M2 Research internship:
Deep dimensionality reduction to unveil health status of humans through metagenomic gut data

Biological context: A typical human intestinal microbiome contains several kilograms of bacteria and around 100 times more genes than the human genome [5]. It has co-evolved with us for all our history and, in recent years, the **intimate connection** between our **gut flora** and **our health** has emerged as a central theme in medicine.

Goal: We propose here to investigate associations at the functional level by **discovering communities** of bacteria using **deep dimensionality reduction** methods. We will further mine **associations between** those **communities** and the **health status** of humans.

Method: Metagenomics allows to study microbial environments by directly sequencing the genomic DNA without the need for prior cultivation in the lab. The association between the genetic information of the bacterial community and the health status of a person is known as MWAS (Metagenome-wide association study). The usual method associates the abundance/depletion of a bacteria to the presence of a disease. We propose here to investigate the association at a higher level: how the co-presence of certain bacteria can impact or witness the health status of an individual? During this internship, you will only deal with N-dimensional vectors representing abundances of bacteria in a human gut sample. The human gut **data is readily available** through the very complete bioconductor package [4] providing gut species abundances together with their sample metadata (e.g. health status).

Looking for communities in abundance data:

This novel approach requires the use of mixture models or deep variational autoencoders to reduce the high dimensionality of the matrix of strain abundance vs. samples to a lower dimensional space interpreted as *communities* of associated bacteria. We plan first to investigate simple **linear admixtures** (using techniques like Sparse Non-negative Matrix Factorization) as they are easy to implement and to interpret: the participation of a bacterial species to a module is defined as its mixture coefficient. The next step will be to use a non linear variant such as **autoencoders (possibly deep)**. The interpretation of their weights is however not straightforward, and *in sillico* sampling of the input space is necessary to infer the communities associated to a module (technically, to the activation of a given neuron of the middle layer of the autoencoder).

Looking for associations between modules and phenotypes:

The modules found in the previous steps generalize the idea of enterotypes[2] by allowing both a module to be composed of several bacteria and a bacterial species to be involved in several modules. This allows us to investigate more deeply the potential associations between enrichment or depletion of a module with the phenotype of the host. The association method can be developed through an international collaboration with the Max Planck Institute in Göttingen as they have both expertise in metagenomic studies [6,7] as well as association studies in GWAS [8].

Dates:
At any time from now on. Don't hesitate to contact us for discussing your availability.

Profile:
No prior knowledge about metagenomics is required.
- Computer scientist, statistician, or bioinformatician background
- Programming experience (e.g. Python) and scripting (e.g. bash)
- Interest for data science, statistics and/or machine learning
- Interest for impact on biological/medical applications

Salary:
- Between **546€/month** and **800€/month** depending on funding.

Contact:
Clovis Galiez LJK, SVH team
clovis.galiez@grenoble-inp.fr

References

- [1] Nielsen et al, Nature Biotechnology 2014, https://doi.org/10.1038/nbt.2939
- [2] Arumugam et al, Nature 2011, https://doi.org/10.1038/nature09944
- [3] Knights et al, Cell Hist Microb. 2014, https://doi.org/10.1016/j.chom.2014.09.013
- [4] Pasolli et al, Nature Methods 2017, https://doi.org/10.1038/nmeth.4468
- [5] Gilbert et al, Nature med. 2018, https://doi.org/10.1038/nm.4517
- [6] Steinegger et al, Nature Biotech 2017, https://doi.org/10.1038/nbt.3988
- [7] https://github.com/soedinglab/mmseqs2
- [8] Banerjee et al, BioRxiv 2018 https://doi.org/10.1101/198911