Workshop phage genome sequencing and annotation → 9/10/2020

Participants:

	Nom	Adresse mail	Affiliation	divers
1	Marie-Agnès Petit	marie-agnes.petit@inra.fr	INRA, Jouy en Josas	Anim.
2	Mireille Ansaldi	mireille.ansaldi@imm.cnrs.fr	CNRS, Marseille	Anim.
3	Angélina Trotereau	angelina.trotereau@inra.fr	INRA, Nouzilly	
4	Fernando Clavijo	Fclavijo@imm.cnrs.fr	CNRS, Marseille	
5	Sylvain Gandon	Sylvain.GANDON@cefe.cnrs.fr	CNRS, Montpellier	
6	Rémy Froissart	remy.froissart@cnrs.fr	CNRS, Montpellier	
7	Clara Torres-Barceló	Clara.TorresBarcelo@inra.fr	INRA, Montfavet	
8	Marie Vasse	marie.vasse@env.ethz.ch	ETH, Zurich	
9	Antoine Culot	antoine.culot@agrocampus-ouest.fr	AgroCampus Ouest	
10	François Gatchitch	Francois.GATCHITCH@cefe.cnrs.fr	CNRS, Montpellier	
11	Frédérique Leroux	fleroux@sb-roscoff.fr	IFREMER, Roscoff	
12	Yannick Labreuche	ylabreuche@sb-roscoff.fr	IFREMER, Roscoff	
13	Sabine Chenivesse	schenivesse@sb-roscoff.fr	IFREMER, Roscoff	
14	Damien Piel	dpiel@sb-roscoff.fr	IFREMER, Roscoff	
15	Nicolas Ginet	nginet@imm.cnrs.fr	CNRS, Marseille	
16	Julien Lossouarn	julien.lossouarn@inra.fr	INRA, Jouy en Josas	
17	Ariane Bize	ariane.bize@irstea.fr	IRSTEA, Antony	
18	Ahlem Djedid	adjedid@imm.cnrs.fr	CNRS, Marseille	
19	Hoang NGO	hoang.ngo@irstea.fr	IRSTEA, Antony	
20	Jack Dorling	Jack.DORLING@i2bc.paris-saclay.fr	I2BC, Gif-sur-Yvette	
21	Audrey Labarde	Audrey.LABARDE@i2bc.paris-saclay.fr	I2BC, Gif-sur-Yvette	
22	Mai Chatain-Li	chatain@vetophage.fr	Vetophage, Lyon	
23	Luis Ramirez	luis-maria.ramirez-chamorro@i2bc.paris-saclay.fr	I2BC, Gif-sur-Yvette	

Context:

The workshop "phage genome sequencing and annotation" took place in Grenoble, just after the annual meeting of the Phages.fr network. We deliberately kept the number of participants around 20 to favor discussions. No real workplan had been set up, the objective being to share experience and ask basic questions for those who never practiced phage genome sequencing and annotation. The idea was also to generate a sort of toolbox and discuss the pitfalls encountered through the different steps.

DNA sample preparation tips:

- Be aware that the drop of chloroform that most people use to kill remaining bacteria can destroy some phages. Can be replaced by a

filtration step (0.2 or 0.45µm use low protein binding, polyethersulfone (PES) support, Acrodisc from Pall work fine).

- Main techniques to concentrate the lysate: on centricons (can be used to wash the sample as well), PEG precipitation + low speed centrifugation, ultracentrifugation (keep in mind that Podoviridae need very high speed 90-120 kg)

Sequencing:

Choice of the companies/facilities. Carefully review the services offered (reads long/short; number; assembly, etc). In-house facilities, MIGS, Eurofins.

For packaging (Cos/headful) information, mechanical fragmentation is required → see (Garneau *et al.*, 2017).

