GenoToul Bioinfo

Analysis of shotgun metagenomic data

Claire Hoede & Philippe Ruiz

CC BY NC SA

MIA TOULOUSE

Contents day 1

- ❑ Tour de table
- ❑ Presentation of concepts and main tools
- ❑ TP on individuals tools

Contents day 2

- Main workflows
- Presentation of metagWGS
- Advantage of workflows manager and containers
- TP on metagWGS
- Next version of metagWGS
- The cluster's carbon footprint

Contents day 3

- ❑ Start cleaning your own data
- ❑ Consider the next steps in the analysis and adapt the configuration
- ❑ What's next ?

Contents day 1

- ❑ Tour de table
- ❑ Presentation of concepts and main tools (coffee break around 10h30)
- ❑ Lunch 12h00 – 13h00 (If you do not have a canteen badge, we will pay for the meal.)
- ❑ TP on individuals tools (coffee break around 15h00)

Plateforme Bioinfo Occitanie Toulouse

Tour de table

Tell me about your project and data:

- ❑ Who are you (name, laboratory) ?
- ❑ Which sequencing technology?
- ❑ Which type of environment? What diversity do you expect?
- ❑ How many samples? How many replicates, how many conditions? How many sequences?
- ❑ Which questions would you like to answer?

Contents day 1

- ❑ Tour de table
- ❑ Presentation of concepts and main tools (coffee break around 10h30)

- ❑ Lunch 12h00 – 13h00
- ❑ TP on individuals tools (coffee break around 15h00)

Genotoul Bioinfo

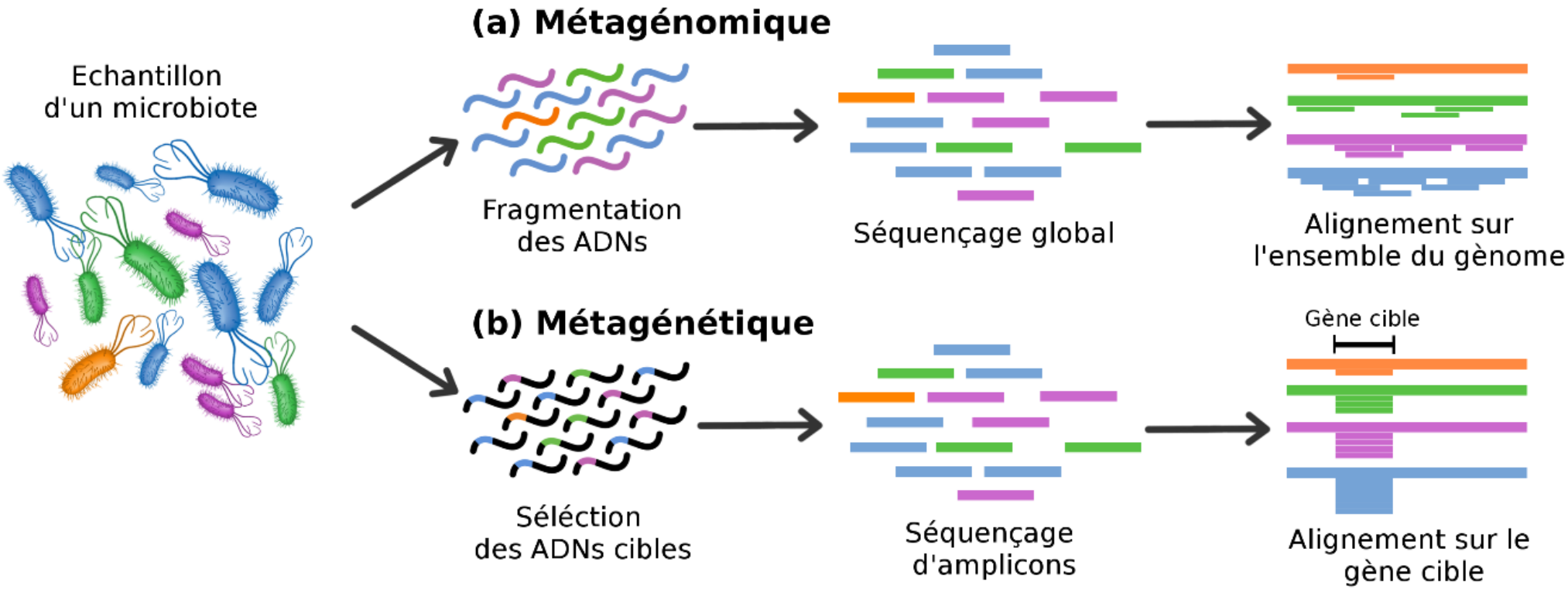
Questions ?

- Who is here ?
- What can they do ?
- What are they doing ?

Questions ?

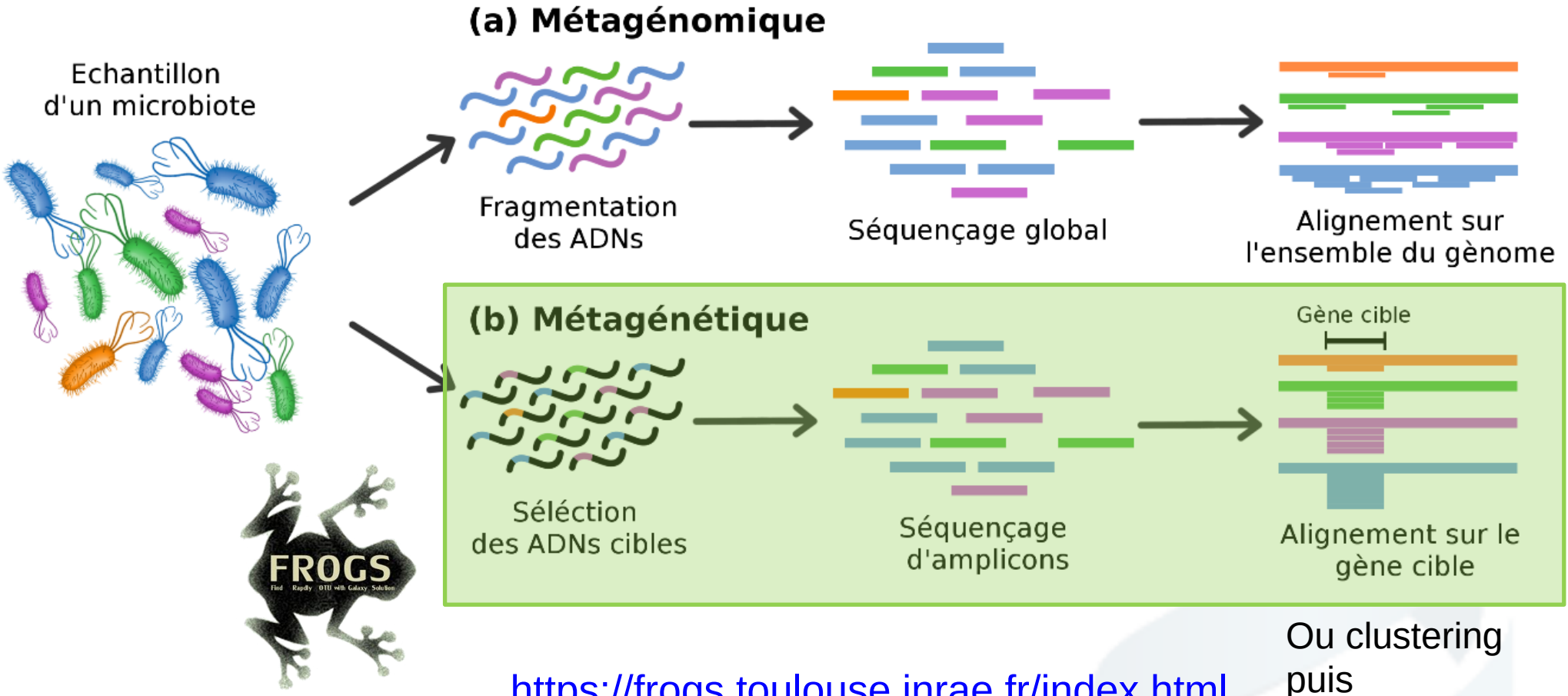
- ❑ Who is here ? ➔ metagenetics or metagenomics
- ❑ What can they do ? ➔ inference from metagenetics or metagenomics
- ❑ What are they doing ? ➔ metatranscriptomics, metaproteomics, metabolomics

Metagenetics or metagenomics ?



