Nano-Stabilics : Workflow & Results

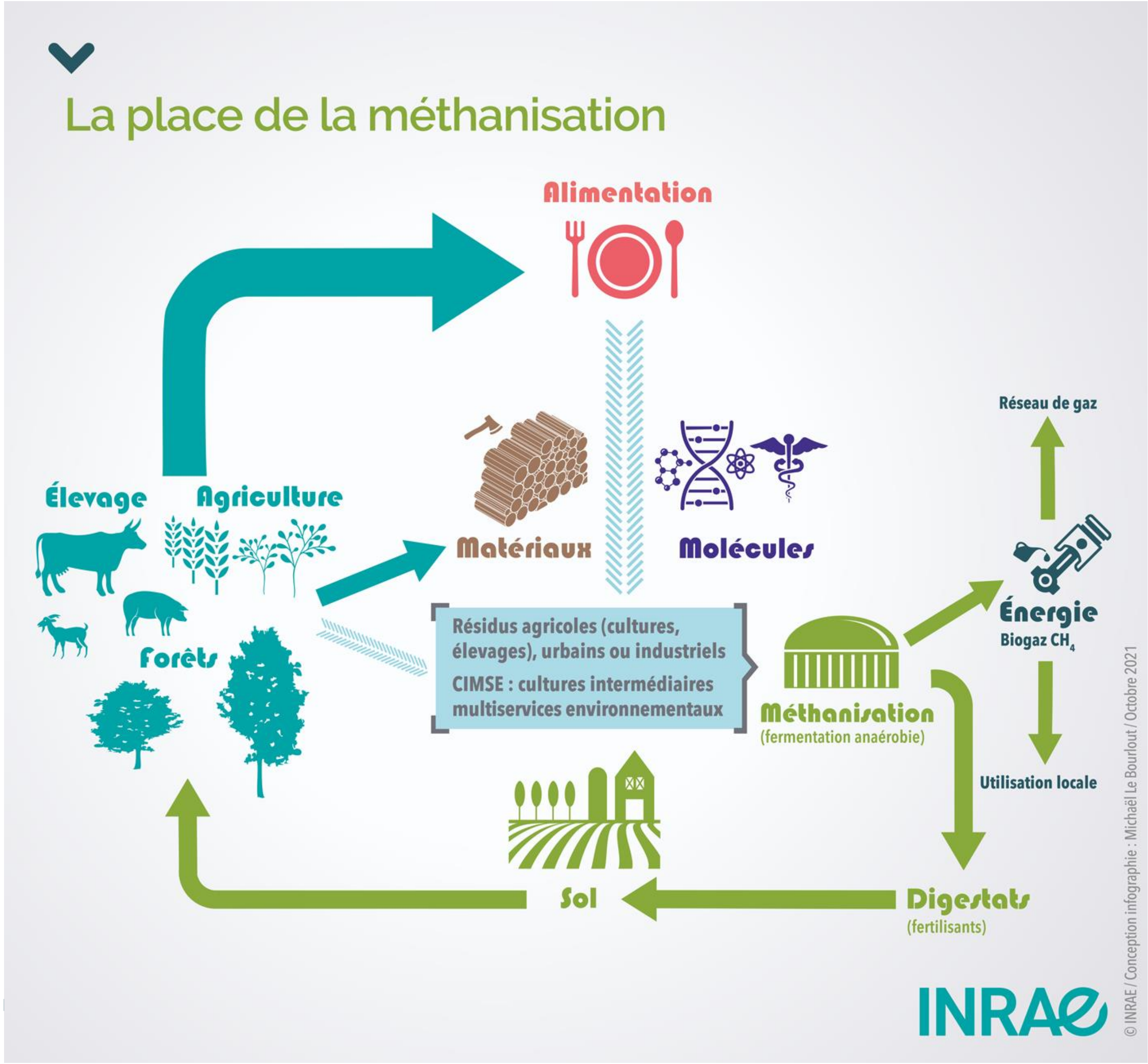
Cédric Midoux[1, 2, 3], Chrystelle Bureau[1], Baptiste Quentin[1] and Olivier Chapleur[1].

1. INRAE, UR 1604 – PROSE
2. INRAE, UR 1461 – MaIAGE
3. BioinfOmics, MIGALE bioinformatics facility

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Context



© INRAE / Conception infographie : Michaël Le Bourlout / Octobre 2021

STABILICS Project

https://anr.fr/Projet-ANR-19-CE43-0003

PRocédés biOtechnologiques au Service de l'Environnement

Nouvelles perspectives dans les déterminants de la stabilité des bioprocédés anaérobies en couplant des approches multi-omiques et statistiques

Olivier Chapleur

Baptiste Quentin

Challenges

- What is the stability of anaerobic microbial bioprocess in front of environmental perturbations?

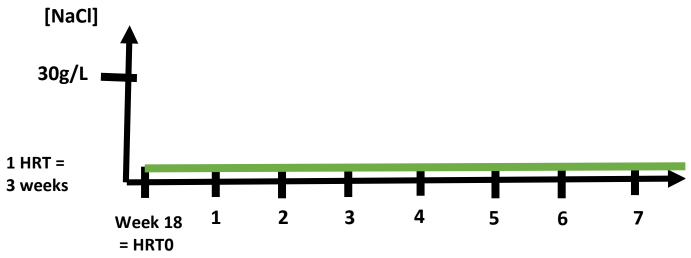
- 4 triplicates of semi-continuous anaerobic digesters
  - Reactors fed with biowaste
  - Working volume = 5 L
  - **4 perturbation scenarii** with NaCl addition
  - Monitoring & sampling for **14 months**

- Bioinformatics questions over time:
  - Community structure and composition? = **MAGs catalog**
  - Genes and metabolites levels? = **Functional annotation**

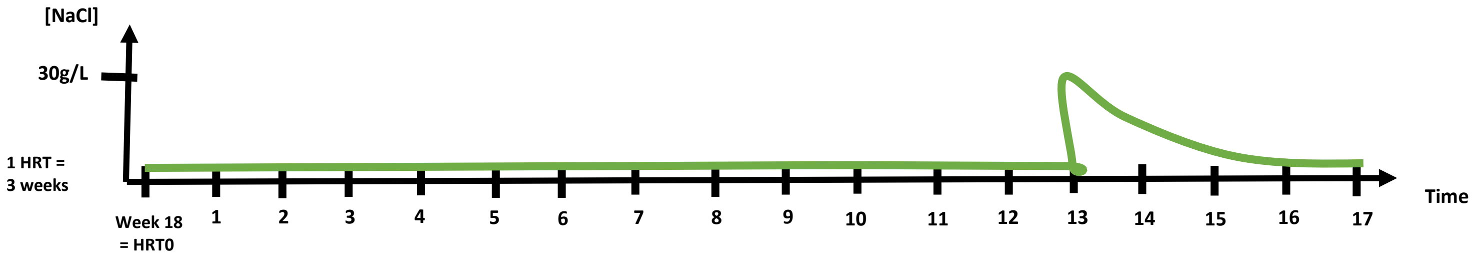
- Biostatistics, data integration, modelling, …

Scenarii of perturbations

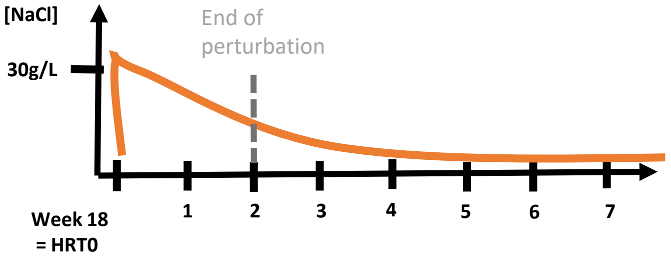
« Control »
Triplicate ABC



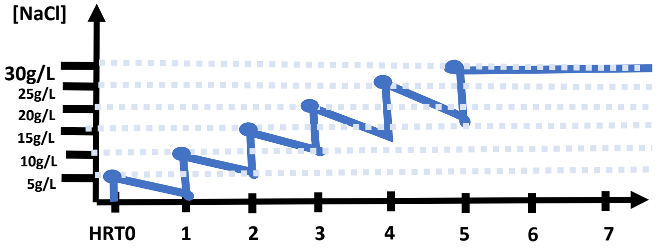
« Late »
Triplicate ABC



« Length »
Triplicate DEF



« Ramp »
Triplicate JKL



Sequencing

Chrystelle Bureau

- 138 samples from anaerobic digesters (replicated time-series)
- Oxford Nanopore Technologies
  - MinION Mk1C
  - Flow-cells R9.4.1
  - Rapid Barcoding Kit
  - Multiplexing
- Without short reads!

