Modular Decontamination of Metagenomic Data

>4728 Samples | >100 Cities | >100 Countries

The **MetaSUB** Consortium is a global organization aimed at identifying the microbiomes of urban environments, particularly public transportation systems like subways.

Global City Sampling Day (gCSD) is an annual project organized by MetaSUB, where researchers and volunteers collect urban microbiome samples.

CAMP, the Core Analysis Modular Pipeline

To analyze microbiome samples collected during gCSD, the MetaSUB Consortium is developing CAMP, a unique analysis system that uses *modularity* to enable assembly of custom metagenomics workflows using standardized components (github.com/MetaSUB-CAMP). This approach provides researchers with a more flexible, extensible, and transparent alternative to traditional "one-click pipelines" which suffer from the black box problem and can be difficult to adapt to research purposes that deviate from their original purpose.

Here, we present decontamination of metagenomic data using two existing CAMP modules plus our own, new module.

Short-read Quality Control → **Short Read Taxonomy** → **Decontamination**

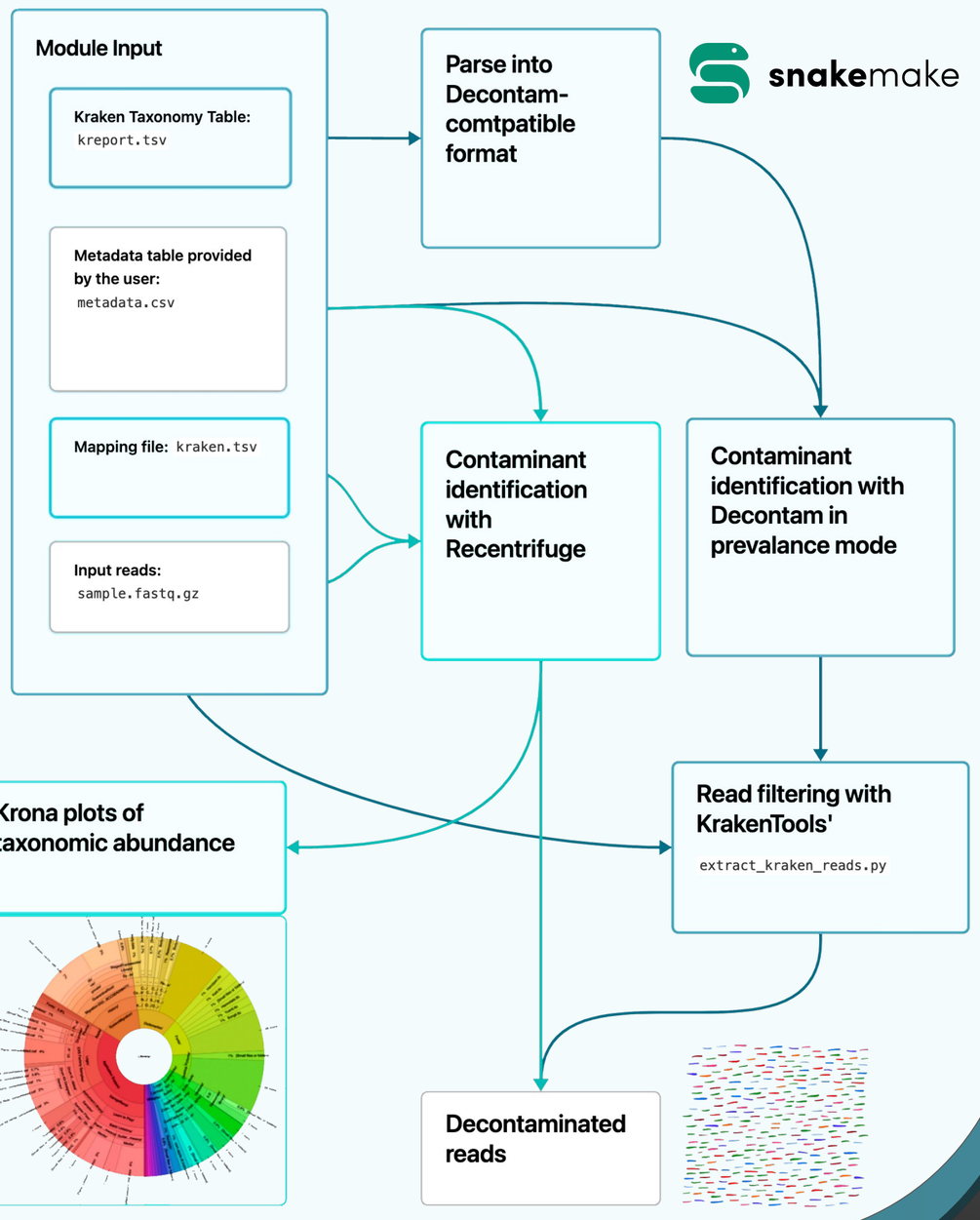
1
- low quality filtration
- adapter trimming
- removal of host reads
- deduplication
- error correction