Ou clustering puis assignation taxonomique par homologie

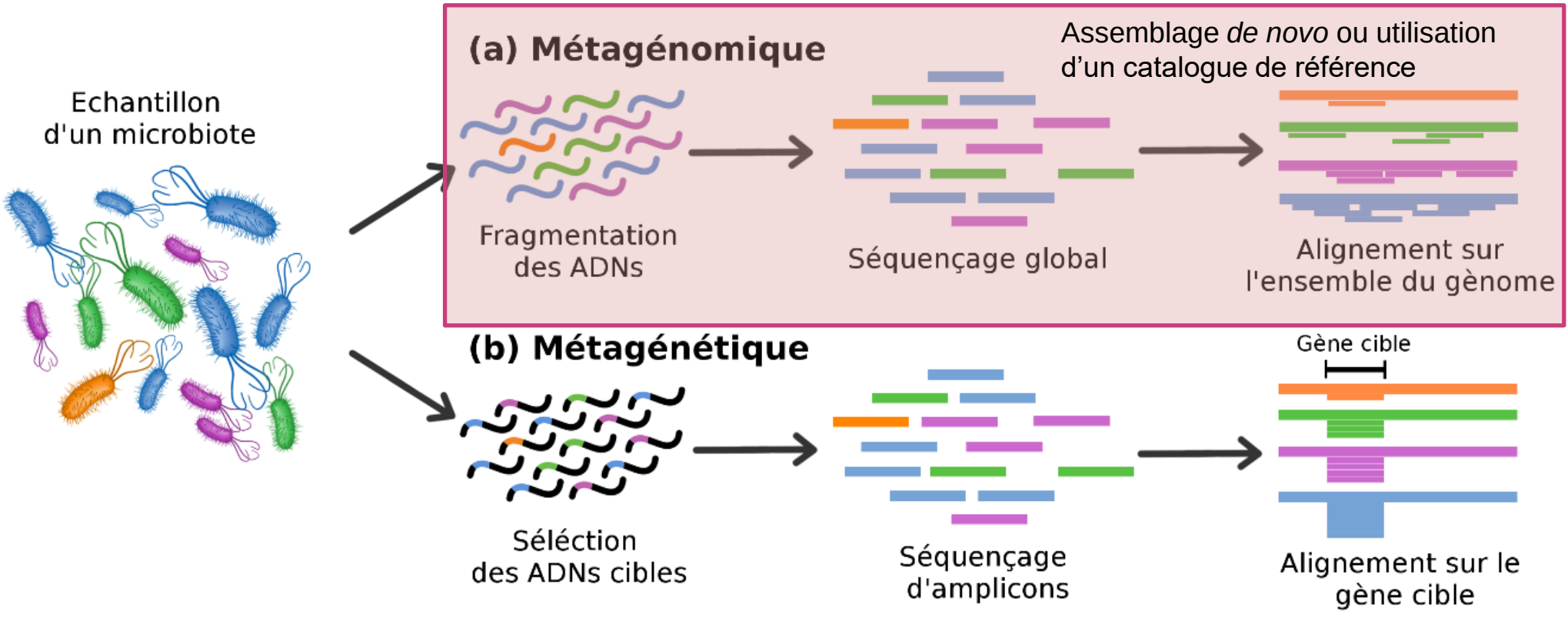
Metagenetics or metagenomics ?



https://frogs.toulouse.inrae.fr/index.html

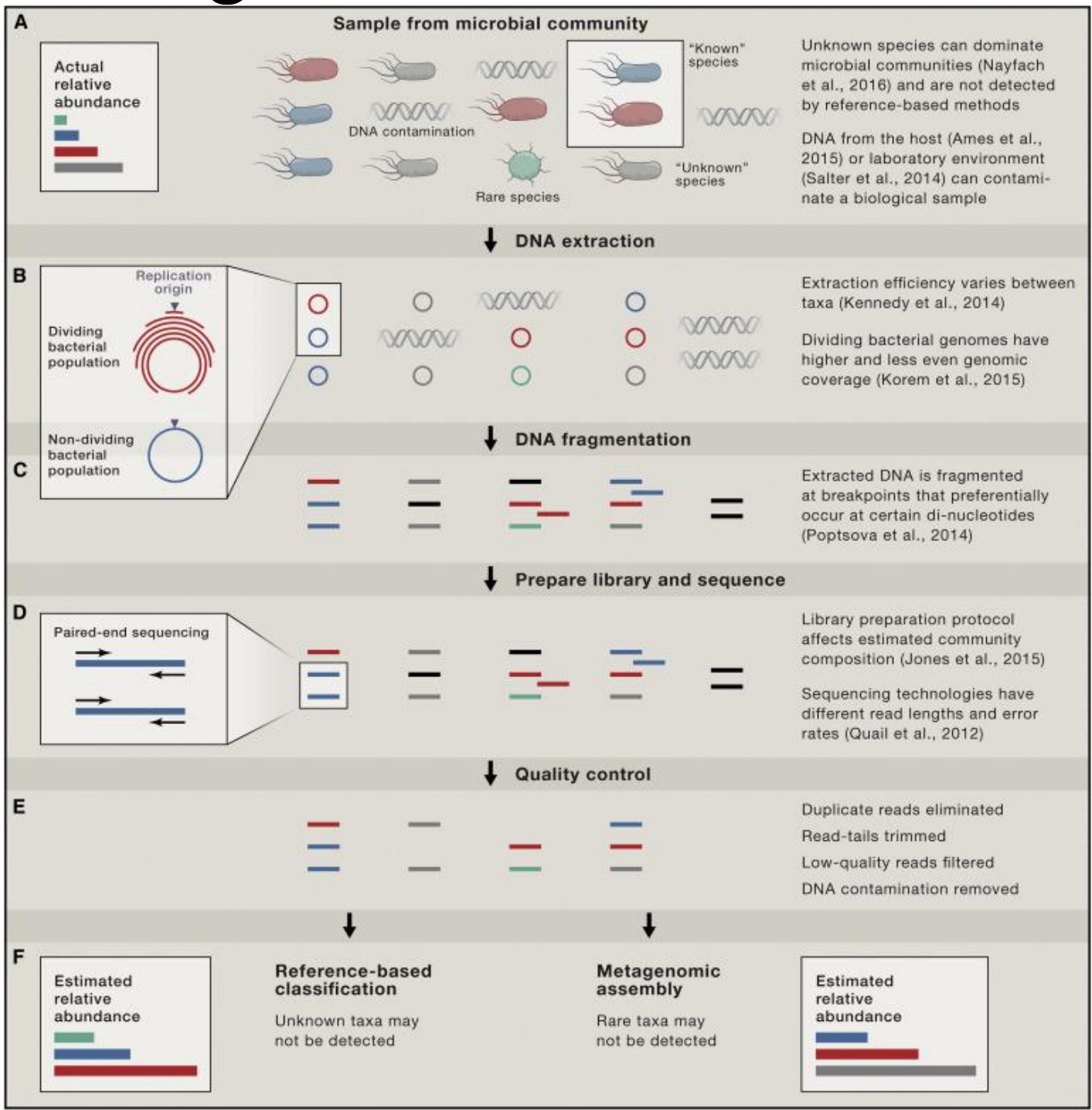
Ou clustering puis assignation taxonomique par homologie

Metagenetics or metagenomics ?



Plateforme Bioinfo Occitanie Toulouse

Metagenomics: data and bias



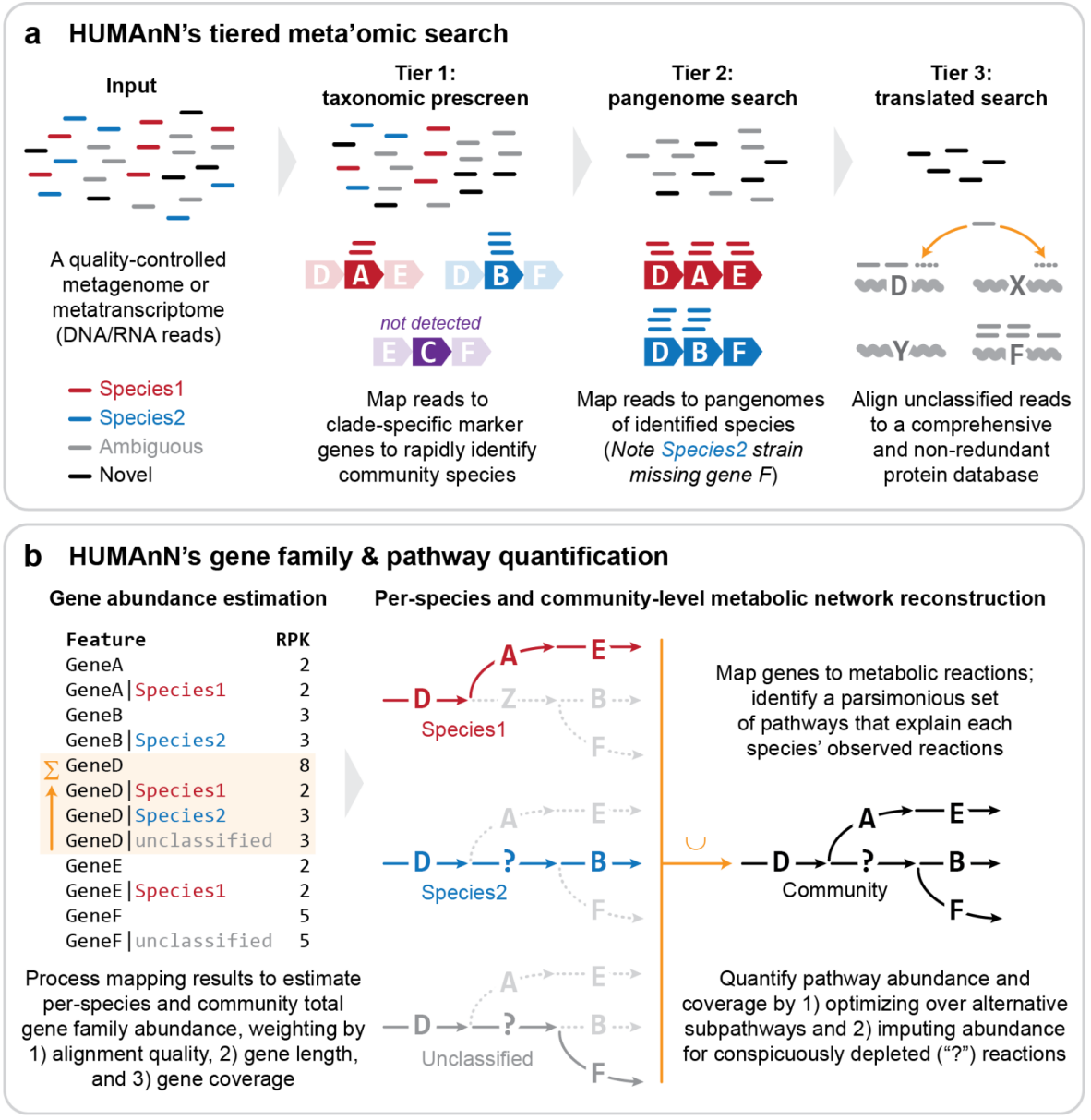
Nayfach, S., & Pollard, K. S. (2016).

