Towards a machine learning approach for automated detection of well-to-well contamination in metagenomics data

Lindsay Goulet, Florian Plaza Oñate, Edi Prifti, Eugeni Belda, Emmanuelle Le Chatelier, Guillaume Gautreau

CONGRÈS JUNIOR PLURIDISCIPLINAIRE

GRADUATE SCHOOL Métiers de la Recherche et de l'Enseignement Supérieur — université PARIS-SACLAY

Towards a machine learning approach for automated detection of well-to-well contamination in metagenomics data

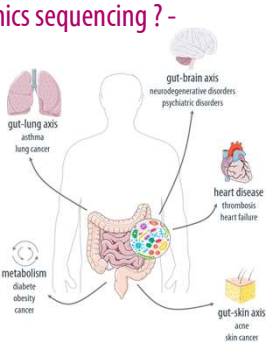
Lindsay Goulet[1], Florian Plaza Oñate[1], Edi Prifti[2,3], Eugeni Belda[2,3], Emmanuelle Le Chatelier[1] and Guillaume Gautreau[1]

Background

- What is metagenomics sequencing ? -

Metagenomics sequencing allows characterization of microbial communities, such as the human intestinal microbiota, without prior organism isolation or culture, by determining the nucleotide composition of randomly selected DNA fragments [1].

The gut microbiota plays a crucial role in human health and is closely associated with various diseases [2].



The intestinal microbiota is involved in the functioning of many organs. Dysbiosis can therefore affect these organs and allow the progression of related diseases.

Le French Gut project, with the sequencing of 100 000 fecal samples, aims to define the heterogeneity of healthy gut microbiomes, the environmental and lifestyle factors impacting them, and their deviations seen in chronic diseases.

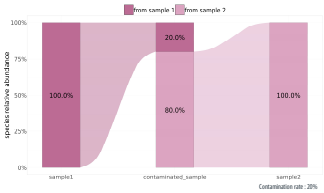
- What is contamination ? -

Contamination refers to the presence of DNA that does not originate from the biological sample under study. It can be due either to :

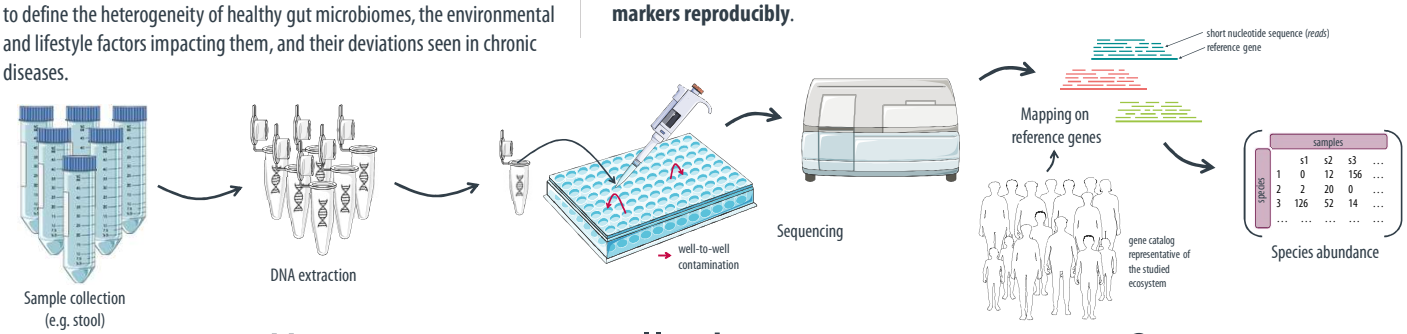
- DNA from an external source (environmental DNA [3] or lab reagents)
- DNA from **another sample** processed on the same plate (**well-to-well contamination**).

Well-to-well contamination occurs during wet lab steps (DNA extraction, sequencing).

After contamination, most abundant species in the source sample can be detected in the contaminated sample, depending on the contamination rate and sequencing depth.



Although well-to-well contamination is a common problem, it remains understudied. It can lead to **biased results** and eventually to **false conclusions** if not detected and it is also a serious impediment to **identify markers reproducibly**.



How to automatically detect contamination ?

- Specific patterns are associated with contamination in species abundance profiles -

Comparison of species abundance profiles reveals **specific patterns** associated with well-to-well contamination.

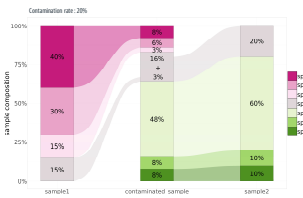
Species not present initially (0) will appear on the abundance profile as a diagonal (0) and the contamination rate can be estimated from the value of the intercept.

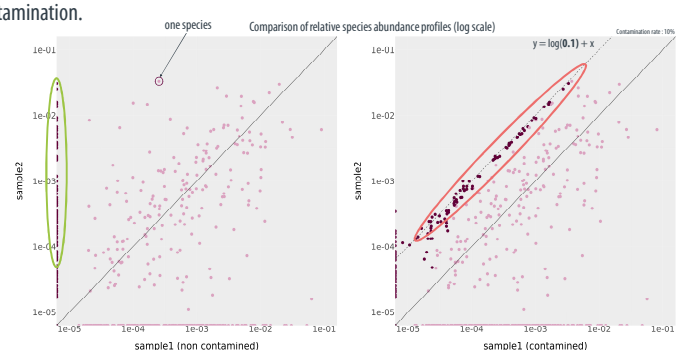
For each species, if sample1 is contaminated by sample2 :
$A = A_1 \times (1\text{-}c) + A_2 \times c$

If the species were not present initially in sample1 :
$A = A_2 \times c$

In log scale : $\log A = \log A_2 + \log c$

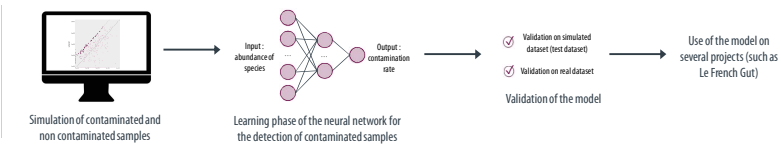
Where :
A_x : abundance of the species in the sample x
A : abundance of the species in the contaminated sample
c : contamination rate





- Towards a machine learning approach -

Human eye is very good at identifying contamination profiles. In contrast, the semi-automated procedure developed by our lab to select suspect cases for manual inspection among thousands of candidates (10 000 pairs in a 100 samples cohort) **lacks sensitivity** and is **time-consuming**. Thus, it is not adapted to the massive treatment of large cohorts, as in Le French Gut project.



[1]Université Paris-Saclay, INRAE, MGP MetaGenoPolis, Jouy-en-Josas, 78350, France
[2]Unité de Modélisation Mathématique et Informatique des Systèmes Complexes, UMMISCO, Sorbonne Université, IRD, Bondy, France
[3]Nutrition and Obesities : Systemic Approaches, NutriOmics, Research Unit, Sorbonne Université, INSERM, Paris, France

[1] Qunice C. *et al.* Nature Biotechology (2017). doi: 10.1038/nbt.3935
[2] Qui J. *et al.* Nature (2012). doi: 10.1038/nature11450
[3] Mzava O. *et al.* Nature Communications (2022). doi: 10.1038/s41467-022-31654-0

contact : lindsay.goulet@inrae.fr