A bioinformatic pipeline to elucidate the links between viruses and their hosts in microbial communities, applied to viruses in anaerobic digestion processes

Vuong Quoc Hoang Ngo, Cédric Midoux, Mahendra Mariadassou, Valentin Loux, François Enault, Mart Krupovic, Ariane Bize

To cite this version:

Vuong Quoc Hoang Ngo, Cédric Midoux, Mahendra Mariadassou, Valentin Loux, François Enault, et al.. A bioinformatic pipeline to elucidate the links between viruses and their hosts in microbial communities, applied to viruses in anaerobic digestion processes. JOBIM 2021 (JOBIM (Journées Ouvertes en Biologie, Informatique et Mathématiques), Jul 2021, Paris, France. . hal-04359920

HAL Id: hal-04359920

https://hal.inrae.fr/hal-04359920

Submitted on 21 Dec 2023

A bioinformatic pipeline to elucidate the links between viruses and their hosts in microbial communities, applied to viruses in anaerobic digestion processes

Vuong Quoc Hoang NGO[1], Cédric MIDOUX[1,2,3], Mahendra MARIADASSOU[2,3], Valentin LOUX[2,3], François ENAULT[4], Mart KRUPOVIC[5], Ariane BIZE[1]

[1] Université Paris-Saclay, INRAE, PROSE, 92761, Antony, France.
[2] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.
[3] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE Bioinformatics Facility, 78350, Jouy-en-Josas, France.
[4] UMR CNRS 6023 Microorganismes : Génome et Environnement, 63177, Aubière, France.
[5] Institut Pasteur, Archaeal Virology Unit, 75015 Paris, France.

Context

- Viruses are key-players in microbial ecosystems. However, predicting hosts from viruses is still a major challenge in microbial ecology.
- We developed a bioinformatic pipeline including the detection of CRISPR protospacers in viral contigs, a method previously used to predict hosts from marine viruses [1].
- We applied our pipeline to anaerobic digestion (AD) ecosystems, in the context of organic waste treatment. We focused on the diversity of viruses infecting methanogens, the latter being the key actors of methane production during AD process.