Plateforme Bioinfo Occitanie Toulouse

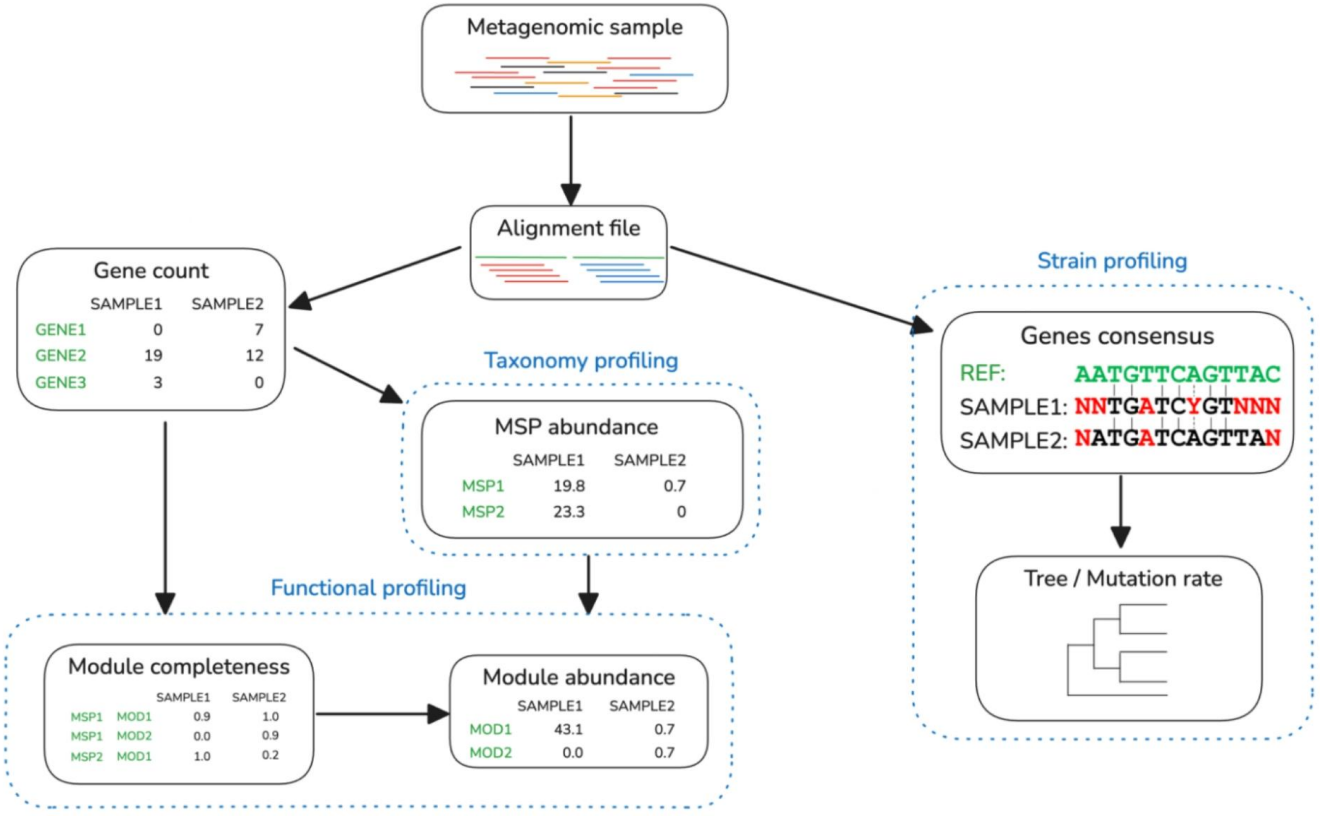
Approaches based on reference catalogues

- HUMAnN3: https://github.com/biobakery/humann

MetaPhlAn and ChocoPhlAn pangenome database
UniRef database provides gene family definitions
MetaCyc provides pathway definitions by gene family
MinPath is run to identify the set of minimum pathways
Bowtie2 is run for accelerated nucleotide-level searches
Diamond is run for accelerated translated searches



Approaches based on reference catalogues



https://doi.org/10.21203/rs.3.rs-6122276/v1

Metagenomic Species Pan-genomes (MSPs)
➔ Binning of genes based on co-abundance

❑ Meteor2: https://github.com/metagenopolis/meteor

- Download or build a reference catalogue
- Structure the raw fastq files
- Map reads against the reference catalogue (bowtie2)
- Compute taxonomical and/or functional abundances
- Strain profiling (SNP calling with freebayes)

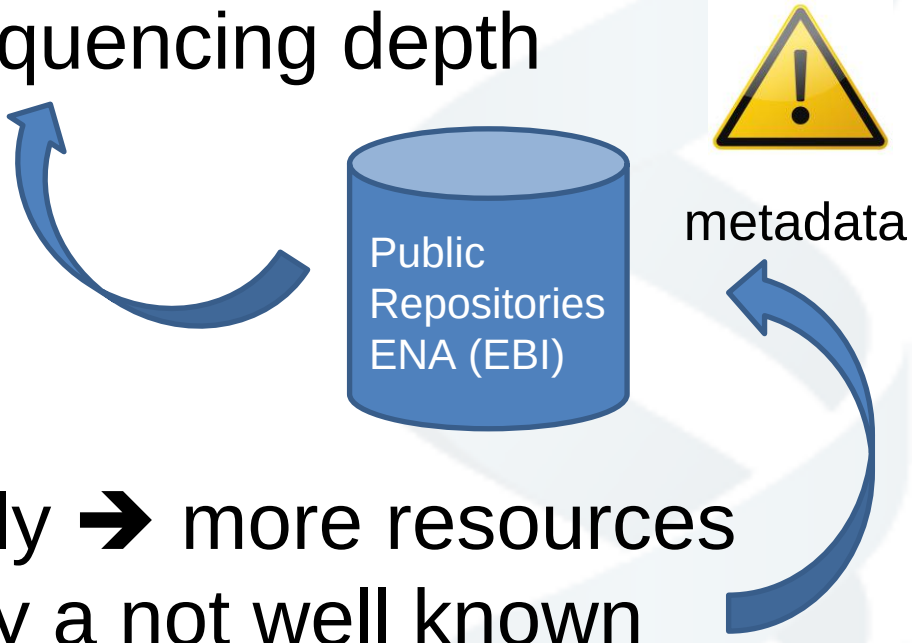
When to use what?

❑ Map on a reference ➔ fast, less resources consuming, when you study known environment and/or if you have a low sequencing depth

❑ Build a *de novo* assembly ➔ more resources consuming, when you study a not well known environment

When to use what?

❑ Map on a reference ➔ fast, less resources consuming, when you study known environment and/or if you have a low sequencing depth



❑ Build a *de novo* assembly ➔ more resources consuming, when you study a not well known environment