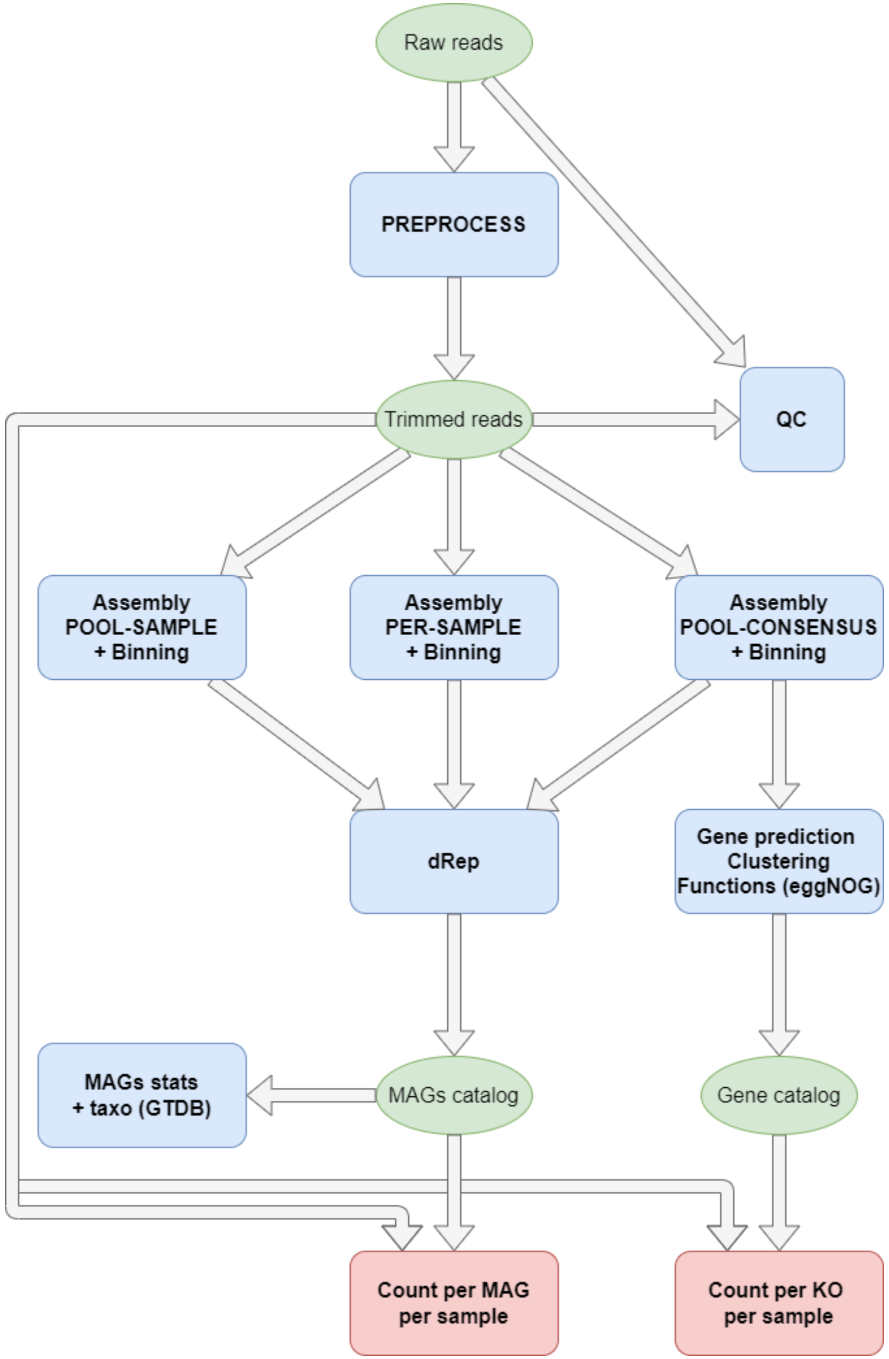
Workflow

Global view

- QC & Preprocess

- Triple approach for MAGs building
- Pool and dereplicate MAGs coming from the different approaches
- MAG taxonomic affiliation and count

- Genes prediction
- Dereplication
- Functional annotation and count

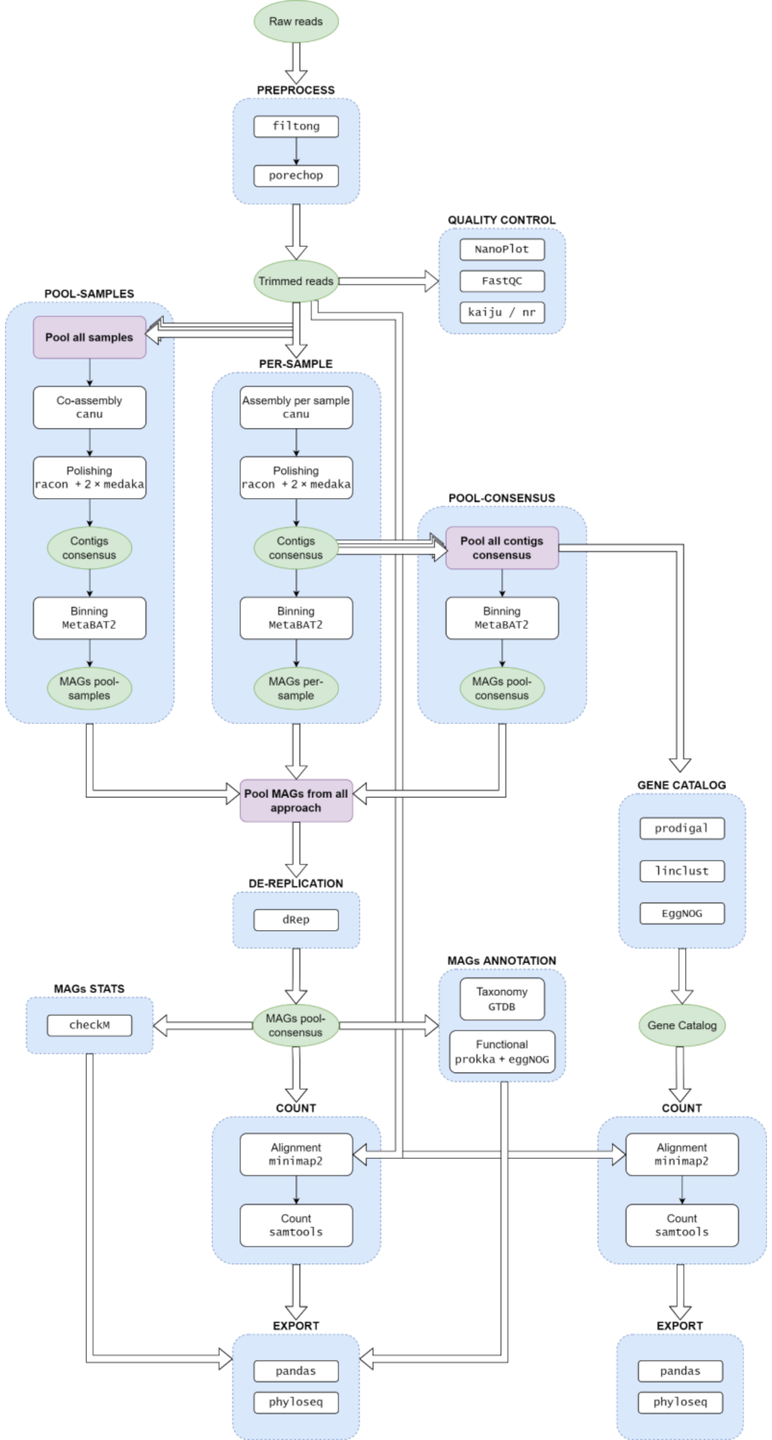

Raw reads
PREPROCESS
QC
Trimmed reads
Assembly POOL-SAMPLE + Binning
Assembly PER-SAMPLE + Binning
Assembly POOL-CONSENSUS + Binning
dRep
Gene prediction Clustering Functions (eggNOG)
MAGs stats + taxo (GTDB)
MAGs catalog
Gene catalog
Count per MAG per sample
Count per KO per sample

Global view

- QC & Preproces

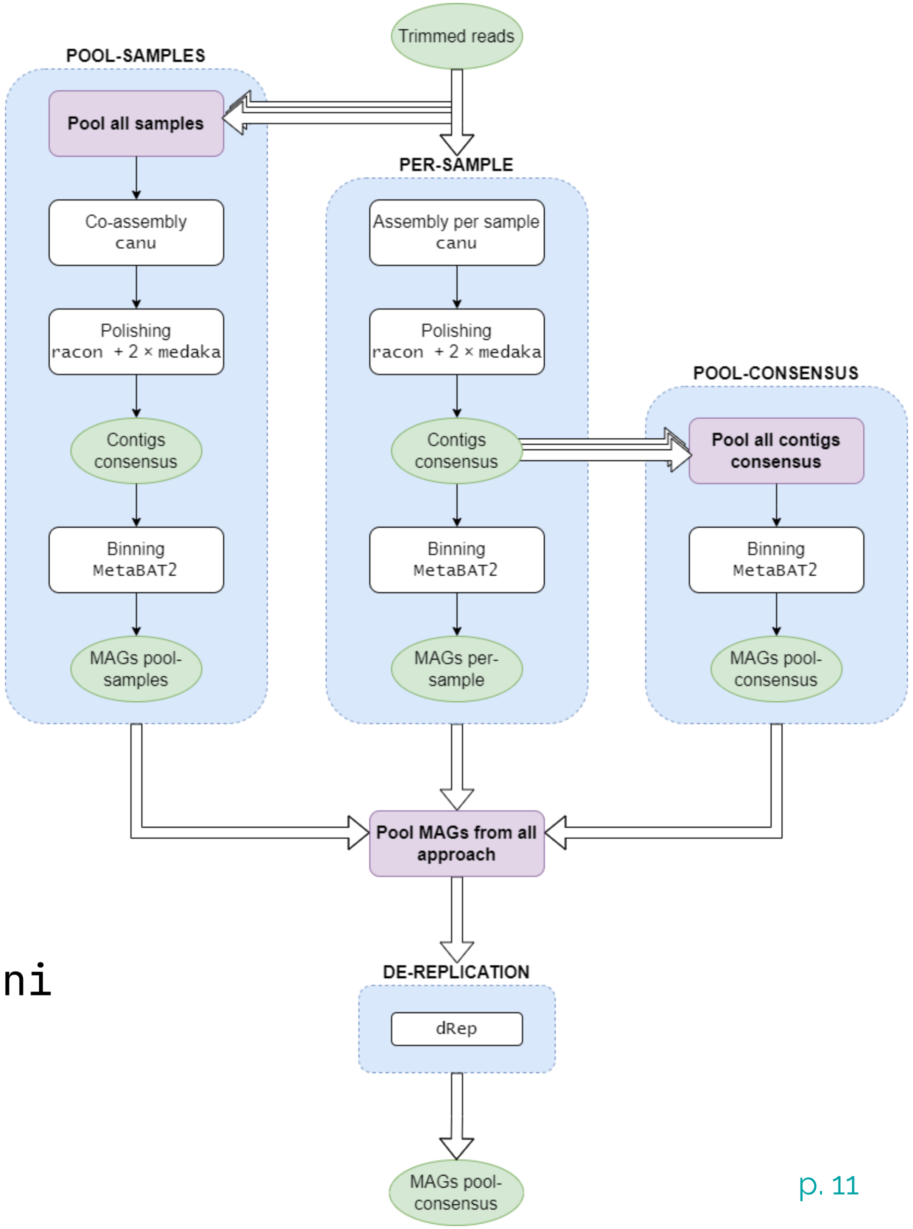
- Triple approach for MAGs building
- Pool and dereplicate MAGs coming from the different approaches
- MAG taxonomic affiliation and count