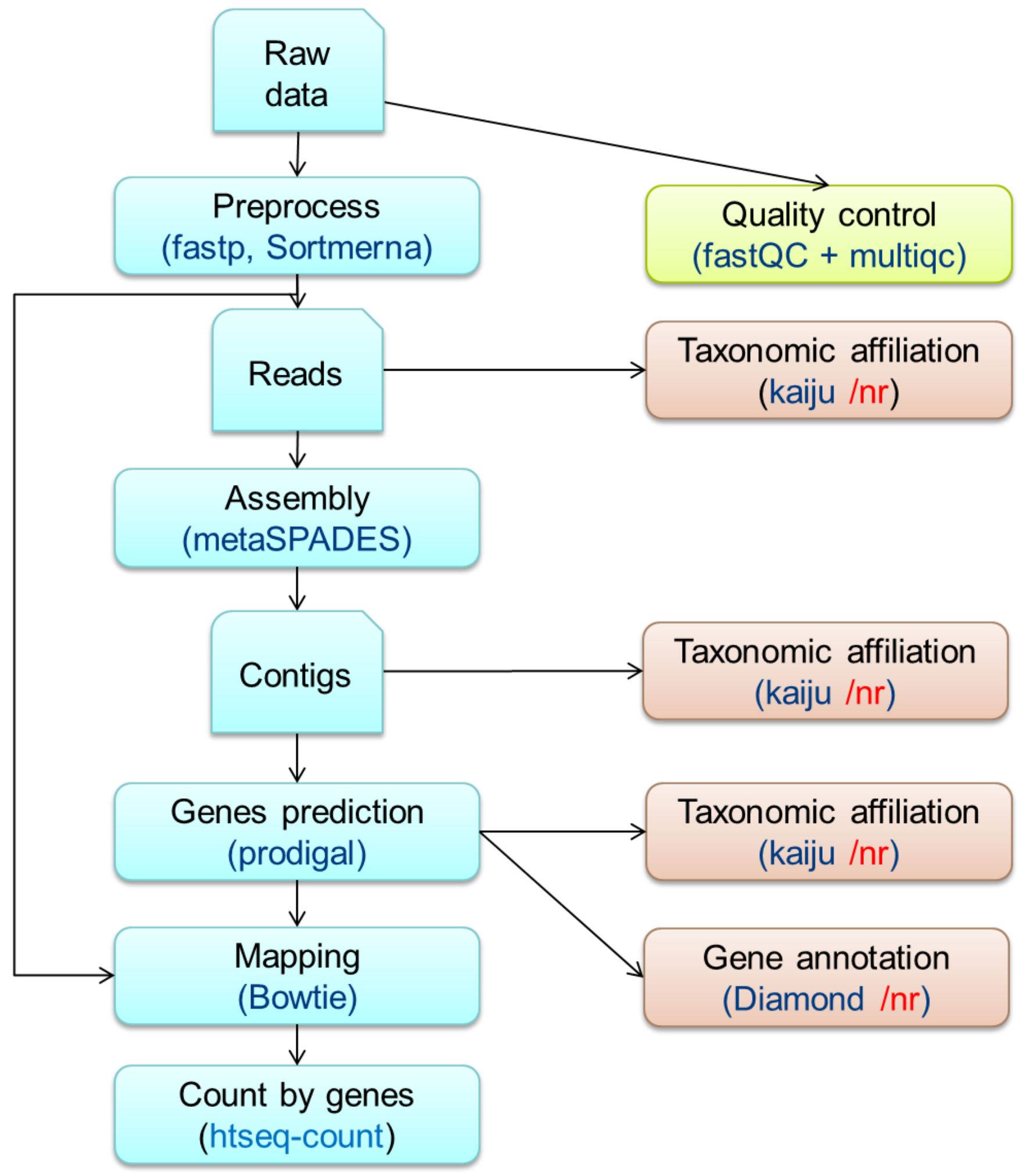
Data

- Samples originating from 2 distinct AD microcosms
- Cellular and viral DNA extracted and sequenced by Illumina NextSeq (2 * 150bp): ≈40M read pairs per sample for cellular metagenomes and ≈20M for each metavirome.