Caution image designed by Freepik

Plateforme Bioinfo Occitanie Toulouse

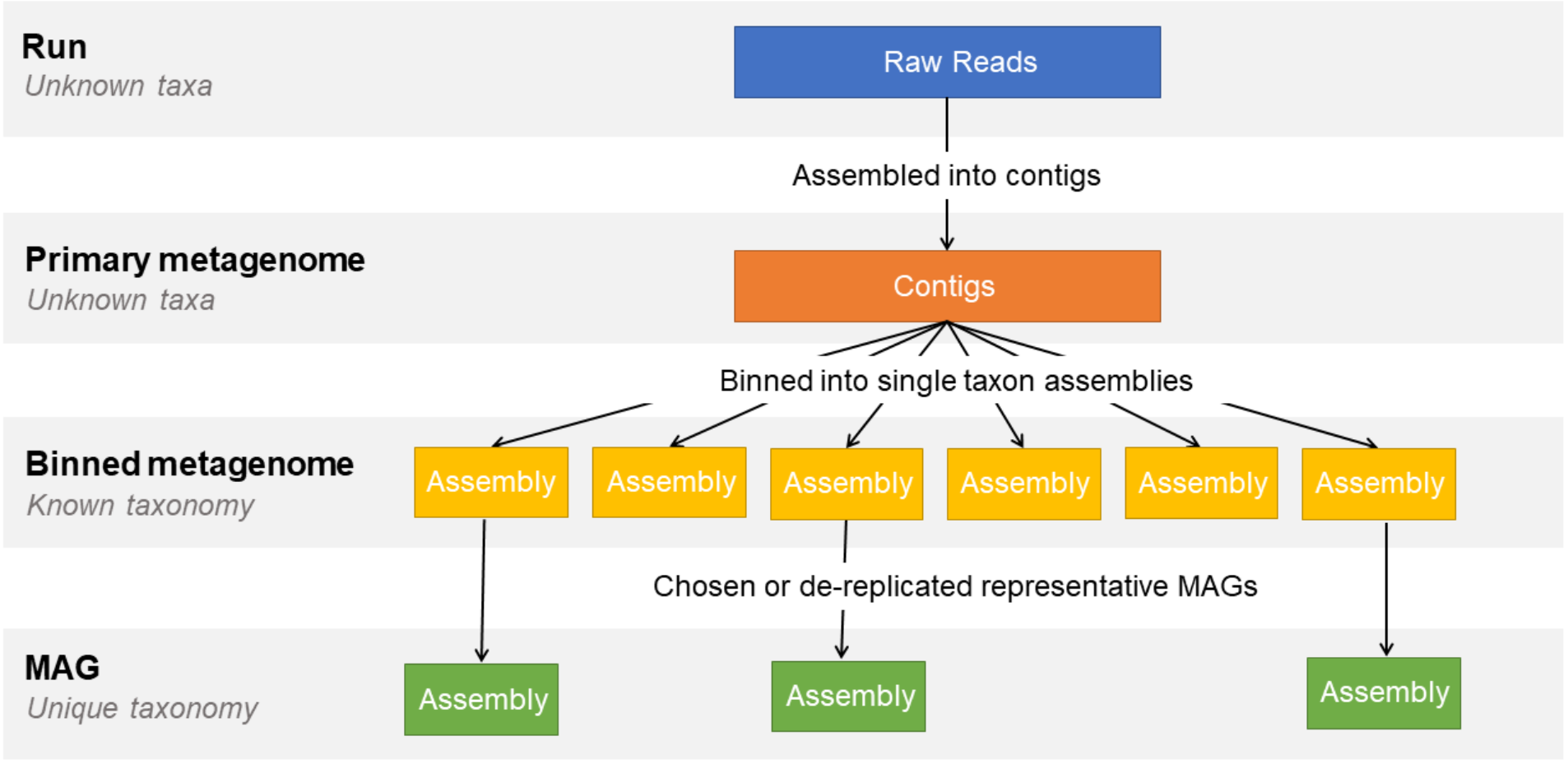
The challenge of meta-assembly

Genotoul Bioinfo

Plateforme Bioinfo Occitanie Toulouse

The challenge of meta-assembly

Vocabulary



https://ena-docs.readthedocs.io/en/latest/submit/assembly/metagenome.html

Main steps for *de novo* approach

1. Pre-process
2. Assembly and Alignment
3. Contigs annotation and quantification
4. Gene catalogue
5. Binning of contigs
6. Taxonomic and functional abundance matrices

Main steps for *de novo* approach

1. Pre-process
2. Assembly and Alignment
3. Contigs annotation and quantification
4. Gene catalogue
5. Binning of contigs
6. Taxonomic and functional abundance matrices

Short / long (HiFi) if specific tool
Quality check:
- fastQC
Remove adapters:
- cutadapt, fastp …
Trim bases on quality:
- sickle / Smrtlink (Pacbio, lors du séquençage)
Taxonomic composition from reads:
- kraken2, metaPhlan4, Kaiju… / Megan-LR (Huson et al. 2018), Pb-metagenomics-tools
Remove contaminating sequences (bwa-mem2, bowtie2, minimap2)

Main steps for *de novo* approach

Pre-process 1

Assembly and Alignment 2

Contigs annotation and quantification 3

Gene catalogue 4

Binning of contigs 5

Taxonomic and functional abundance matrices 6

Assembly:

- De Bruijn graph ==> potential chimeras
  - MetaVelvet, IDBA-UD, MetaSPAdes, Megahit…
  - MetaFlye (ONT too), Hifiasm-meta, HiCanu, MetaMDBG ...

Alignment:

- bwa mem2 or bowtie2 / minimap2 against genomes or genes

Main steps for *de novo* approach

Pre-process 1

Assembly and Alignment 2

Contigs annotation and quantification 3

Gene catalogue 4

Binning of contigs 5

Taxonomic and functional abundance matrices 6

Structural annotation:
- Prokka (Prodigual), FragGeneScan...

Quantification:
- featureCount, HtSeqCount...

Functional annotation:
- diamond (homology)
- Interproscan, EggNOG (HMM protein domains and families),

⇒ eggNogMapper (Cantalapiedra *et al.*, 2021)
- KEGG, metaCyc (metabolic pathways...)
- COG (cluster of orthologs genes)

Main steps for *de novo* approach

Pre-process 1

Assembly and Alignment 2

Contigs annotation and quantification 3

Gene catalogue 4

Binning of contigs 5

Taxonomic and functional abundance matrices 6

Proteins clustering:

- CD-Hit, mmseq2 ….

Main steps for *de novo* approach

Pre-process

Assembly and Alignment

Contigs annotation and quantification

Gene catalogue

Binning of contigs

Taxonomic and functional abundance matrices

Binning of contigs:

- drafts genomes
- CONCOCT, MetaBat2, MaxBin2, solidBin, Vamb, semibin2 use kmer, depth, marker genes (Alneberg et al, 2014, Wu et al, 2015, Kang et al., 2019, Wang et al. 2019, Nissen et al. 2021, Pan et al. 2023)
- Launch several binners and combine them with DASTool or binning_refiner (Song et al, 2017 ; Sieber et al, 2018 ; bin_refinment de metawrap Uritsky et al. 2018) ; Binette (Mainguy et al. 2024) etc
- Quality of bins (CheckM2, Chklovski *et al.*, 2022)
- dRep to choose the best bins among all samples (Olm *et al.*, 2017) ⇒ one set de bins for all samples ⇒ MAGs (Metagenome Assembled Genome)

Main steps for *de novo* approach

Pre-process 1

Assembly and Alignment 2

Contigs annotation and quantification 3

Gene catalogue 4

Binning of contigs 5

Taxonomic and functional abundance matrices 6

Taxonomic affiliation (genes, contigs) :
- diamond : homology suivi d'un script algo LCA (lowest common ancester) ex CAT et BAT (Bastiaan von Meijenfeldt *et al.* 2019) **+ quantification**

Taxonomic affiliation (bins):
- Gtdb-tk (Chaumeil *et al.*, 2022) **+ quantification**

Functional abundance:
- Orthologues (Kegg orthologie, COG, NOG)
- Pathways profiles (Kegg or MetaCyc pathways or GO terms)
- Cluster of genes profiles

+ quantification

Contents day 1

- ❑ Tour de table
- ❑ Presentation of concepts and main tools (coffee break around 10h30)

- ❑ Lunch 12h00 – 13h00
- ❑ TP on individuals tools (coffee break around 15h00)

TP1 – individual tools

Statement:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP_1_metaG
Correction:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP_1_metaG_corrected

What did you think of this approach tool by tool?

- What are the benefits?

- Disadvantages?

What did you think of this approach tool by tool?

- What are the benefits?

➢ You control every step and every parameter
➢ You can check each step and each output before continuing

- Disadvantages?

➢ not very efficient (slow)
➢ difficult to trace what has been done
➢ frequent human errors

End of the day

- Thank you for your attention

- See you tomorrow at 9.00 am

Contents day 2

- ❑ A recap of yesterday's action (type of data, main steps …)
- ❑ Automation and reproducibility (coffee break around 10h30)

- ❑ Lunch 12h00 – 13h00
- ❑ TP on metagWGS (coffee break around 15h00)

- ❑ Next version of metagWGS
- ❑ The cluster's carbon footprint

Genotoul Bioinfo

What do you remember from yesterday ?

Genotoul Bioinfo

To automate: workflows

- ❑ Build the commands
- ❑ Organize output files (naming, directories....)
- ❑ Run them in parallel on the cluster
- ❑ Enables error recovery (will only restart what has not been completed)
- ❑ Generally based on containers, which freeze dependencies (and improve reproducibility)