Reads treatment:

- FastQC: scans raw data – estimates the number of adapters remaining – gives quality statistics
- Trimmomatic (you can provide the tag sequences to remove, trim the reads below a certain quality threshold)
- Assembly : lots of information and assembler comparison in (Sutton et al., 2019): Choice of assembly software has a critical impact on virome characterization.
- Treat Fastq files → BBRIC / Galaxy
  → fastx-uniques dereplicates the reads (belongs to USEARCH, expensive but some useful tools inside; the 32bit version of USEARCH is free but limited), useful for SPADES assemblies (memory blocks sometimes).
  → SPADES based on De Bruin graphs, k-size; METASPADES, deals with repeat problems in bacterial genomes
  → MIRA and MEGAHIT, both need less memory than SPADES
  → UGENE (free) & GENIOUS (expensive) = packages

To do:
- Define criteria to decide which assembly is the best
- BLAST against a reference genome (if available) to verify the completeness of the genome
- Look in the "trash" file to check what's inside (can be the Kitome!)

What can go wrong?

- Differentiate 2 phage genomes when you have a contaminant phage genome
- Bacterial DNA contamination: treat with DNAseI prior capsid removal: check DNaseI works in your preparation buffer (Mg2+ most often required)– Ambion DNaseI is the most versatile; dialyze on a Millipore microfilter (0.02µm) to change buffer and remove potential DNAseI inhibitors. Essentially, DNaseI needs less than 100 mM NaCl and 1 mM $CaCl_2$;

- check the non-mapped reads and start over the assembly (Bowtie 2 in UGENE for eg.)
- Problems with tag removal

Always perform a BLASTn against nr/nt section of NCBI after those steps, to check if your phage has a close relative.

Where do you place the +1?

- Follow the +1 of another/close genome in a database
- Otherwise, place it before the terminase small subunit encoding gene (large terminase is easier to find though, and small subunit is usually just ahead of the large one)

How do you know the contig is a complete phage genome?

- Look for ~127 nt long repeats at the extremities (with the Repeat tool in clone manager), they are often (but not always) left after a Spades assembly. It is an artifact, so remove one of the repeats, circularize the molecule, and place the +1 according to above guidelines.

Problems with updating? "Freeze" your pipeline in Galaxy whatever the updates made on the different programs to avoid problem of reproducibility

Define the ORFs – gene calling

Keep in mind ORFs can be messy in phage genomes: overlapping starts, ORF inside an ORF, gene fusion…

Most used softwares:

- RAST / virus option → better than PROKKA and multifasta file treatment is fast
- PROKKA → can miss small genes
- MaGe → does not take viruses genomes, you must "trick" the system calling your phage genome a plasmid for eg. (MaGe demo by Nico and Fred)
- MULTIPHATE → includes PHANOTATE, compares different gene callers, finds more genes, more false positives as well?
- PHANTOME → developed for mycophages
- VIRFAM → for specific viral genes; contains a module that allows neck protein identification (usually difficult to find), looks for the neck protein. Associated to MCP-terminase-tail (conserved proteins) – among outputs, a tree placing your phage next to its close relatives in terms of neck proteins.

Be careful, related phage genes may have less than 20% aa-identity but a similar 3D fold, so structure prediction is useful in this case. Alternative to BLAST: HHsearch, makes distant homology search, toolkit HHPred and HHblit (proba 0.9-1 indicates a reliable prediction, between 0.7 and 0.9 can still be relevant), very powerful but time consuming (no automation).

- PhROGs = Phage Remote Orthologous Groups, is a web site under construction by F. Enault and coll., conceived as an update of ACLAME, which clusters the proteins extracted from the Virsorter phage dataset into super-families of remote homologs, and annotate them. Should come out soon!

What problems can arise from the annotation step?

- Presence of INTRONS
- Distinguish between virulent and temperate: look for a lysogeny module (repressor, Cro, integrase genes); try PHASTER predictions (not always accurate when genome assembly is poor); look for AMGs – Auxiliary Metabolic Genes (virulent only); Blast the MCP protein and find out if it's found in bacterial genomes (temperate) see (Grose and Casjens, 2014)
- SNP detection → BRESEQ, SNIPPY, PINDEL, MAUVE, …

And remember, better to give a putative function that nothing since it helps to curate the databases!

Phage genome submission EMBL-ENA / NCBI (tutorial by Julien)

- Create an account, add all the persons involved as an author list submission
- Name the project NCBI or the study ENA
- Be careful with the release date, it cannot be changed once published
- File format conversion tool from RAST

TO PLAN:
Have a meeting with the MICROSCOPE team (C. Médigue, D. Vallenet) to include phage genomes in MaGe (N. Ginet, F. Leroux).