Substrate = Formate

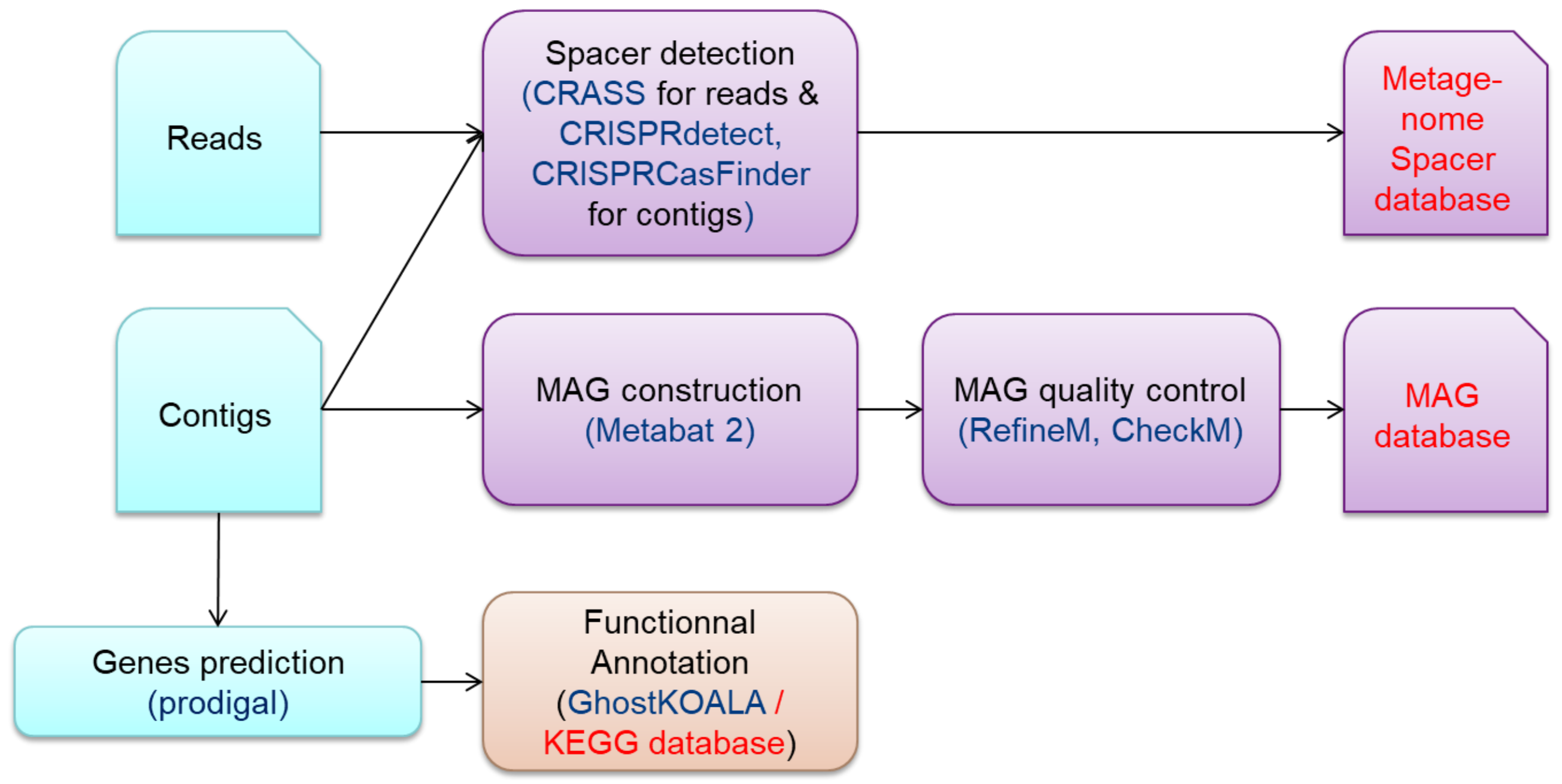
Pipeline

- The most generic steps were scripted as a *snakemake* workflow (https://forgemia.inra.fr/cedric.midoux/workflow_metagenomics) to favor reproducible and scalable data analysis. It was run on the cluster of the INRAE MIGALE bioinformatics platform.
- Several steps specifically dedicated to the prediction of hosts from viral contigs were performed using *bash* and *python* scripts.
- For the cellular metagenomes, a non-redundant spacer database was built from the spacer sequences obtained from Spacer detection step. In addition, a metagenome-assembled genome (MAG) database was constructed from cellular metagenomic data with *Metabat2*.
- Host prediction was performed by using alignment-free (*WiSH*) or alignment-dependent (*Blastn* or *SpacePharer*) methods.