Plateforme Bioinfo Occitanie Toulouse

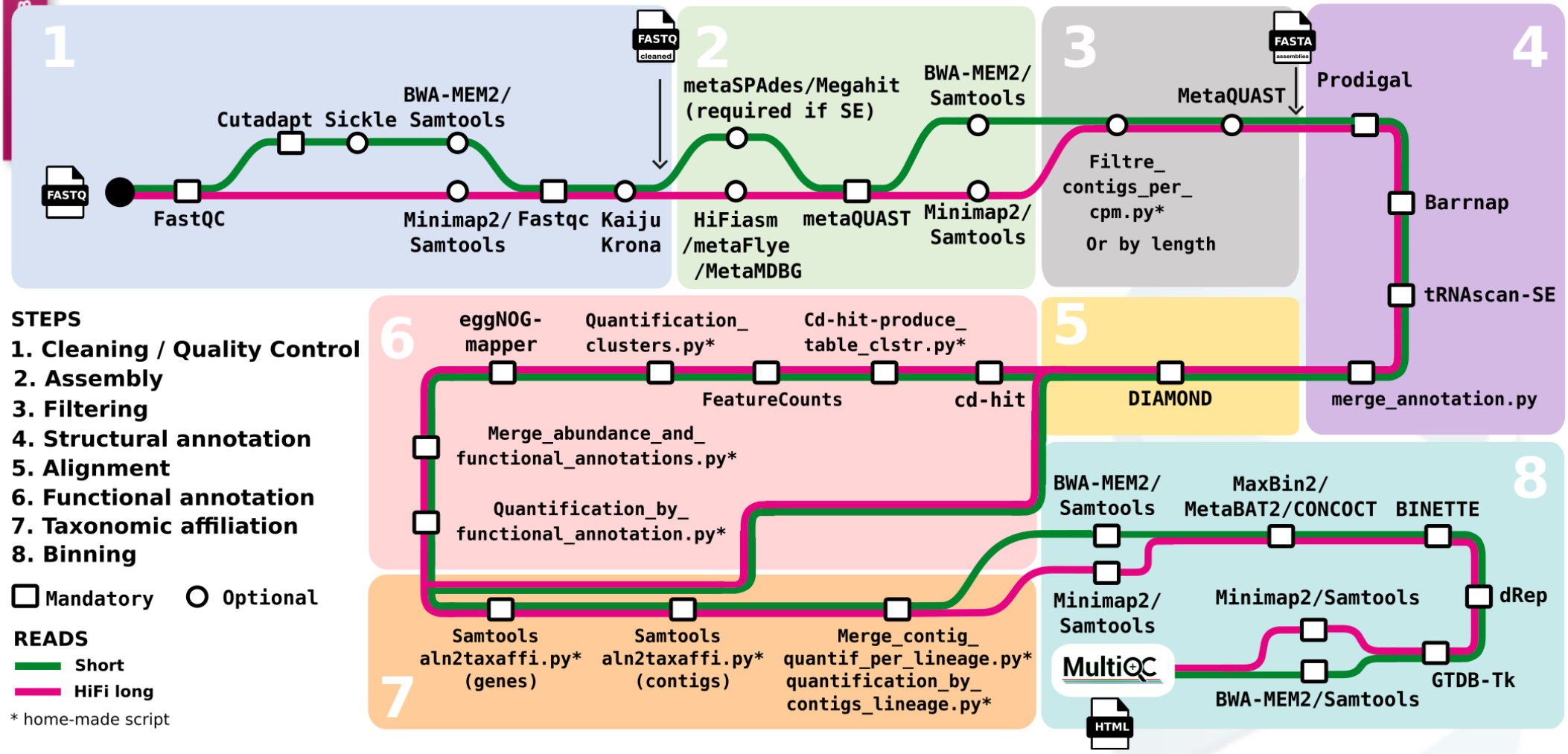
Genotoul Bioinfo

Metagenomics workflows

Level	Reads		Genes in all contigs	All contigs		MAGS / bins	
Tools	**taxo**	**function**	**taxo**	**taxo**	**genes function**	**taxo**	**genes function**
MAG (nf-core) Hybrid assembly possible	YES	NO	NO	NO	NO	YES	YES (but not rRNA and tRNA)
Metawrap	YES	NO	NO	YES	NO	YES	YES
VEBA 2.0 (short and long reads: ONT & Pacbio)	NO	Only with bins found	NO	NO	NO	YES	YES
Atlas	NO	NO	NO	NO	YES	YES	YES
Anvi'o Metagenomic workflow	YES	NO	NO	NO	YES	YES	YES
HiFi-MAGs-pipeline (**HiFi** reads only, binning only)	NO	NO	NO	NO	NO	YES	NO
metagWGS (short_reads & **HiFi**)	**YES**	NO	**YES**	**YES**	**YES**	**YES**	**YES (via contigs)**

Bioinfo Occitanie Toulouse

metagWGS



STEPS

1. Cleaning / Quality Control
2. Assembly
3. Filtering
4. Structural annotation
5. Alignment
6. Functional annotation
7. Taxonomic affiliation
8. Binning

□ Mandatory ○ Optional

READS

Short

HiFi long

* home-made script

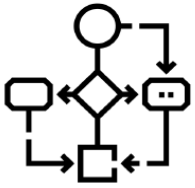
Plateforme Bioinfo Occitanie Toulouse

metagWGS

Type of NGS data:
Whole genome shotgun sequencing

Short reads, paired **or single (New in v. 2.6.0)**

PACBIO HiFi reads, single-ends



Workflow:
A scalable and reproducible metagenomics analysis with nextflow pipeline using singularity containers

Fully documented
https://forge.inrae.fr/genotoul-bioinfo/metagwgs
https://genotoul-bioinfo.pages.mia.inra.fr/metagwgs/master/index.html

Genotoul Bioinfo

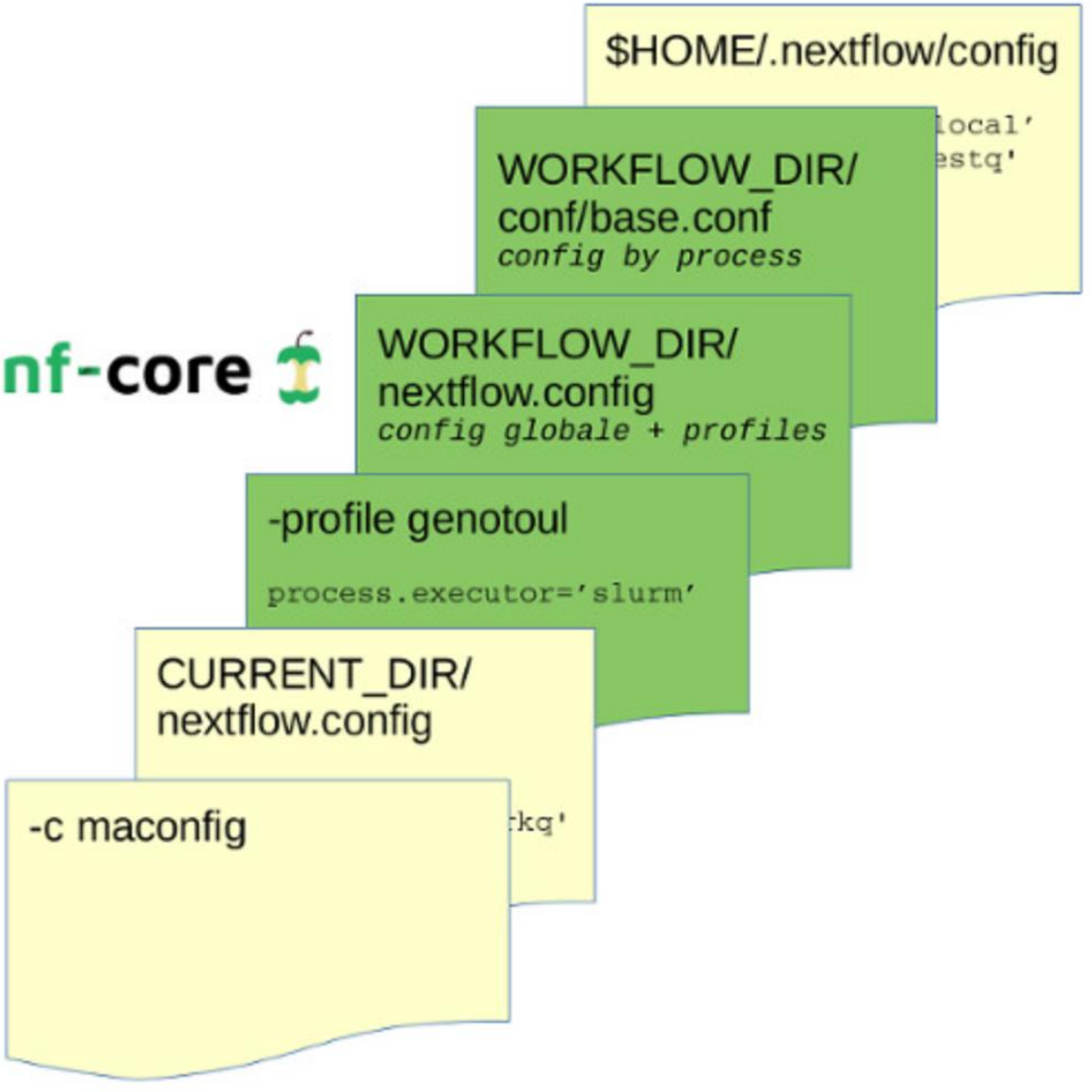
nextflow

- Develop by CRG (Centre for Genomic Regulation) Barcelona
- In Groovy
- Parallelization on multiple infrastructure (Cloud, HPC, local…)
- Used to restart after an error and/or correction. Does not restart what has been completed successfully and what is not affected by the modification.
- Few manual configurations

Genotoul Bioinfo

nextflow

- Config files



❑ Outputs

```
$ ls -la
drwxr-xr-x  3 bleuet BIOINFO  4096 23 dec.  12:11 .nextflow
=> internal files
-rw-r--r--  1 bleuet BIOINFO   108 23 dec.  12:07
nextflow.config  => config file
-rw-r--r--  1 bleuet BIOINFO 10962 23 dec.  12:11 .nextflow.log
=> log file with all intermediate directories
drwxr-xr-x  2 bleuet BIOINFO  4096 23 dec.  12:11 results
=> final results
drwxr-xr-x  8 bleuet BIOINFO  4096 23 dec.  12:10 work
=> working and temporary files
```

Genotoul Bioinfo

nextflow

❑ Work directory

```
$ ls -la work/40/944143ebbcc45aa0e4bf2f8ba9dab6/
total 4
drwxr-xr-x 3 bleuet BIOINFO 4096 24 mars  10:00 ..
-rw-r--r-- 1 bleuet BIOINFO 2514 24 mars  10:00 .command.run
-rw-r--r-- 1 bleuet BIOINFO   36 24 mars  10:00 .command.sh
=> the command to execute
-rw-r--r-- 1 bleuet BIOINFO    0 24 mars  10:00 .command.begin
-rw-r--r-- 1 bleuet BIOINFO    0 24 mars  10:00 .command.err
=> error file
-rw-r--r-- 1 bleuet BIOINFO   13 24 mars  10:00 .command.log
-rw-r--r-- 1 bleuet BIOINFO   13 24 mars  10:00 .command.out
=> output file
drwxr-xr-x 2 bleuet BIOINFO 4096 24 mars  10:00 .
-rw-r--r-- 1 bleuet BIOINFO    1 24 mars  10:00 .exitcode
```

- Result directory:

The organisation of output files will depend on the workflow.

❑ Useful options:

Option	Description
-resume	It allows to rerun metagWGS from the lastest process uncorrectly ended or from a process where input or output files have changed.
-with-report	Generates a report.html file describing the use of memory and cpus for each process.
-with-timeline	Generates a timeline.html file describing the duration of each process.
-with-trace	Generates a trace.txt file describing the location of cache directory and metrics for each process.
-with-dag	Generates a dag.dot file, a graph representing the pipeline.

Containers

- **A container** allows you to run one or more Linux applications in an isolated, reproducible environment that depends only on the Linux kernel of the machine you are running. A container is similar to a virtual machine, except that it does not necessarily have a complete operating system on board, which means that it can be launched in a few seconds and is lighter.

- **Singularity / Apptainer**: The initial aim is to offer a containerisation solution tailored to the needs of scientists who need to run containerised applications on computing clusters (HPC). Unlike other container systems (such as Docker), Singularity requires no administrator rights, no daemons, does not virtualise the network and talks directly to its host's file system. Each container is launched and stopped at the same time as the application it encapsulates.