- Genes prediction
- Dereplication
- Functional annotation and count



MAGs bulding

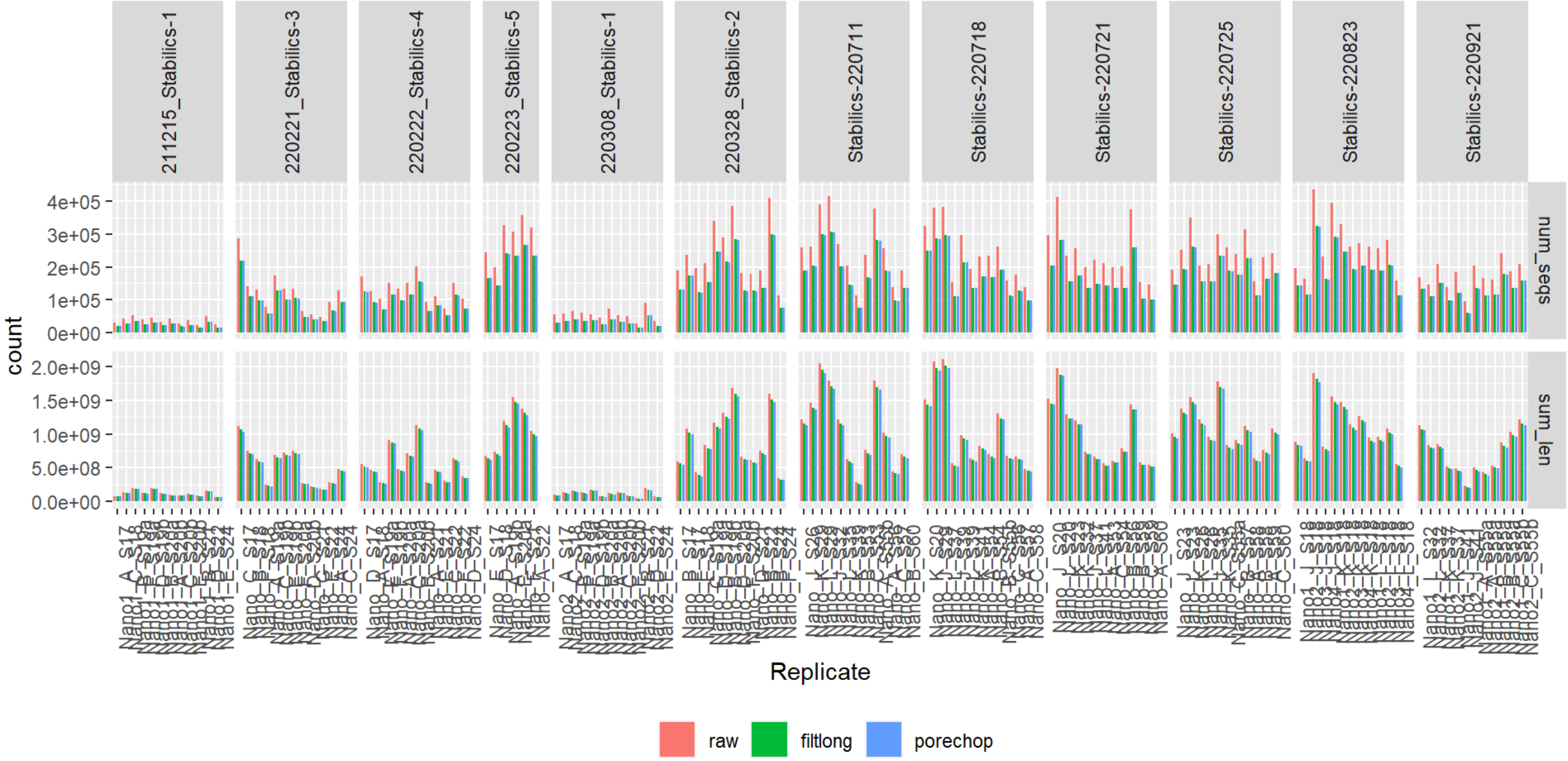
- Triple approach
  - **PER-SAMPLE**: Assembly and binning individually for each sample
  - **POOL-SAMPLES**: Co-assembly and binning
  - **POOL-CONSENSUS**: Assembly per sample and co-binning
- Assembly & Polishing
  - `canu genomeSize=5m`
  - `minimap2` + `racon --include-unpolished` + 2 × `medaka_consensus`
- Binning
  - `metabat2 --minContig 1500 --maxEdges 500`
- Pool MAGs and dereplication
  - `dRep dereplicate --S_algorithm fastANI --P_ani 0.9 --S_ani 0.95 --completeness 0 --contamination 10`



Results

Preprocess

- Number of reads and bases per sample, group per sequencing run (chronologically)
- Quality and depth depend on runs and increases over our experience
  - 220308_Stabilics1 made on the same (bad) extraction than 211215
- After preprocessing, we only loose short reads



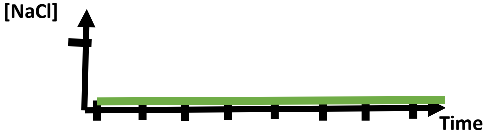
Number of reads

Number of bases

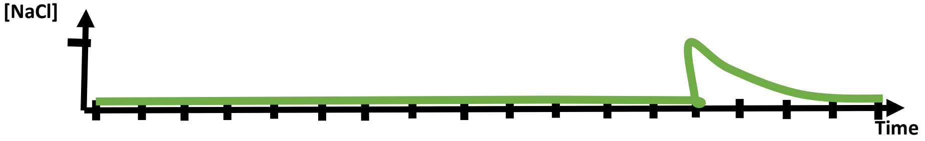
Taxonomic annotation / nr_euk

- Taxonomic barplot per sample
  - Facet grid : Days × Condition
  - Triplicat (Pilote)
  - Level = Kingdom
  - Boxborder = [NaCl]

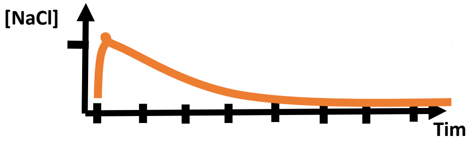
« Control »



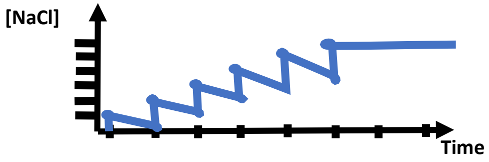
« Late »

