A platform for quantitative metagenomic profiling of complex ecosystems

Nicolas Pons[1],Jean-Michel Batto[1], Sean Kennedy[1], Mathieu Almeida[1], Fouad Boumezbeur[1], Bouziane Moumen[1], Pierre Léonard[1], Emmanuelle Le Chatelier[1], Sébastient Monot[2], Benjamin Rat[2], Tarik Saidani[2], S. Dusko Ehrlich[1] and Pierre Renault[1]

[1] Institut MICALIS, INRA CRJ, Domaine de Vilvert, 78352 Jouy-en-Josas, France
[2] AS+, 22, rue René COCHE, 92170 Vanves, France

Contact : nicolas.pons@jouy.inra.fr

Quantitative Metagenomic

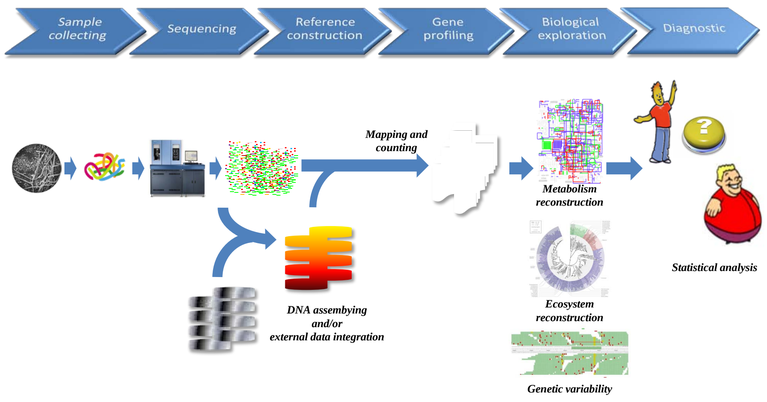


Figure 1. Schematic of quantitative metagenomic analysis using high-throughput sequencing. The study of complex microbial ecosystems by a quantitative metagenomic approach has been made possible by advancements in high-throughput sequencing technologies. Quantitative metagenomics relies on deep sequencing to construct an ecosystem profile using gene and genome counts. Next generation sequencing (NGS) technologies such as SOLiD or Illumina produce millions of short sequences (35 to 75bp) which can be used as tags to establish gene profiles. This approach requires the use of a specific reference catalog which should be composed of genes present in the ecosystem of interest. The use of classical bioinformatic methods for the analysis of such large amounts of data is not feasible as we overpass the expected dataset size of common tools. We have therefore developed an integrated metagenomic analysis pipeline called METEOR (METagenome ExploratOR).