Singularity/Apptainer: vocabulary

- **Image**: As with virtual machines, an "image" is a static description of a container, a sort of photograph of a machine, which you can exchange with your collaborators, and from which you can instantiate and run containers. Singularity has its own image mechanism, but can also interface with Docker images.

- **Container**: A lightweight, memory-loaded virtual machine used to run an application within an isolated, reproducible environment. A container is instantiated from an image.

- **Registry**: Warehouse where ready-to-use images are stored. Singularity's official central registry can be consulted on the web at https://singularity-hub.org/

TP2 – metagWGS

Statement:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP_2-MetagWGS-on-a-very-small-dataset

Correction:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP_2-MetagWGS-on-a-very-small-dataset---correction

Assembly metrics

Consider 9 contigs (size 10, 9, 8, 7, 6, 5, 4, 3, 2),

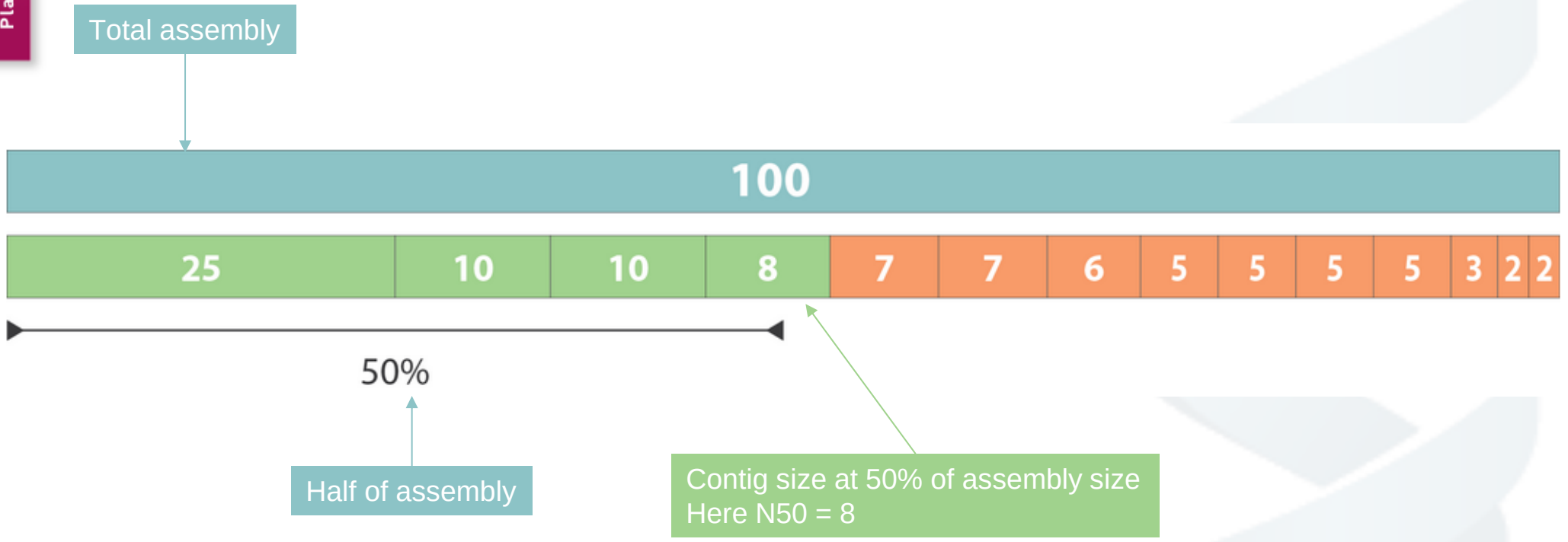
assembly length = 54.

50% of assembly = 27

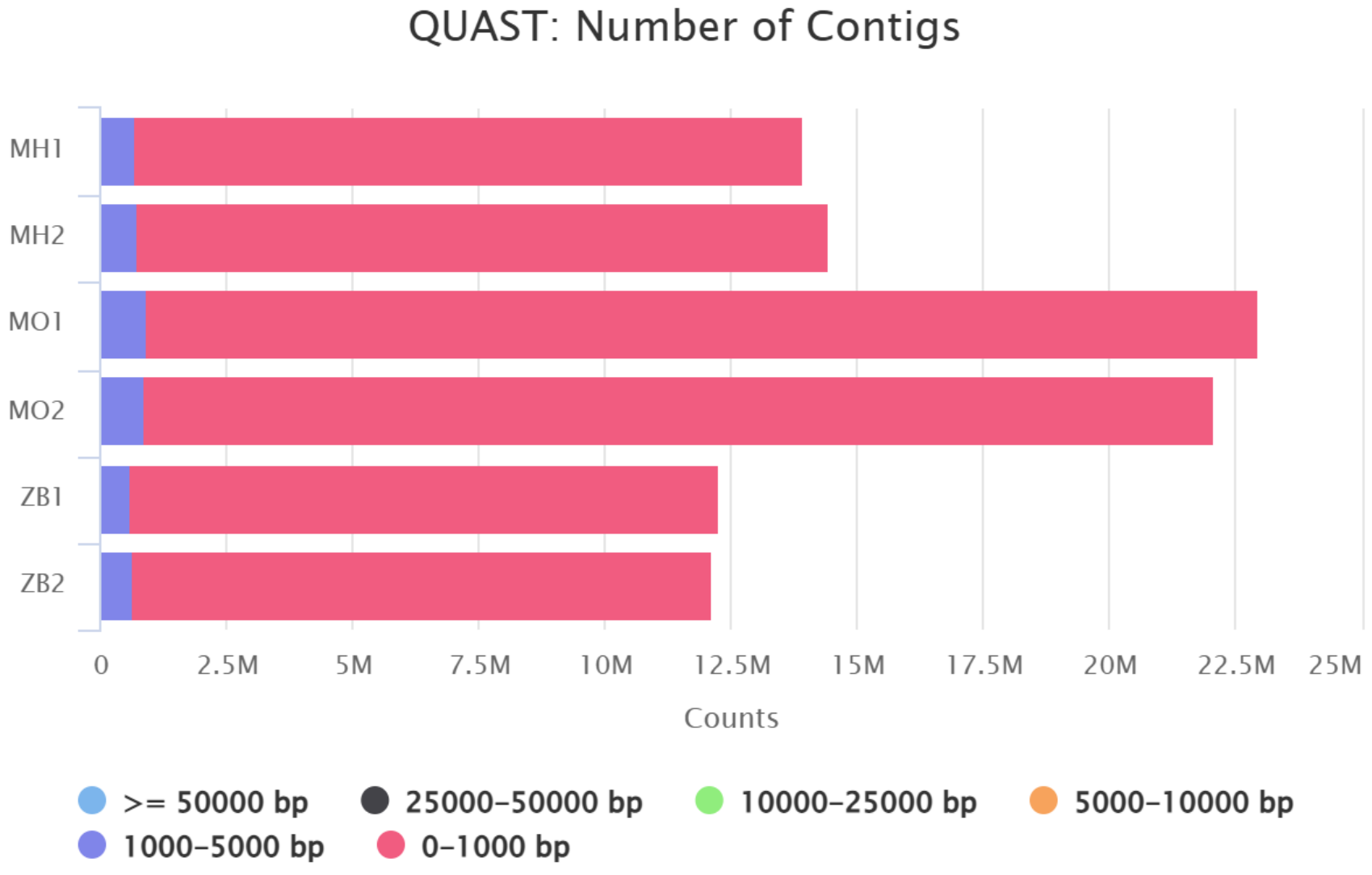
10 + 9 + 8 = 27.

N50 = 8; L50 = 3

N50 ?


Total assembly
100
25
10
10
8
7
7
6
5
5
5
5
3
2
2
50%
Half of assembly
Contig size at 50% of assembly size
Here N50 = 8

Filter the assembly ?



Unfiltered assemblies on soil samples (%mapped reads between 68% and 80%)

