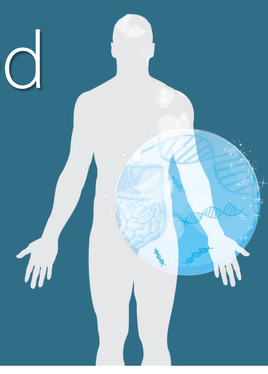
Towards a machine learning approach for automated detection of well-to-well contamination in metagenomic data

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

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



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

Introduction

Background

The gut microbiota plays a crucial role in human health [1]. Metagenomic sequencing allows a deep characterization of microbial communities without prior organism isolation or culture.

Several massive sequencing projects are now on the launchpad as Le French Gut which aims to analyze 100 000 fecal samples to define the heterogeneity of healthy gut microbiota, the environmental and lifestyle factors impacting them, and their deviations seen in chronic diseases.

In this context, the detection of well-to-well contaminations is a crucial but time-consuming task. Checks must therefore be AI-assisted to ensure data quality at scale.

Well-to-well contamination

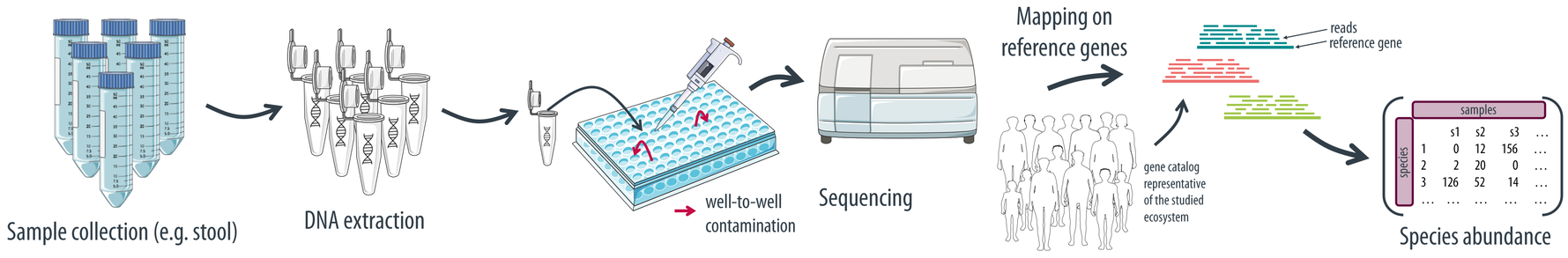
Contamination refers to the presence of DNA that does not originate from the biological sample under study. It can be due either to :

- DNA from an external source (environmental DNA [2] or lab reagents)
- DNA from another sample processed on the same plate (well-to-well /cross-sample contamination).

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

Although well-to-well contamination is a common problem, it remains understudied. It can lead to biased results (i.e. overestimation of α diversity, false strain sharing events) and eventually to false conclusions if not detected and it is also a serious impediment to studies reproducibility.

We introduce CroCoDDeeL, a deep-learning tool to automatically detect well-to-well contamination. Contrary to state-of-the-art approaches [3][4], CroCoDDeeL works with related samples that may naturally share strains (*e.g.*: mother/child), discriminates contamination sources from contaminated samples, estimates contamination rates and does not require costly negative controls.



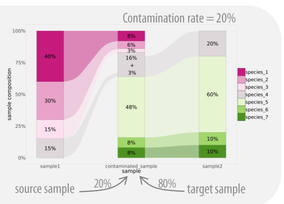
Methods

In-house correlation-based method

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

1.

Above a certain threshold, all the abundant species of the contamination source sample are present in the contaminated sample.