METEOR platform design

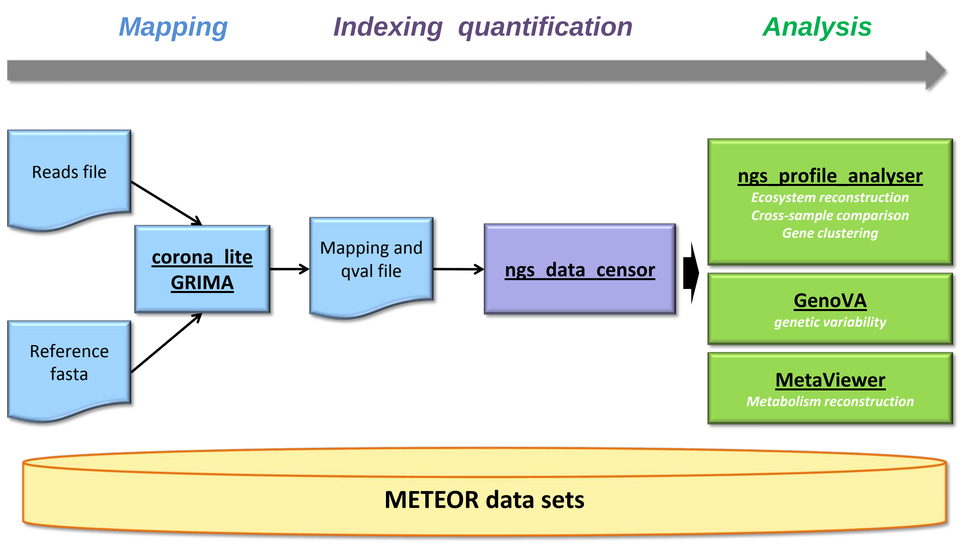


Figure 2. METEOR pipeline. Quantitative analysis of metagenomic samples by METEOR is divided into 4 distinct steps: **Mapping** is the process of identifying individual SOLiD reads by locating a match in the reference sequence. **Indexing** consist to store the position of each match in ISAM files. **Quantification** refers to the identification and enumeration of genes in the reference. **Analysis** step is any downstream use of the data including cross-sample comparisons, metabolic reconstruction, reference gene clustering, gene/species diversity and SNPs.

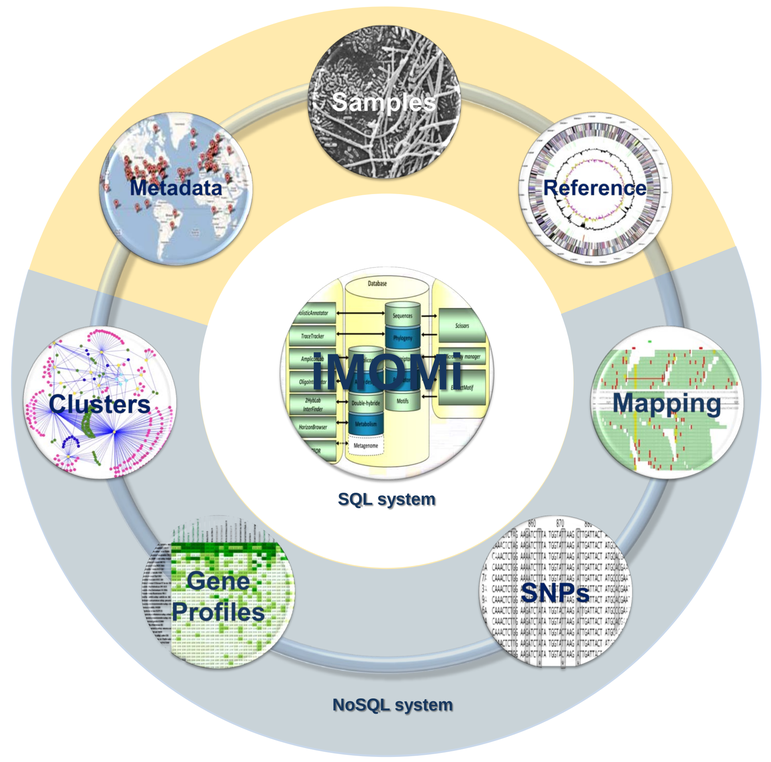
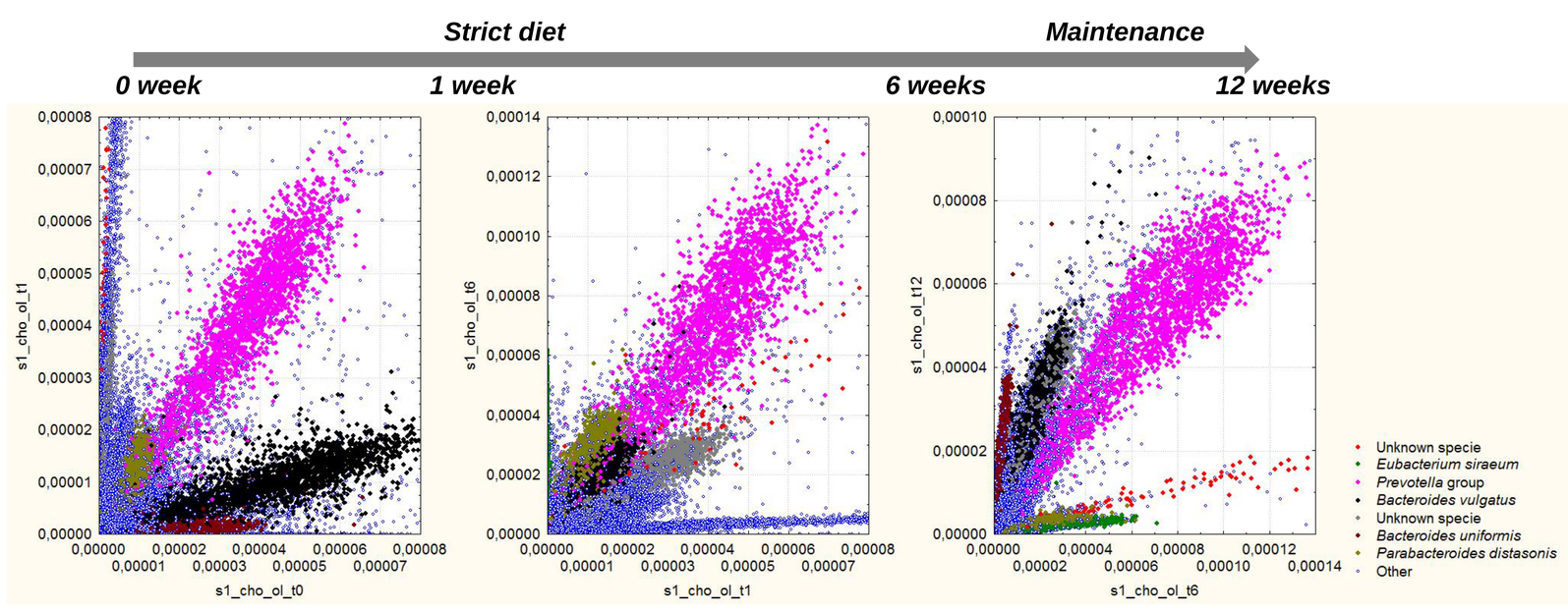