Plateforme Bioinfo Occitanie Toulouse

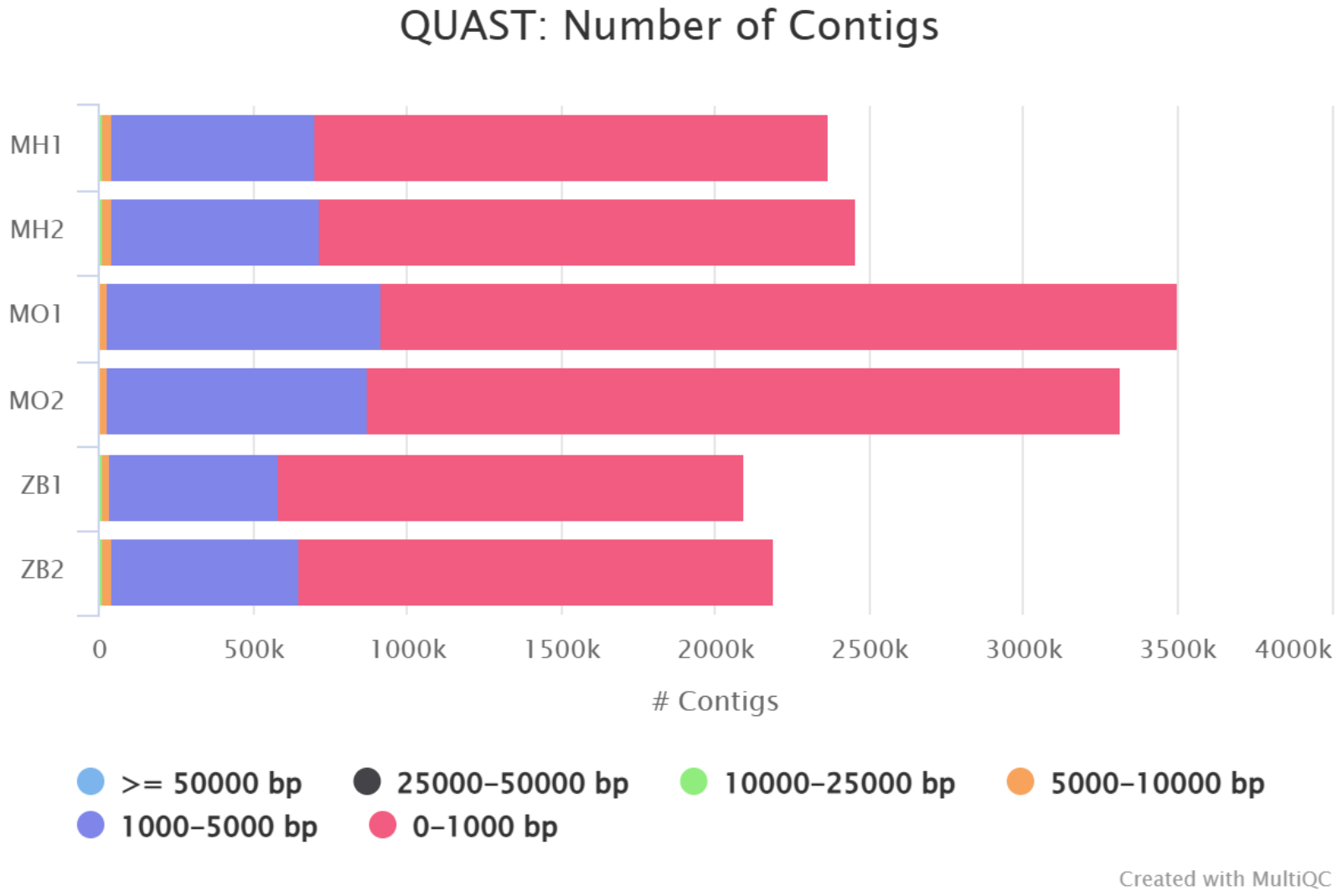
Filter on cpm or with contigs length ?

In metagWGS, default assembly filter is with contigs length in step 3 (to be prioritised in the case of highly fragmented assembly).

It’s possible to use a cpm threshold instead (useful for beautiful assembly).

Plateforme Bioinfo Occitanie Toulouse

Filter on cpm or with contigs length ?



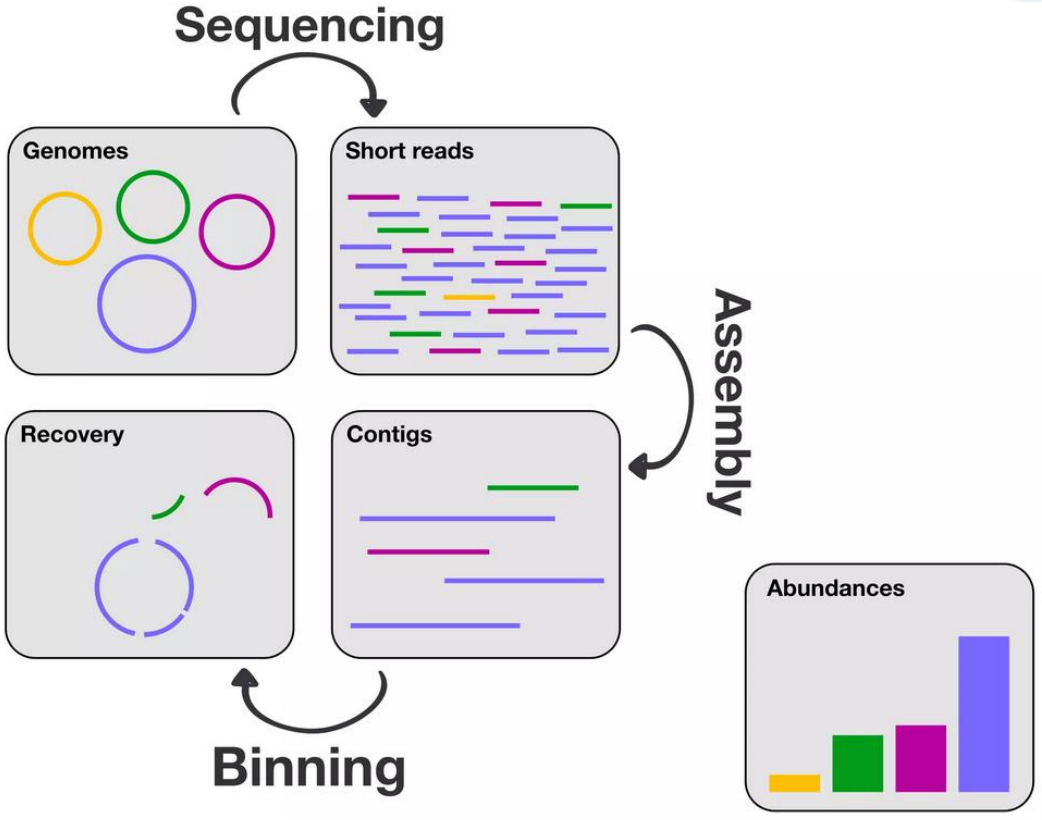
Filter by contigs length > 500pb, but be aware to the quantity of mapped reads (between 53 – 70%).

Back to the TP2

Binning ?

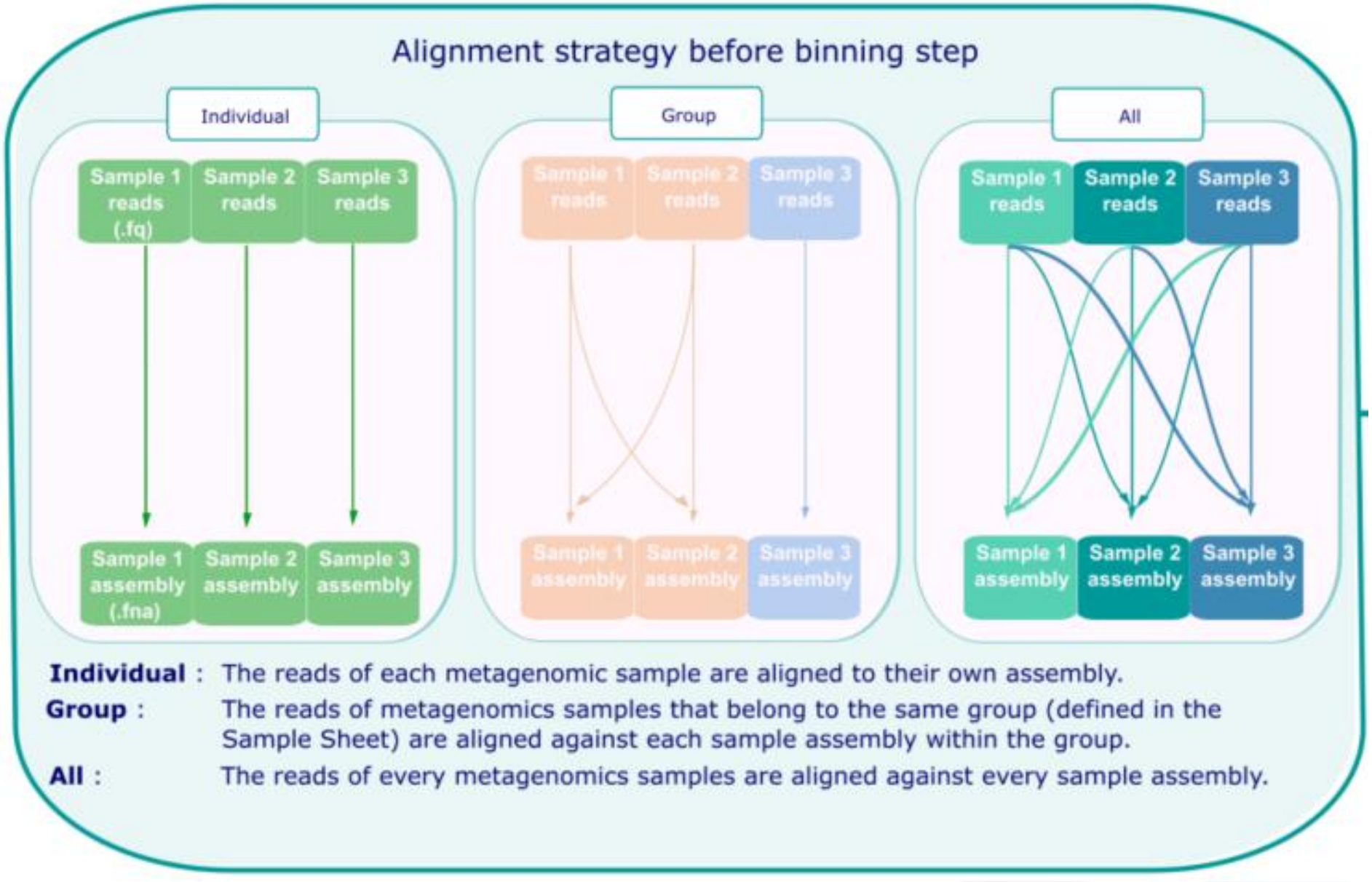
In metagenomics, **binning** is the process of **grouping reads or contigs** and assigning them to **individual genome**.

Binning methods can be based on either **compositional features** or alignment (**similarity**), or **both and co-abundances**.