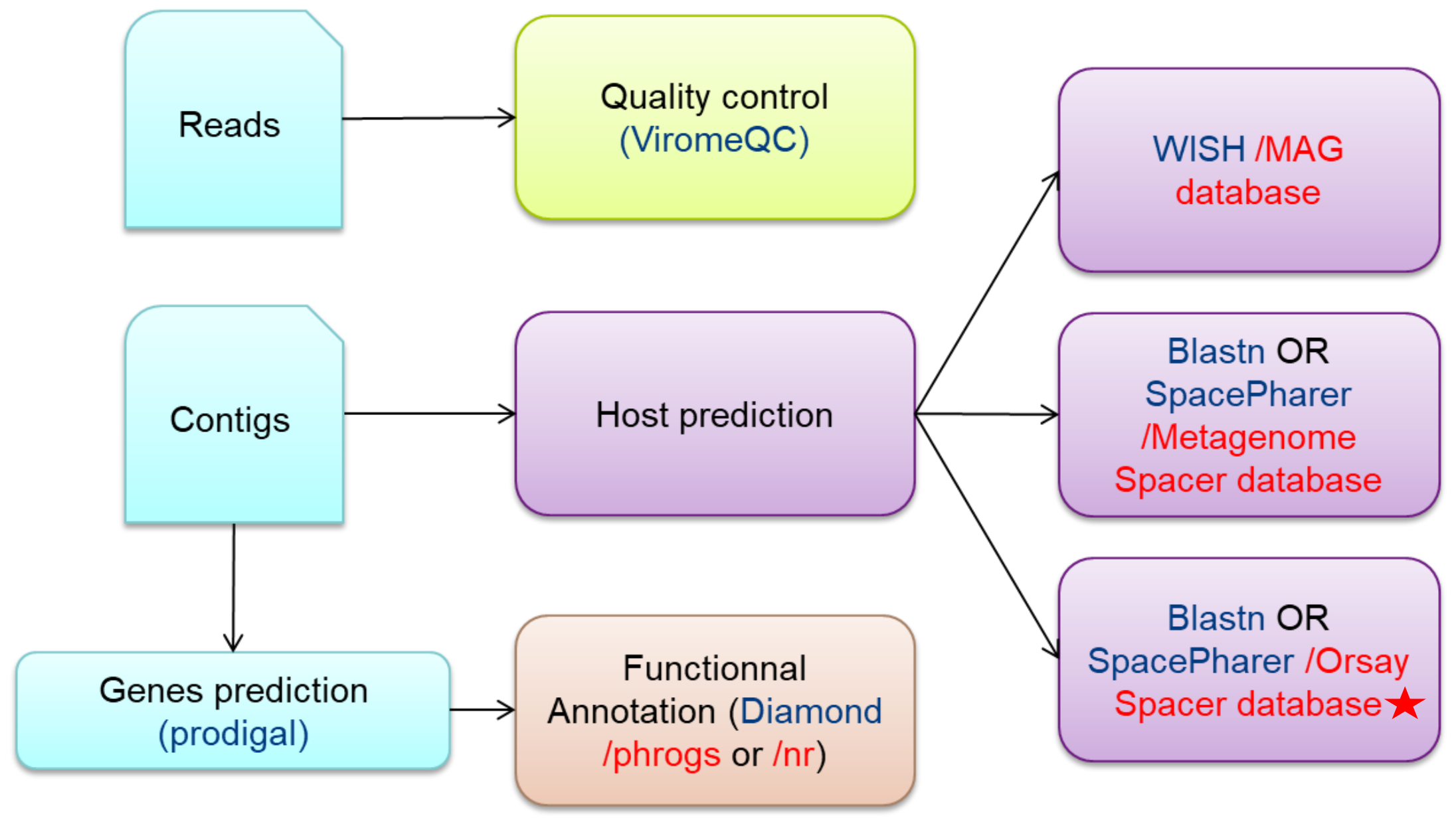
Raw data
Preprocess (fastp, Sortmerna)
Quality control (fastQC + multiqc)
Reads
Taxonomic affiliation (kaiju /nr)
Assembly (metaSPADES)
Contigs
Taxonomic affiliation (kaiju /nr)
Genes prediction (prodigal)
Taxonomic affiliation (kaiju /nr)
Mapping (Bowtie)
Gene annotation (Diamond /nr)
Count by genes (htseq-count)

Generic pipeline for both cellular and viral metagenomes (snakemake)



Reads
Spacer detection (CRASS for reads & CRISPRdetect, CRISPRCasFinder for contigs)
Metage-nome Spacer database
Contigs
MAG construction (Metabat 2)
MAG quality control (RefineM, CheckM)
MAG database
Genes prediction (prodigal)
Functionnal Annotation (GhostKOALA / KEGG database)