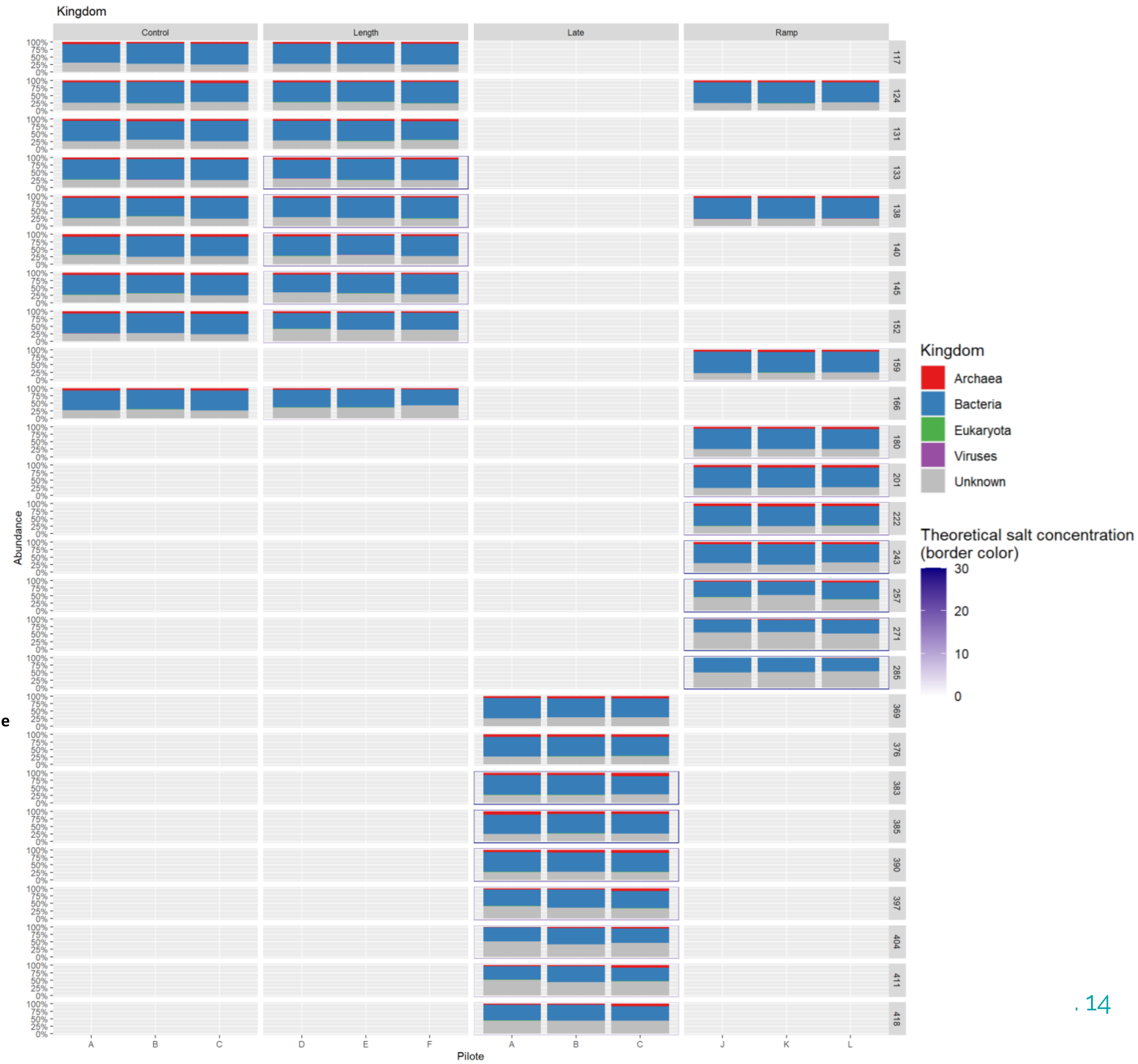


« Length »



« Ramp »

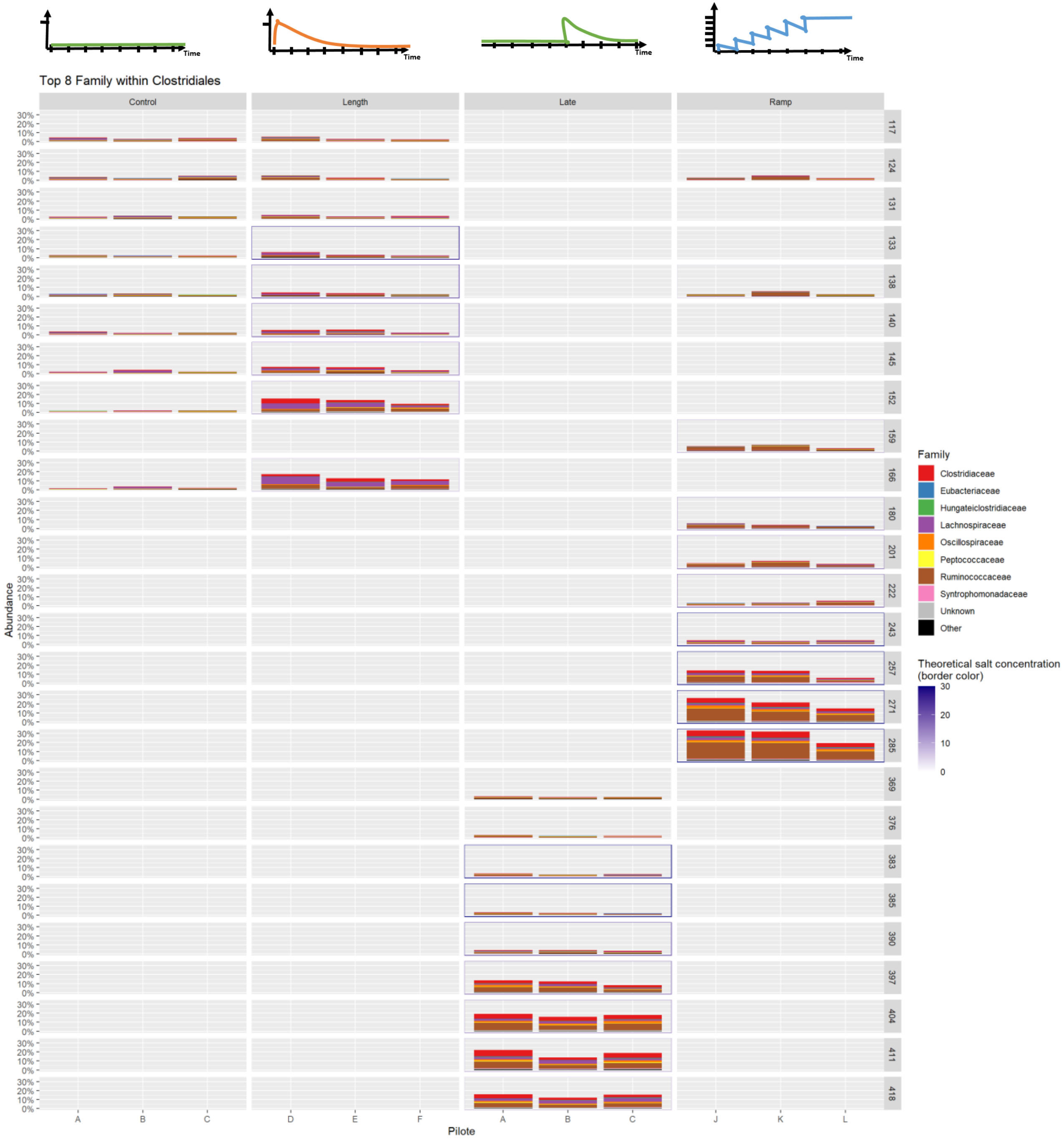




Taxonomic annotation / nr_euk

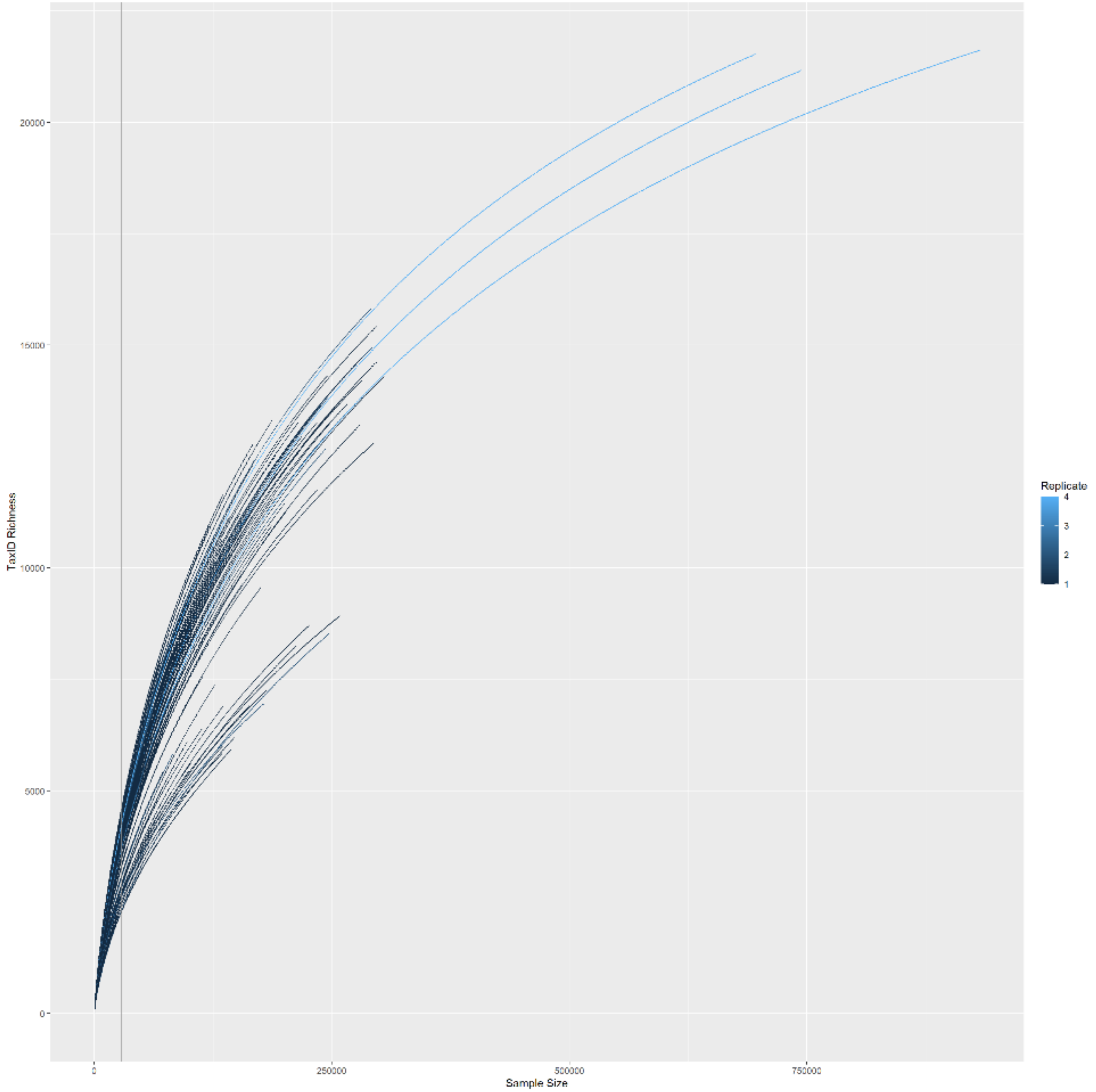
- At family level, inside *Clostridiales*
  - Relative abundance of *Clostridiales* increases after perturbation

