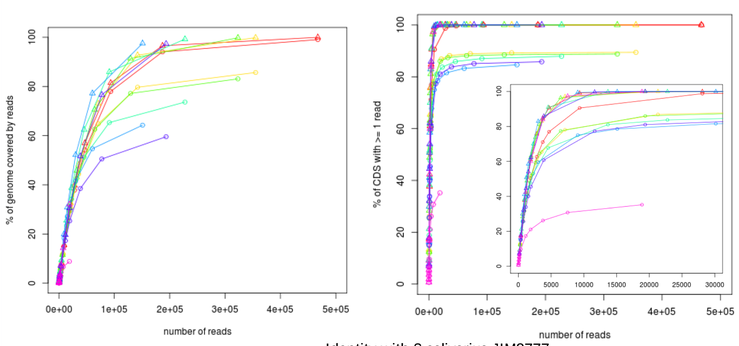
Comparison between observed and expected coverage for genome and/or CDS ⇒ Identification and quantification of species

Validation of the method

We have tested if we can identify the proper species from reads extracted from a strain when (i) its genome is in our database, and (ii) its genome is not, but genomes of more or less closely related species. For this purpose, we simulated reads from the genome of *Streptococcus salivarius* JIM8777 and mapped them independently on its genome and on 6 genomes of the same genus sharing between 80% and 95% identity with JIM8777.



Read count on simulated data



Identity with S salivarius JIM8777

Streptococcus_salivarius_JIM8777	100%
Streptococcus_salivarius_SK126	~ 95%
Streptococcus_salivarius_57_I	~ 95%
Streptococcus_vestibularis_ATCC_49124	~ 90%
Streptococcus_thermophilus_LMD.9	~ 90%
Streptococcus_sp__C150	~ 90%
Streptococcus_infantarius_subsp_infantarius	~ 85%

Observed ○
Expected Δ

For genomes with >100 000 reads, the genome coverage indicates:
- right genome (coverage >=80%)
- close species (coverage >= 50%)
- absent or very distant species (coverage < 50%)

For genomes with >1000 reads, the CDS coverage indicates:
- right genome (CDS coverage >=80%)
- close species (coverage >= 60%)
- Absent or very distant species (coverage < 60%)

The ecosystem of the Camembert cheese

We have applied our method on a sample of core Camembert made from raw milk. We selected species with more that 1000 reads. The species having >=80% of CDS with more than one read were considered as present, whereas species with between 60% and 80% of CDS were considered as close to species present in the ecosystem.

species	% CDS ≥ 1 read	nb reads
Lactococcus_lactis_subsp__lactis_Il1403	0,998708	4042722
Lactococcus_lactis_subsp__cremoris_SK11	0,998003	21872777
Lactobacillus_casei_ATCC_334	0,96716	184871
Lactobacillus_helveticus_R0052	0,916823	184263
Penicillium_camembertii	0,875772	40392
Brevibacterium_linens_BL2	0,875085	34199
Pediococcus_acidilactici_DSM_20284	0,873142	125403
Lactobacillus_plantarum_subsp__plantarum_ATCC_1491	0,799065	19802
Streptococcus_thermophilus_LMG_18311	0,742192	335410
Leuconostoc_pseudomesenteroides_KCTC_3652	0,727637	270789
Hafnia_alvei	0,727041	22286
Leuconostoc_mesenteroides_subsp__cremoris_ATCC_192	0,715755	301420
Leuconostoc_mesenteroides_subsp__mesenteroides_ATC	0,65586	48564
Arthrobacter_arilaitensis_Re117	0,654249	6909

Conclusion

- The method was tested on simulated data set and works quite well
- We can identity species having their genome in the database and species with a close species genome in the database. Furthermore, we can exclude misidentification of species absent of the ecosystem but whose genomes displays cross matching with genomes present in the database.

Perspectives

- Several improvements will be done:
  - Improving the threshold to identify present species
  - Taking into account mismatches to compute a distance between the species of the ecosystem and reference genome of the database
- The method will be validated with data from sequencing project with well studied ecosystems, before being applied to complex data

FoodMicrobiomes consortium