2
- taxonomic profiling via marker genes (MetaPhlAn4) and/or k-mers (Kraken2 + Bracken or XTree)

3
- identifies microbial contaminants
- filters contaminant reads
- visualizes taxonomic profiles post-decontamination

Decontamination Steps

Like other CAMP modules, our decontamination module uses Snakemake rules to define the order of steps in the module. The current version of our module follows the structure below.



Module Construction & Testing

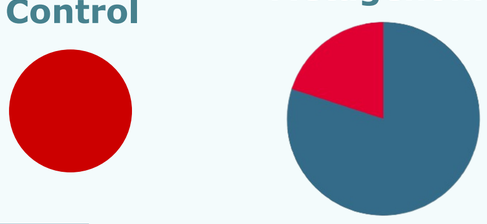
To integrate our module into CAMP, we built the public the CAMP Module Template with Cookiecutter. To test it, we began preparing two forms of input: a simulated urban metagenome with CAMISIM and data containing public reagent contaminants from Salter *et al.* (2014).

Input Data

Our testing data

Negative Control

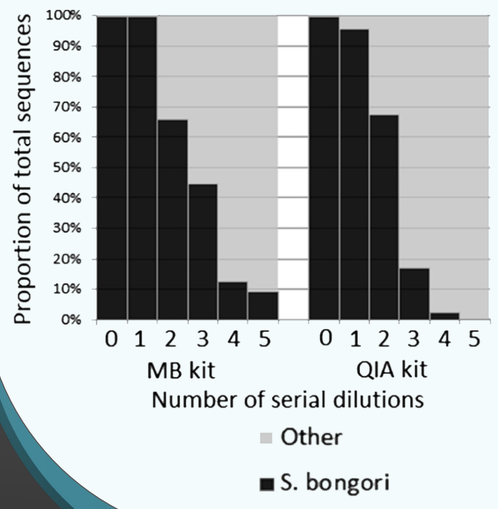
Simulated Metagenome



Core Urban Microbiome *(ex. Saccharomyces cerevisiae)*

Contaminant *(ex. Cutibacterium acnes)*

Salter *et al.* data



Reagent Samples

MP BIO & QIAGEN kits

Negative Control

Lab-grade "ultrapure" water

Config Files

Dependency management

decontamination.yml
(main conda environment)

decontam.yml
(decontam environment)

recentrifuge.yml
(recentrifuge environment)

parameters.yml
- Command-line arguments for each tool
- File paths to project databases and executables

resources.yml
Computing resources per tool (RAM & # threads)

Core Files

Snakefile
Defines the module structure

decontamination.py
Used to run the module

samples.csv
Contains the paths to the sample input

utils.py
Holds additional helper scripts


Braden Tierney PhD

Lauren Mak
SMTB
Tartu Ülikool
Zimin Foundation
Howard Hughes Medical Institute