Figure 3. METEOR data sets. Imported (in orange) and experimental (in blue) data are indexed in an embedded NoSQL database (AdvantageDB, Sybase) around the iMOMi framework and organized in a dedicated file system. This optimization facilitates different data processing.

Micro-Obes – an example of application

Figure 4. Micro-Obes study of obese individuals. In Micro-Obes, we investigate the changes of gut microbiota in a human model of weight loss induced by restrictive diet in moderately obese subjects. DNA isolated from 215 faecal samples of 49 obese subjects collected at different date (start of the study, 1 and 6 weeks after a restrictive diet and 12 weeks) have been sequenced with SOLiD technology yielding about 300 gigabases. Short reads have been indexed against the 3.3 millions genes of the human gut microbial gene catalog (Qin et al, Nature 2010, MetaHIT consortium). Statistical analysis of the gene profiles generated indicate significant variations in gene and genome frequencies during the first 6 weeks of dieting (particularly after 1 week) and a subsequent stabilization after 12 weeks according to the observed success of patients dietary. By integrating these results with clinical data from the same group of individuals, we hope to learn more about how the intestinal microbiome interacts with the host and changes during dieting. In the future, we will be able to tailor both diet and nutrition to aid obese individuals lose weight and maintain a healthy BMI.



CONCLUSION

METEOR is used in many metagenomic project supported by the MetaQuant platform in INRA (Jouy-en-Josas) such as Micro-Obes for characterizing the human intestinal microbiome of obese individuals following a restrictive diet, Food-Microbiomes for studying the ecosystem of fermented food like French traditional cheeses, Metaflora for comparing the flora of different starters and BB-Allergy for understanding the different between intestinal microbiome of healthy and allergic infants. With the sequencing of a mix of several strains of a same specie, the pipeline is also used for studying the genetic variability of *Streptococcus thermophilus* and *Streptococcus salivarius*. More than 300 samples have been sequenced yielding about 500 Gigabases (more than 10 To of space-disk with processed data).

Future improvements of high-throughput sequencers require to develop more efficient algorithms or methods for processing a larger amount of samples in reasonnable time. In this context, we are associated to the project called OpenGPU. With AS+ company, our task consists to adapt and improve the actual pipeline onto massive parallel computing platform (for example, hybrid platform including GPGPU).