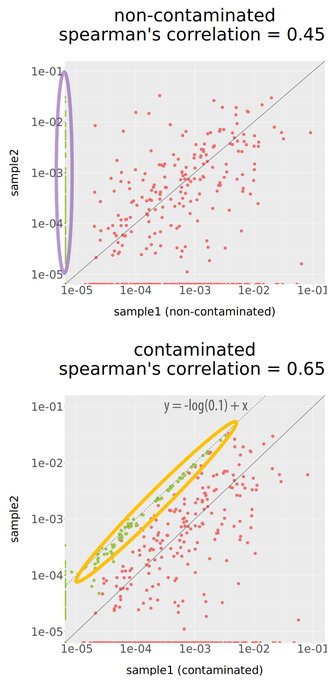


2.

A subset of these shared species have a proportional abundance between the two samples and form a contamination line (0).
This line corresponds to species that were not present before the contamination event (0).

3.

The contamination line is used to detect contamination events, and the contamination rate can be estimated from the value of the intercept.



- This in-house procedure requires a labor-intensive inspection to spot the contamination line.
- Thus, we restrict our inspection to the highest correlation given the combinatory explosion of all-vs-all comparisons.
- But correlation has pitfalls and is unsuitable for certain kinds of cohorts (diverse body sites, longitudinal samples).

⇒ Intelligent assistances are required to limit inspection time.

Deep Learning method : CroCoDDeeL

Cross-sample **Co**ntamination **D**etection using **Dee**p **L**earning

- Input data -

Species abundance profiles of all sample pairs are compared : matrix of pairwise abundance of 128 species :

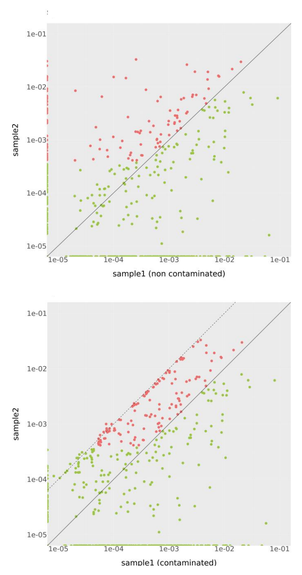
- selection only in the upper triangle: allows the detection of the contamination direction (which is the source, which is the target).
- selection of the 128 most abundant species in the source sample: in contaminated cases, allows the selection of the species forming the contamination line.

Keras model
7 dense layers
1 Batch Normalization layer
1 LeakyReLu layer
1 DropOut layer

Batch size = 128
Number of epochs = 15

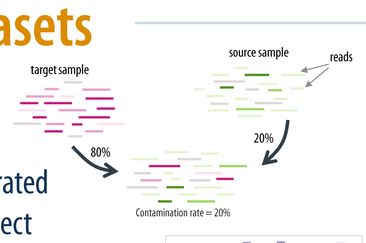
- Output -

Class prediction (contaminated/non-contaminated)



CroCoDDeeL's results

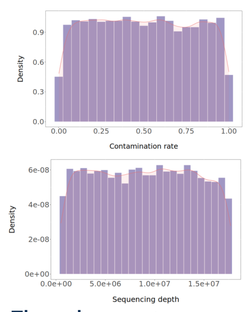
Learning datasets



– Contaminated dataset –

A semi-simulated training dataset was generated by mixing reads from manually-curated real metagenomic data of the MetaCardis project (PRJEB38742).

- Contamination rate : from 0.1% to 100%, following a uniform distribution.
- Sequencing depth : from 1M to 20M reads, following a uniform distribution.



– Non contaminated dataset –

Part of the data uses sample pairs of the MetaCardis project for which we simulated varying sequencing depth. The other part uses pairs of samples from distinct cohorts (human and animal gut) that cannot be contaminated because they have never been processed together.

Split data

Training data – 80%	Test data – 20%
16K contaminated samples	4K contaminated samples
16K non-contaminated samples	4K non-contaminated samples

Performance

Performance was evaluated on several real metagenomic datasets.