- It’s a good observation (*but 16S experiments maybe sufficient for that*)



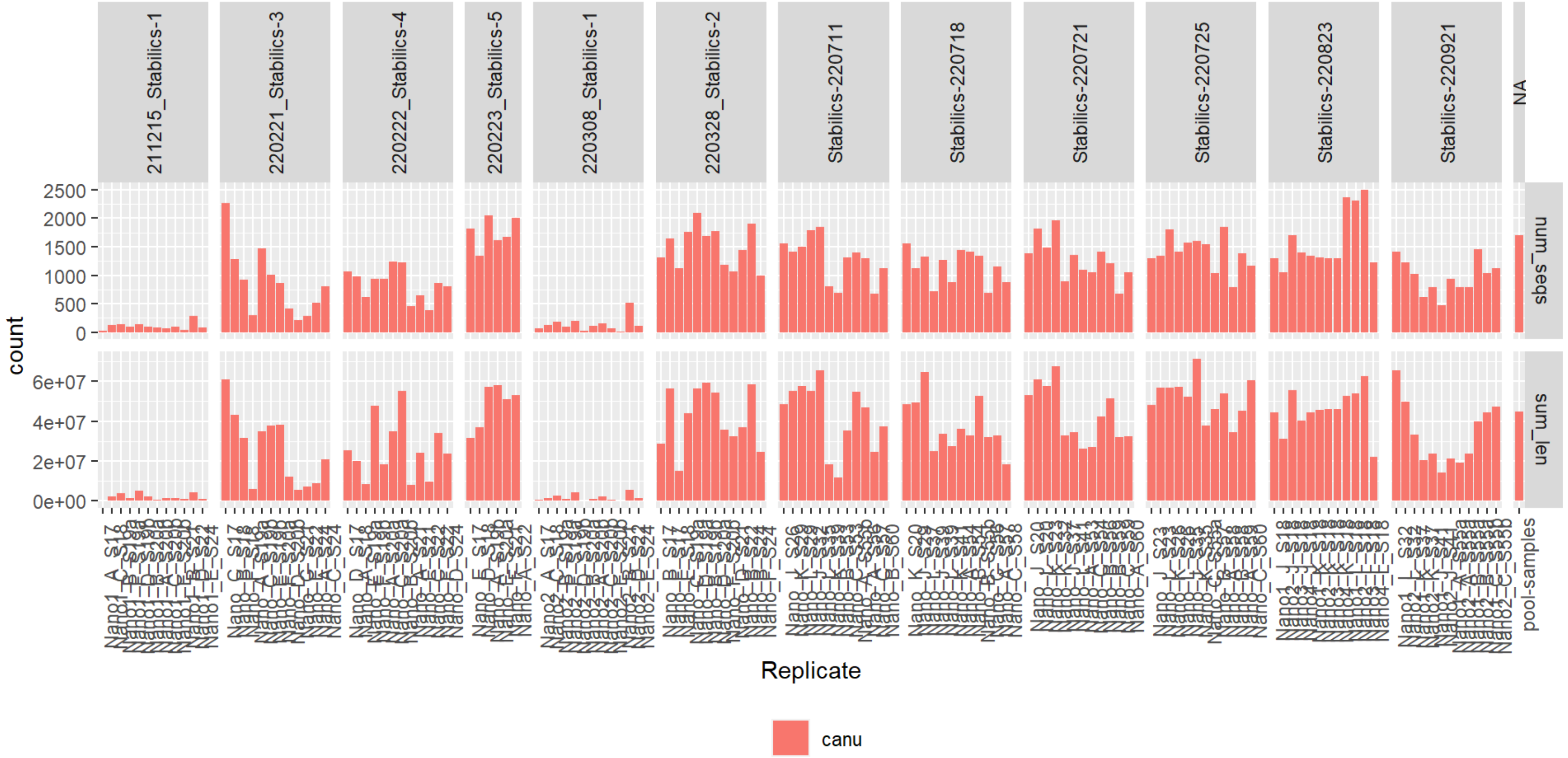
Taxonomic annotation

- Rarefaction curve on TaxID of raw reads
  - Nanopore approach is not sufficient to observe the rare biosphere of bioprocess



Assembly PER-SAMPLE

- The size of assembly (per-sample with canu) follows the number of reads
- Some samples (with low data) don’t assemble well
- We try to build a common MAGs catalog to be able to analyze all samples



MAGs catalog

RÉPUBLIQUE FRANÇAISE
*Liberté
Égalité
Fraternité*

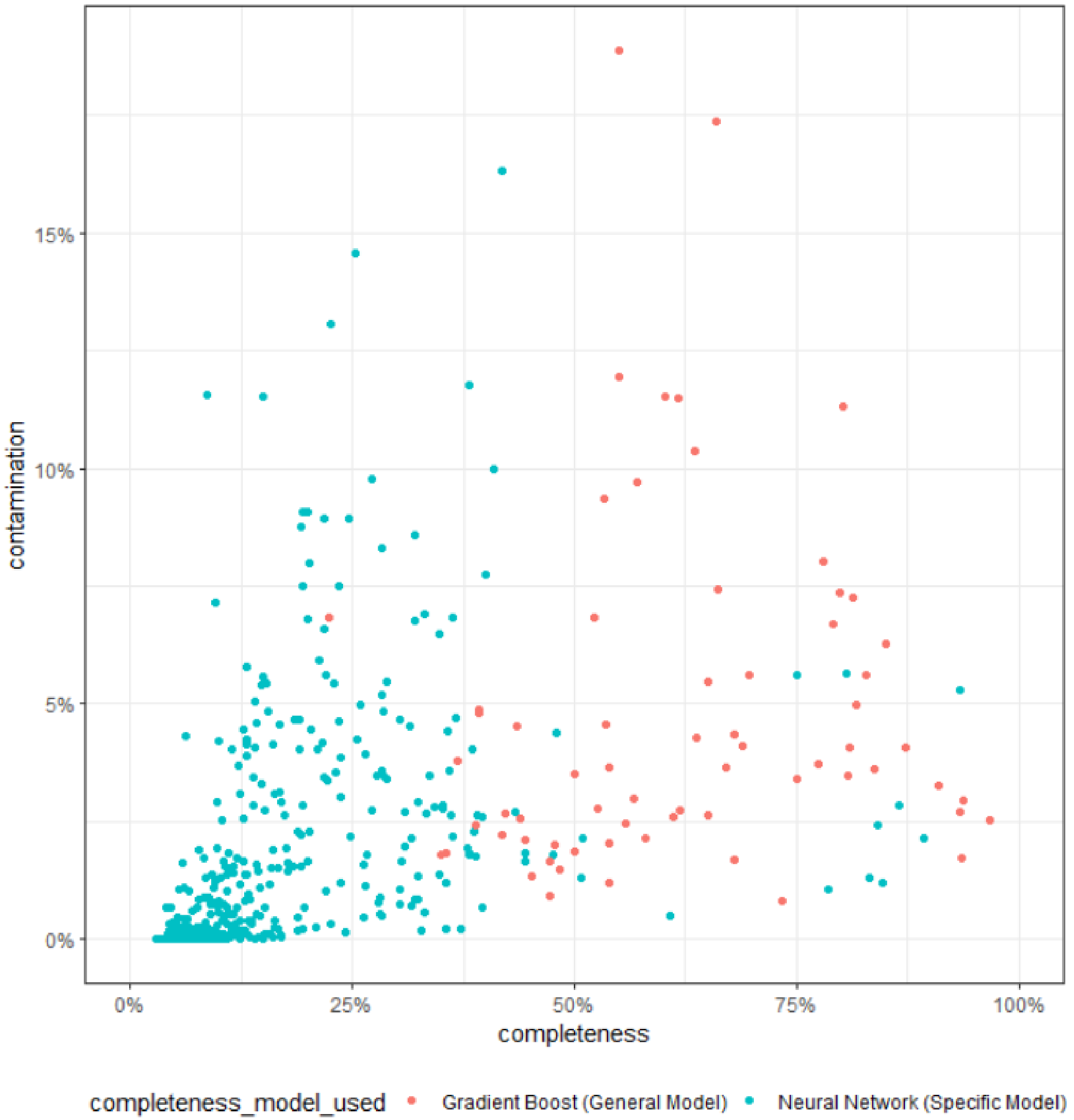
INRAE

Binning – METABat2 checkm2 dRep

- Dereplication: 2 754 genomes were given to dRep
  - 1~39 bins(/sample) from PER-SAMPLE
  - 29 bins from POOL-SAMPLES
  - 477 bins from POOL-CONSENSUS