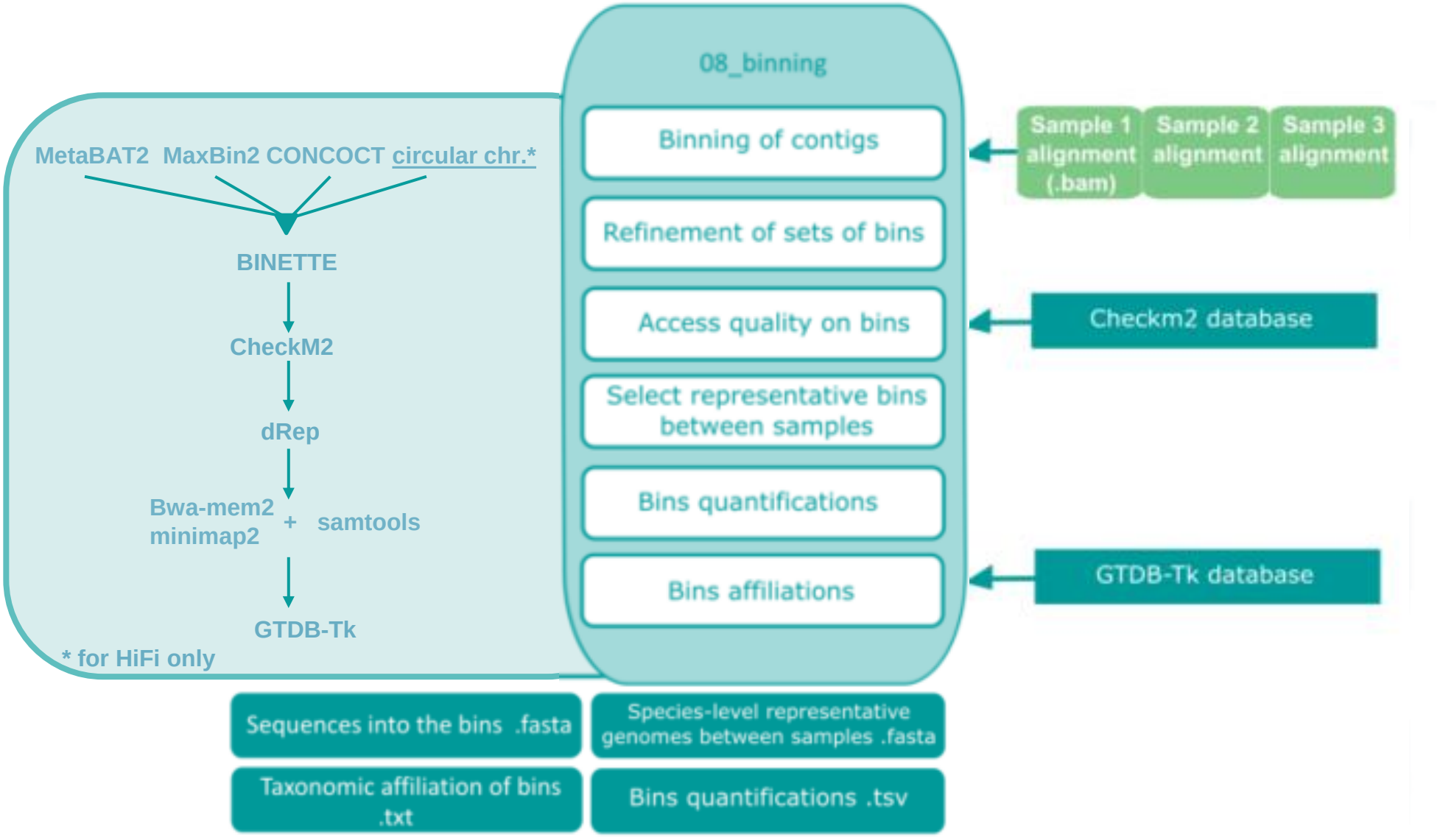
Ref: https://www.slideshare.net/AMuratEren/intro-to-metagenomic-binning

The binning : alignment



Plateforme Bioinfo Occitanie Toulouse

The binning of metagWGS



Back to the TP2

What will happen in the new version of metagWGS?

- Improve MultiQC report
- Use bowtie 2 instead bwa mem 2
- Use minimum coverage threshold for gene and MAGs to count reads
- Scheduled for end of June

- In the longer term, the plan is to :

- Ability to deal with ONT reads
- Deal with micro-eukaryotes genomes
- Improve binning strategie

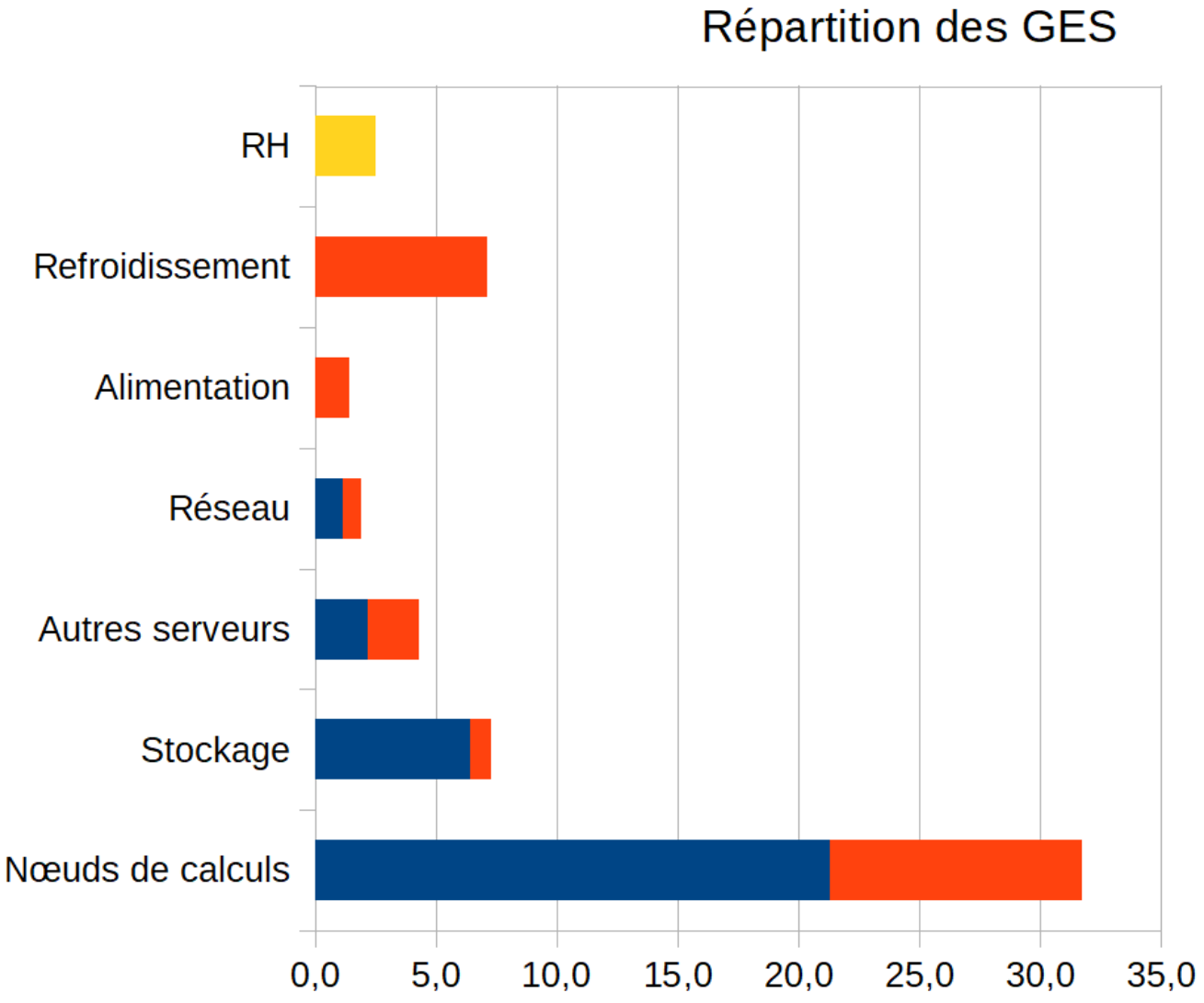
The cluster's carbon footprint (scope)

- Computing cluster acquired in 2023/2024:
  - Compute node
  - Storage servers (temporary Work)
  - Other servers: Frontends, Monitoring, Website, etc.
  - Network equipment (switch)
  - Power equipment (PDU)
  - 2.7 FTE
- Footprint taken into account:
  - Electricity consumption
  - Manufacturing
  - Transport
  - EOL

The cluster's carbon footprint (scope)

- Not taken into account
  - Save space
  - Refrigerant gas (lack of information)
  - Generators and their consumption
  - Building manufacturing (DROCC)
- Parameters
  - Lifetime 7 years
  - PUE 1.5 (Power Usage Efficiency)
  - Energy mix emission factor: 0.06 kgCO2e/kWh (Base Carbone V19; source year 2020)
  - Effective calculation hours (2024): 15,709,439

The cluster's carbon footprint (results)



3.6 gCO2eq / hCPU

555 555h de calcul = 2t CO2eq

The cluster's carbon footprint (results)

- Display on connection screen

```
============================================================
Informations sur le compte choede (02-05-2025 04:05)
============================================================
[redacted]

CO2 equivalent : .5436 kg (~ 0.0% of your carbon budget of 2000 kg CO2e per year and per human being)
(CPU hours * 3.6 g CO2e. See https://hal.science/hal-02549565v5 for more details)
```

The cluster's carbon footprint (results)

- Not taken into account
  - Save space
  - Refrigerant gas (lack of information)
  - Generators and their consumption
  - Building manufacturing (DROCC)
- Parameters
  - Lifetime 7 years
  - PUE 1.5 (Power Usage Efficiency)
  - Energy mix emission factor: 0.06 kgCO2e/kWh (Base Carbone V19; source year 2020)
  - Effective calculation hours (2024): 15,709,439

Contents day 3

- ