CroCoDeEL : accurate detection of cross-sample contamination in metagenomic data

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

CroCoDeEL: accurate detection of cross-sample contamination in metagenomic data

Lindsay Goulet[1], Florian Plaza Oñate[1], Pauline Barbet [1], Edi Prifti[2,3], Eugeni Belda[2,3], Emmanuelle Le Chatelier[1] and Guillaume Gautreau[1,4]

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 cross-sample contaminations is a crucial but time-consuming task. Checks must therefore be AI-assisted to ensure data quality at scale.

Cross-sample contamination

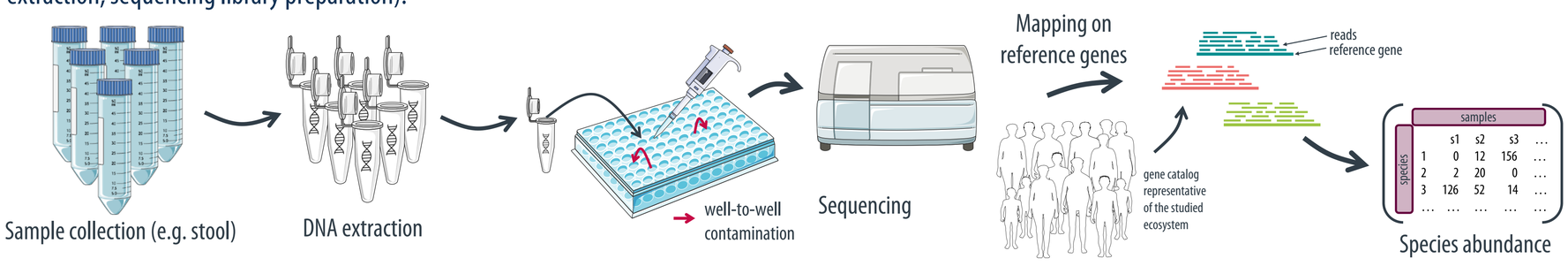
Contamination refers to the presence of DNA that does not originate from the biological sample under study. It can be due either to :

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

Cross-sample contamination occurs during wet lab steps (DNA extraction, sequencing library preparation).

Although cross-sample 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.

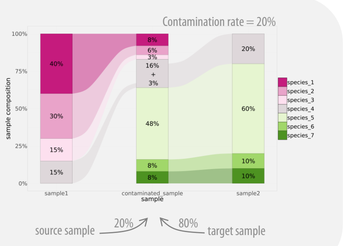
We introduce CroCoDeEL, a tool based on a supervised pre-trained model to automatically detect cross-sample contamination. Contrary to state-of-the-art approaches [3][4], CroCoDeEL works with related samples that may naturally share strains (e.g.: mother/child), discriminates contamination sources from contaminated samples, estimates contamination rates and does not require costly negative controls.



Methods

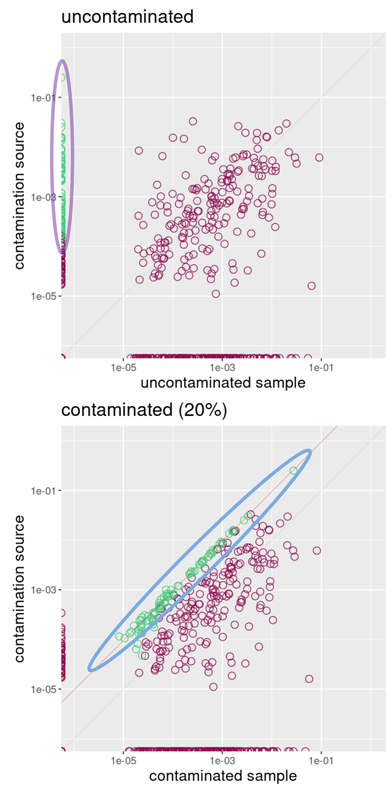
How to detect cross-sample contamination ?

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 relative abundance ratio of species that constitute it.