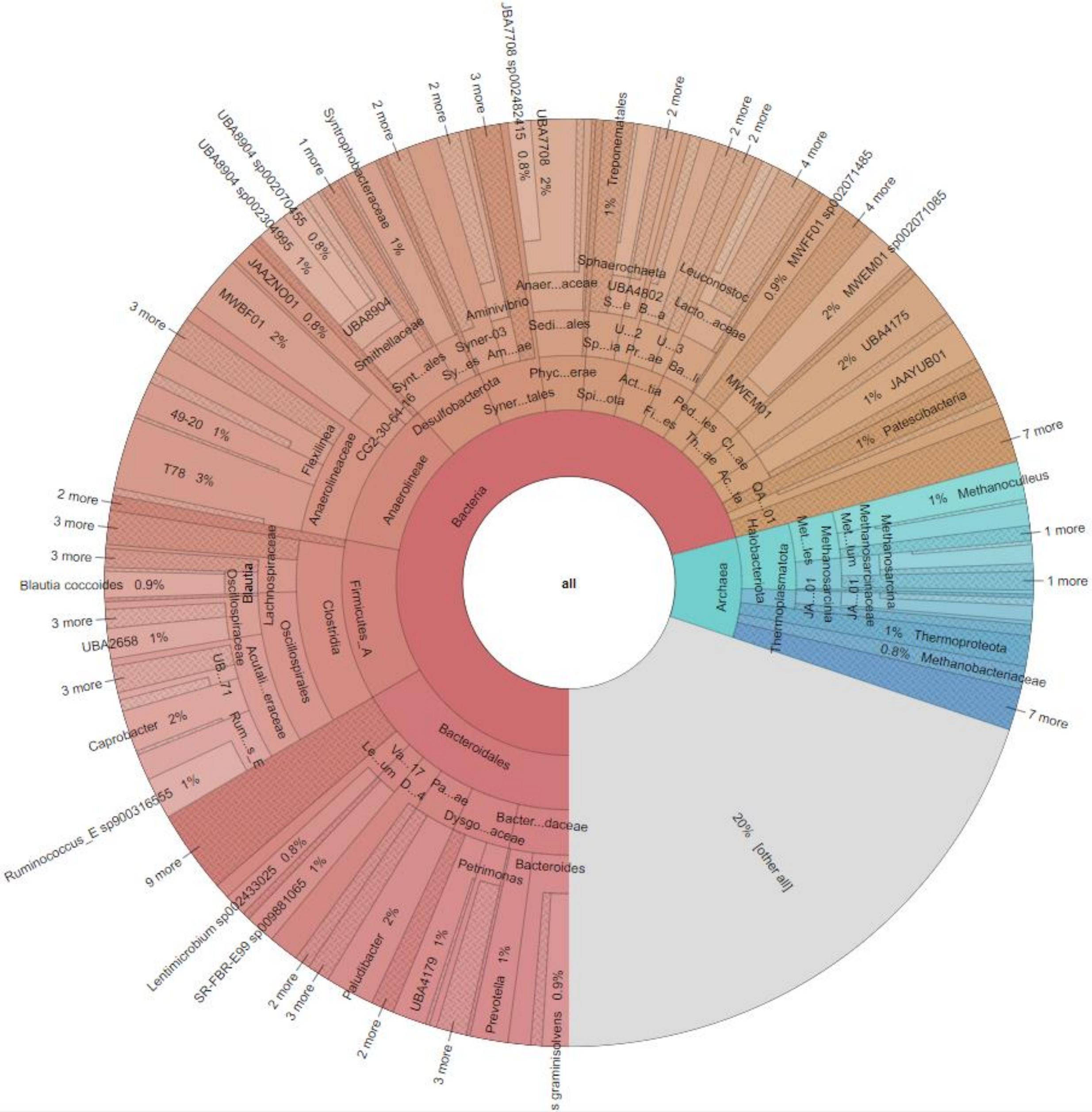
- Binning: 770 MAGs
  - No filter on completeness
  - HQ-MAG: comp>80% & conta<10%
  - ➔ 20 HQ-MAGS

- A large part of the reconstructed MAGs are incomplete.



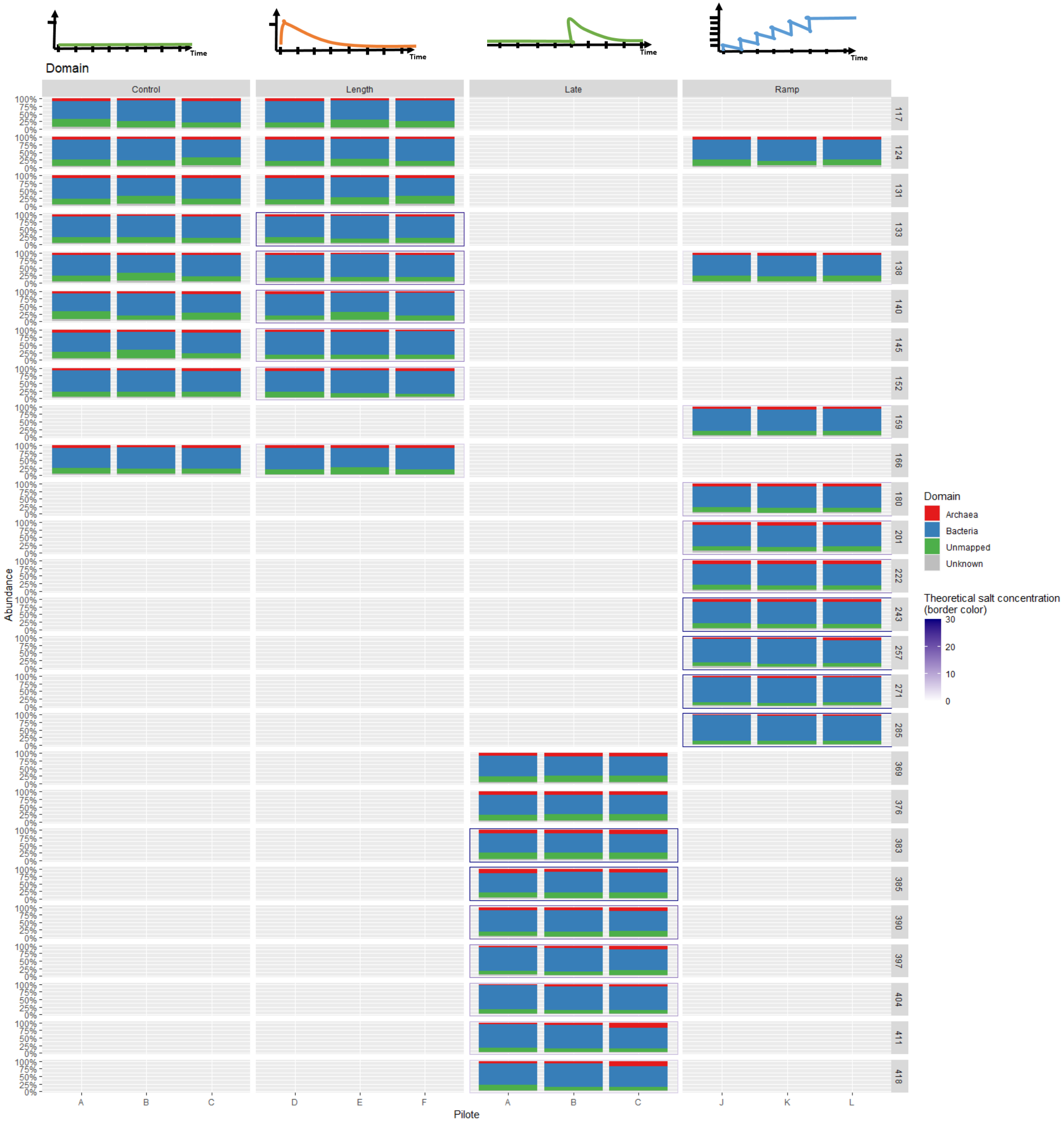
MAGs Annotation – GTDB-Tk

- We can find known bioprocess microorganisms in the MAGs annotation.
- 20% of MAGs are “Unknown”