– Low correlated samples –

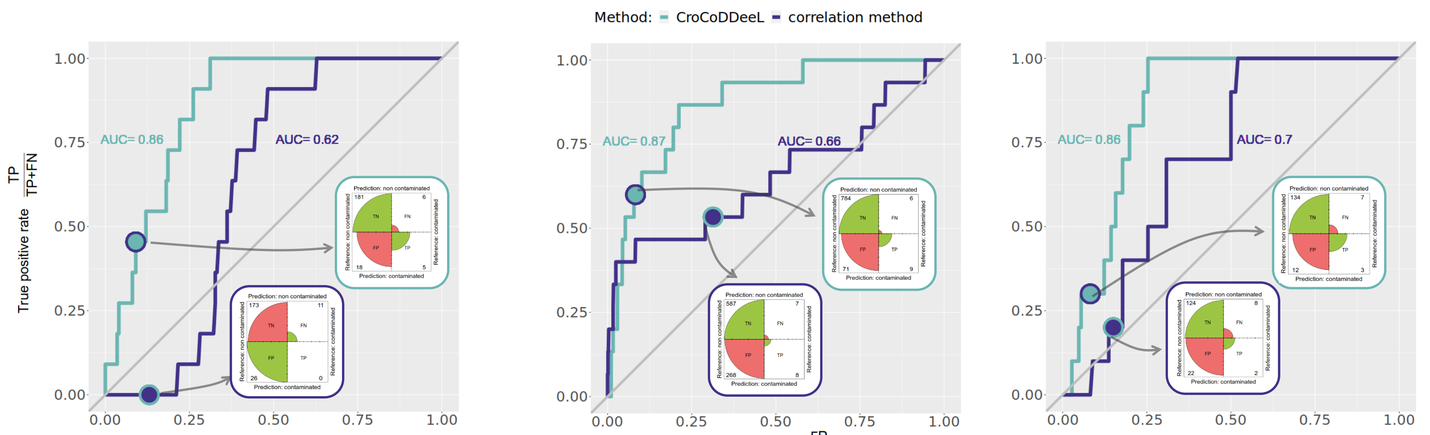
- 15 samples from the Human Microbiome Project (PRJNA48479)
- human gut and oral microbiota

– Medium correlated samples –

- 30 samples from the COVIDbiome Project (PRJNA792726)
- human gut microbiota

– Highly correlated samples –

- 13 samples from the Metachick projects
- chicken gut microbiota



Since the aim of CroCoDDeel is to detect as many contaminated pairs as possible, we are looking for the compromise that maximizes the number of true positives, without giving too many false positives to manually inspect.

Conclusion and future work

CroCoDDeeL outperforms our in-house correlation-based method in terms of sensitivity. However, it still produces many false positives without a contamination line, suggesting that the model uses its own criteria beyond the ones used by humans. We anticipate opportunities for improvement through an enhanced training dataset.

Next efforts will extend the use of this modeling approach in the way of identifying contamination levels and working towards the automatic decontamination of samples.

- As this time, we recommend CroCoDDeeL as an aid for human inspection, rather than a fully autonomous approach.
- CroCoDDeeL can deal with any metagenomic sequencing data without the need for negative or spike-in controls.
- CroCoDDeeL can process species abundance tables generated by any taxonomic profiler.
- CroCoDDeeL will be used to perform quality control of public cohorts and results will be integrated into the MIASSM database [5] (poster B-167).

Centre de recherche de Jouy-en-Josas
Unité MetaGenoPolis
www.mgps.eu

Contact: lindsay.goulet@inrae.fr
guillaume.gautreau@inrae.fr

[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

GitLab: https://forgemia.inra.fr/metagenopolis/CroCoDDeeL

[1] Qin J. *et al*. Nature (2012). doi: 10.1038/nature11450
[2] Mzava O. *et al*. Nature Comm. (2022). doi: 10.1038/s41467-022-31654-0
[3] Minich J. J. *et al*. ASM Journals (2019). doi: 10.1128/mSystems.00186-19
[4] Lou Y. C. *et al*. Microbiome (2023). doi: 10.1186/s40168-023-01477-2
[5] Barbet P. *et al*. Poster ECCB/ISMB (2023).