Steps specific for cellular metagenomes

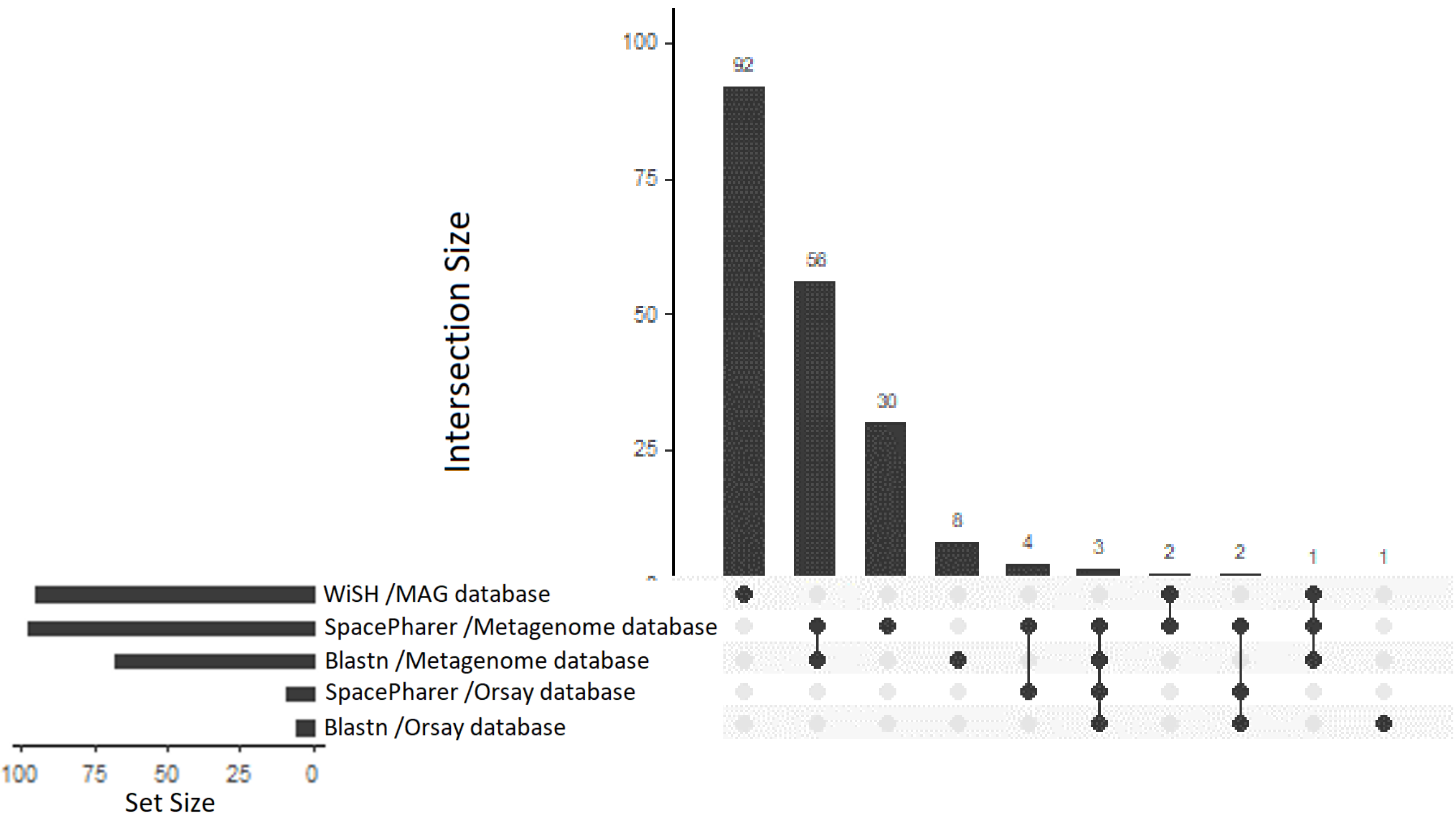


Reads
Quality control (ViromeQC)
WISH /MAG database
Contigs
Host prediction
Blastn OR SpacePharer /Metagenome Spacer database
Genes prediction (prodigal)
Functionnal Annotation (Diamond /phrogs or /nr)
Blastn OR SpacePharer /Orsay Spacer database★