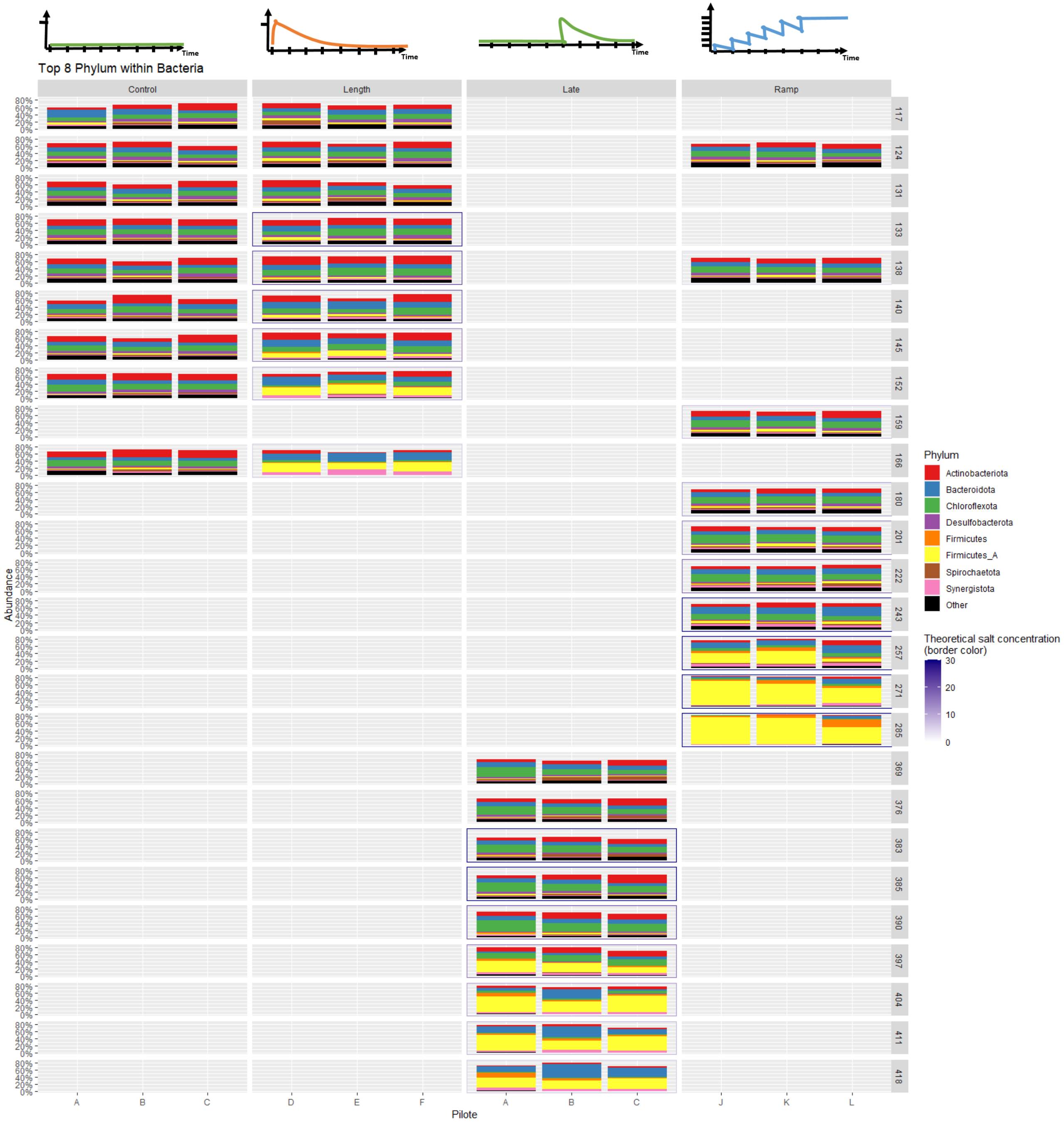
MAGs count

- Primary mapped ≈ 75% - 80%
- Unknown MAGs represents very few reads
- How to correctly normalize the data?



MAGs count

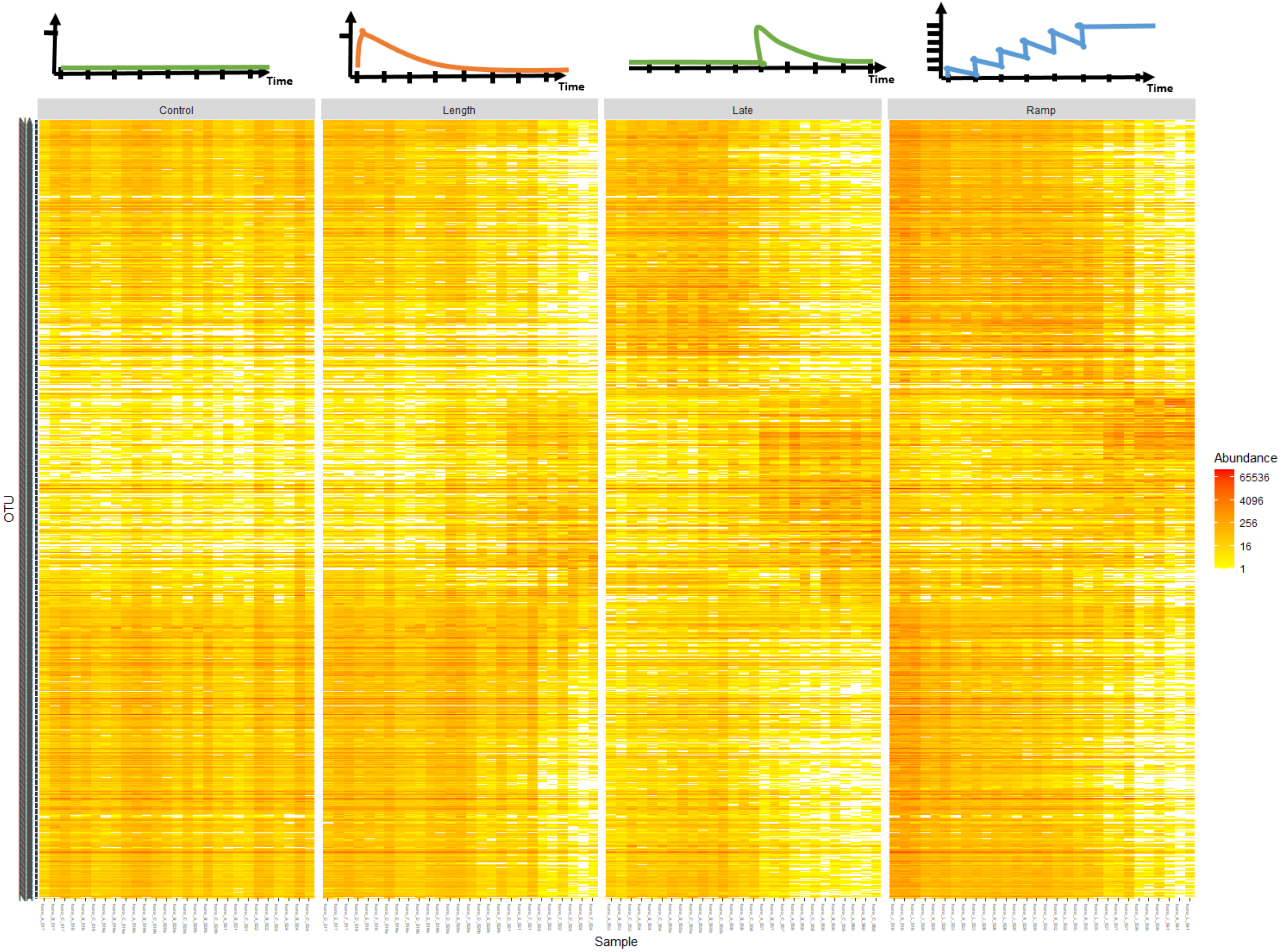
- Inside Bacteria, we can see, for example, the relative appearance of Firmicutes (yellow + orange) after perturbation



MAGs count

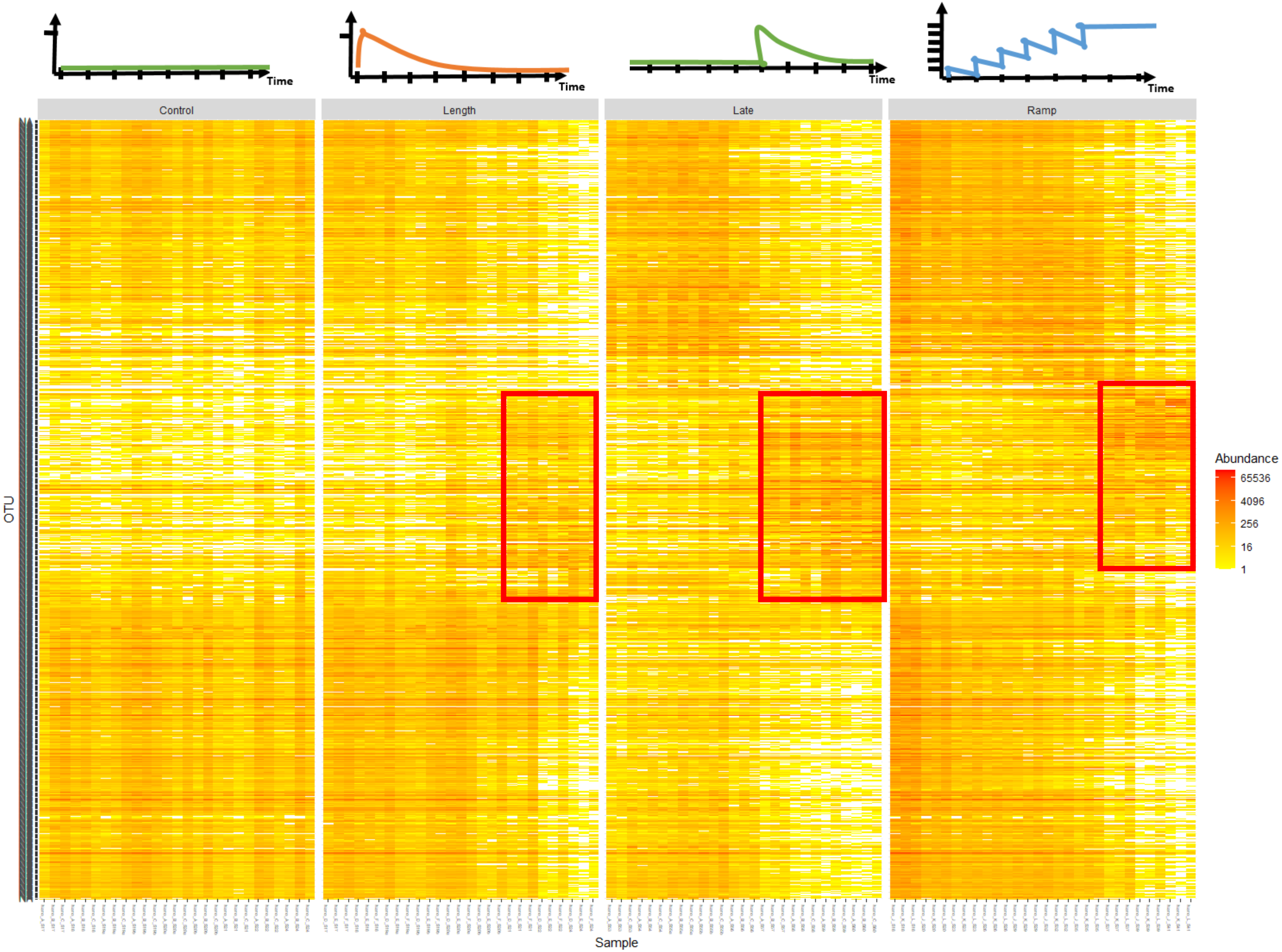
- OTU per line
- Sample per column
  - Group: Perturbation
  - Order: Days