CroCoDeEL

Cross-sample **Co**ntamination **De**tection and **E**stimation of its **L**evel

- Input -

samples

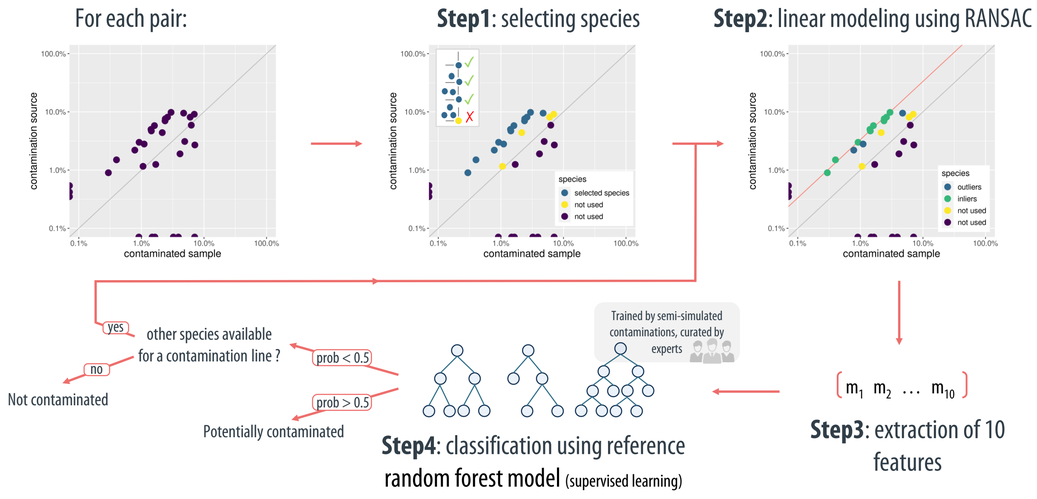
species	s1	s2	s3	…
1	0	12	156	…
2	2	20	0	…
3	126	52	14	…
…	…	…	…	…

Species abundance

- Output -

source	target	rate	score
s1	s2	1%	0,8
s2	s3	2,5%	0,9
…	…	…	…

List of suspicious samples

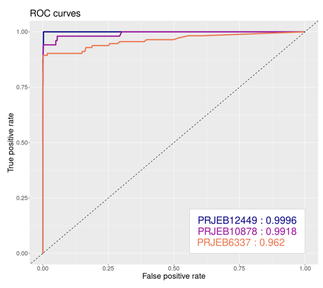


Results

Performance

Performance was evaluated on 3 public metagenomics cohorts, curated by human experts :

		PRJEB12449	PRJEB10878	PRJEB6337
# of comparisons (# of plot×2)		11990	16256	55932
False Positive (FP)		39	38	101
False Negative (FN)		0	3	12
Accuracy		99.70%	99.70%	99.80%
Area Under the ROC Curve (AUC)		0.9996	0.9918	0.962
Runtime (seconds)	8 CPU	133	184	681
	16 CPU	75	103	370
RAM consumption (MB)	8 CPU	205	222	213
	16 CPU	206	206	214



Application

CroCoDeEL has been applied to fifteen public cohorts on colorectal cancer. Here are some non-exhaustive results :

project	paper	# samples	seq. depth (M reads) (mean ± sd)	# contaminated samples	contamination rate (Q1 | med | Q3)		
PRJDB4176	YashidaS_2019	645	21.9 ± 6.3	5 (0.8%)	0.05%	0.08%	0.52%
PRJNA389927	HanniganGD_2018	84	2.8 ± 1.7	4 (4.8%)	5.57%	33.72%	64.94%
PRJEB27928	WirbelJ_2019	82	18.2 ± 8.1	39 (58.5%)	0.25%	0.35%	0.54%
PRJNA531273	GuptaA_2019	30	4.6 ± 2.1	17 (56.7%)	13.40%	36.33%	62.57%

Conclusion and future work

- We systematically found cross-sample contamination in the projects we curated but the proportion of contaminated samples and contamination rates varied significantly.
- These results show that this issue is widespread although some wet lab pipelines are more prone to cross-sample contamination.

- CroCoDeEL accurately detects cross-sample contamination in shotgun metagenomic data.
- CroCoDeEL is fast and works with limited computing ressources.
- We believe that checking for sample cross-contamination should be a mandatory quality control step in metagenomics.
- CroCoDeEL can now be applied to your dataset, even without negative controls.
- Next efforts will focus on the automatic decontamination of samples, where possible.

- CroCoDeEL source code is available on GitHub : https://github.com/metagenopolis/CroCoDeEL
- CroCoDeEL can be easily installed with pip or conda.

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
[4] Université Paris-Saclay, INRAE, Mathématiques et Informatique Appliquées du Génome à l'Environnement, MaIAGE, Jouy-en-Josas, , 78350, France

[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

GitHub repository