Steps specific for metaviromes

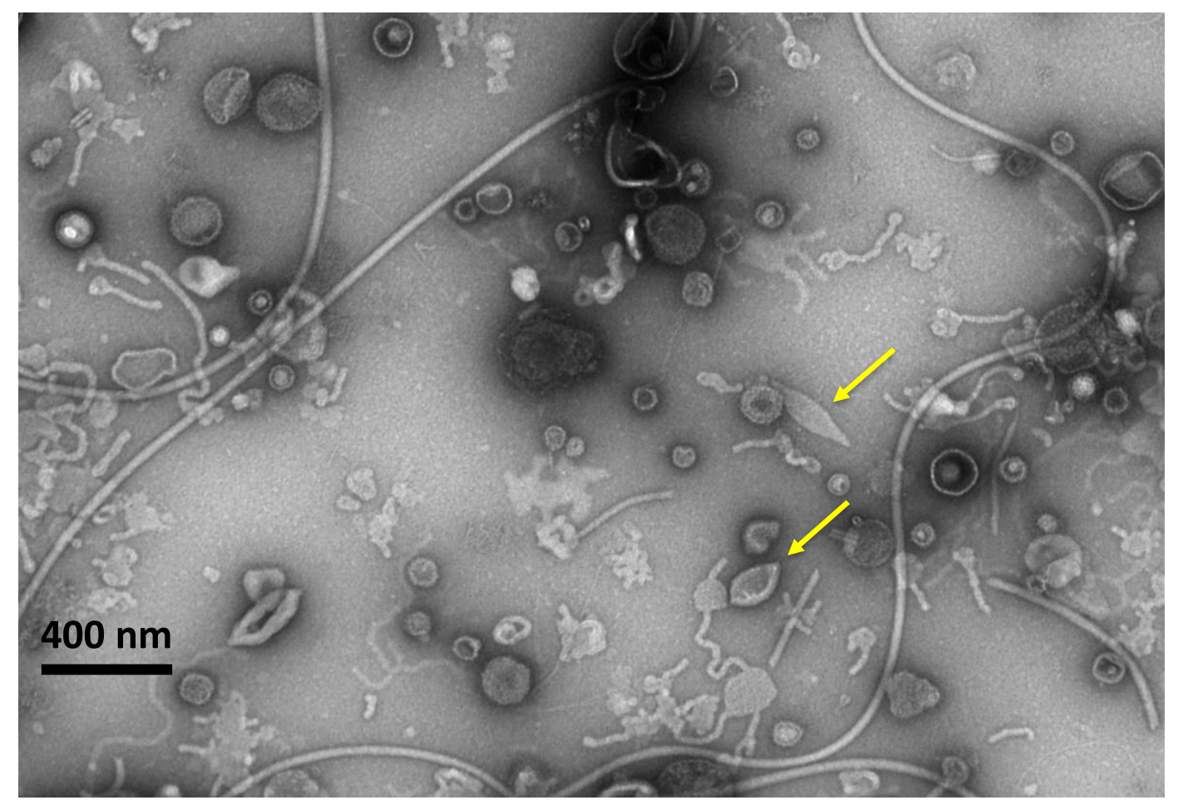
Biological results

- Out of 5570 viral contigs longer than 3 kb, 199 were identified as possibly originating from archaeal viruses thanks to complementary approaches.

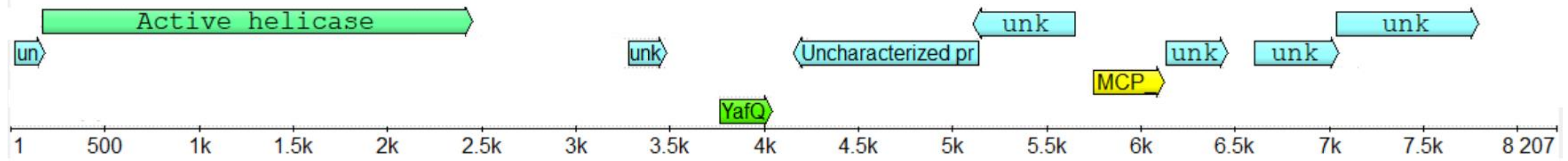


Sets of contigs identified by different approaches

- One of them is likely a novel spindle-shaped virus, a morphotype specific for archaeal viruses. Virus-like particles with this morphotype were also observed by Transmission Electron Microscopy.



MCP_S: Major capsid protein SSV1_like
yaFQ: mRNA interferase



Genomic map of the contig from the putative new fusiform virus

Conclusions

- Using the developed pipe-line, we were able to discover new viral diversity and to identify some host-virus links.
- The pipeline will be used in future metagenomics projects and will continue to be improved by the addition of new features.

References

Felipe H Coutinho, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature communications* 8, no. 1: 1-12, 2017

★ CRISPRCasdb spacer ([https://crispercas.i2bc.paris-saclay.fr/Home/DownloadFile?filename=spacer_34.zip](https://crisprcas.i2bc.paris-saclay.fr/Home/DownloadFile?filename=spacer_34.zip))

Project ANR-17-CE05-0011-01