- Community switch over time after perturbation



MAGs count

- OTU per line
- Sample per column
  - Group : Perturbation
  - Order : Days
- Community switch over time after perturbation
- Statistical analysis to come:
  - Diversity analysis
  - Identification of differentially abundant taxa



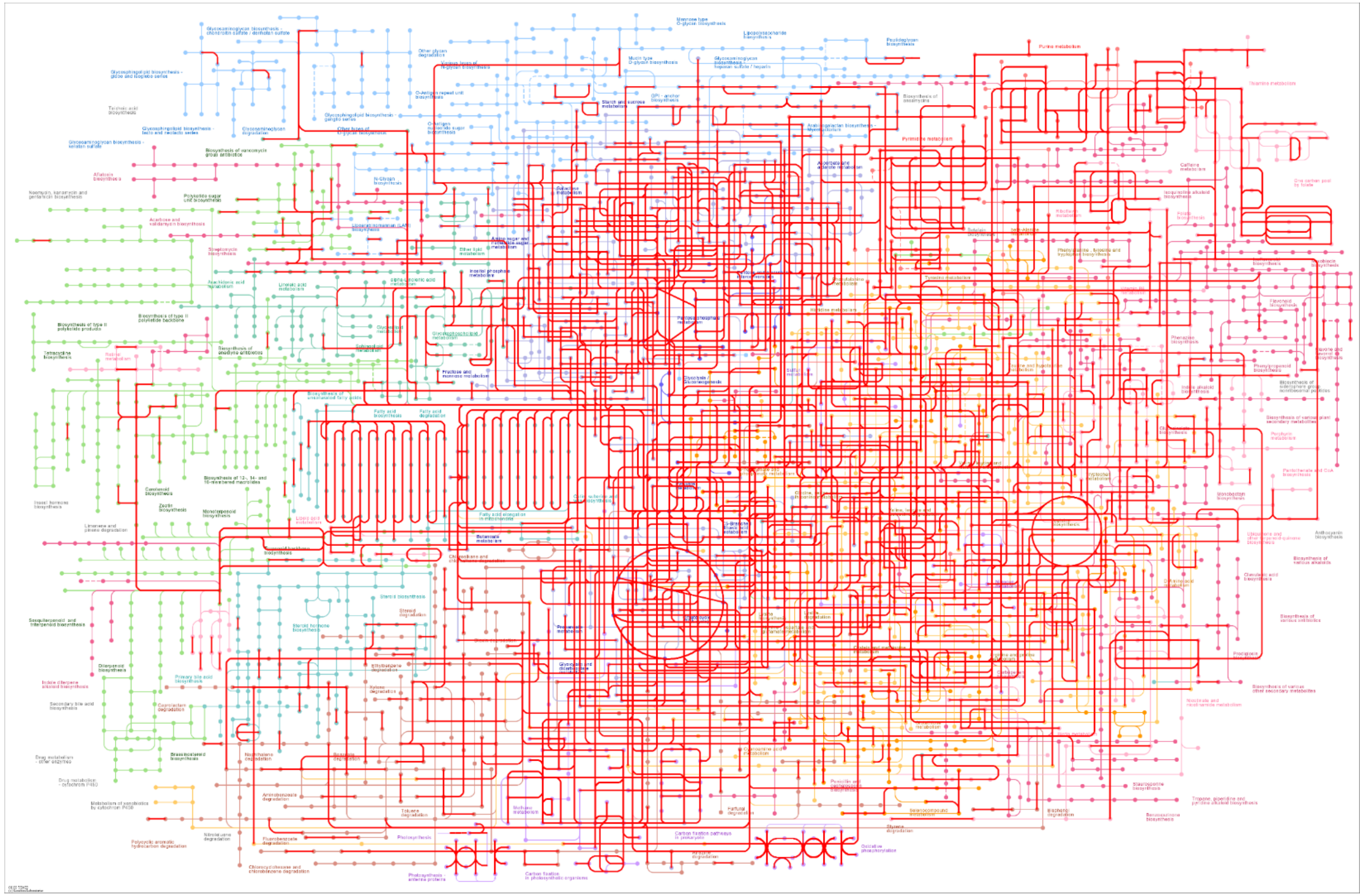
Genes catalog

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Genes prediction – `prokka linclust EggNOG`

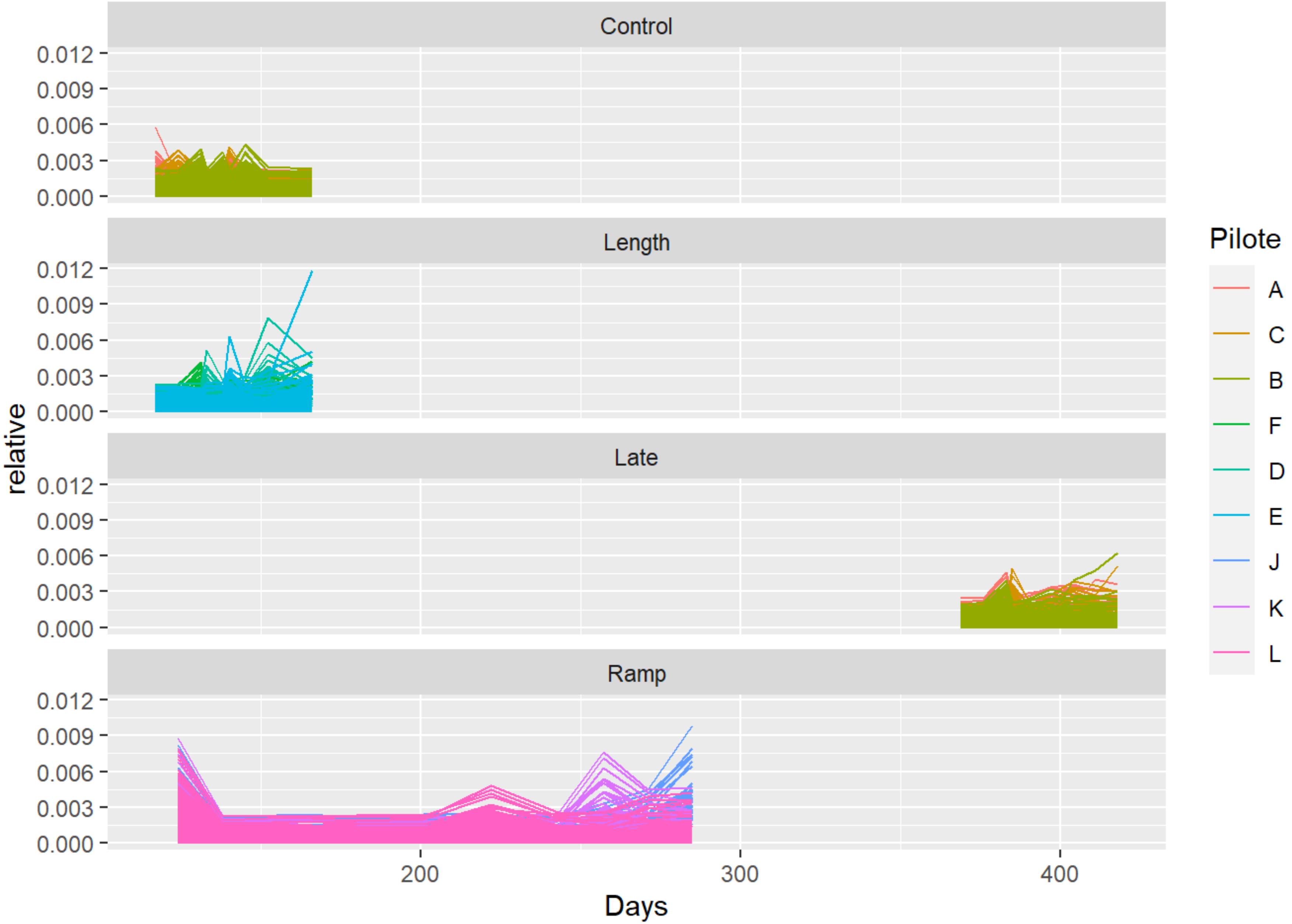
- 6 650 709 CDS
- 3 713 043 representative seq
- 2 767 548 emapper.annotations
- 7909 unique KEGG_ko



KO count per sample

- Relative count of each KO at each time point

- Statistical analysis to come:
    - regroup KOs with the same dynamical profile



Computations

Utils

- Snakemake workflow
  - Open to fork (and to advice) !
- Versionning on ForgeMIA
  - https://forgemia.inra.fr/cedric.midoux/nanosnake
- Running on MIGALE cluster
  - Obviously!
- Reporting with `workflowr`
  - Try it, it's nice! (but a little strict)

Computations

- Some steps required bigmem.q (2To RAM)

- Runtime = 10 days
  - Including 4.6 days for eggNOG mapper (on bigmem.q)
  - The assemblies require a lot of time (112 * few hours and 22h for coassembly) but they can be executed in parallel

- We tried to upgrade the basecalling to “high accuracy” but it’s really long (20 days for 4 runs and not over)

Current work

- Statistical issues
  - How to normalize counts?
- MAG/gene catalog cleaning and validation
- Data integration
  - … with physico-chemical measurements
- Workflow valorization? Standalone paper?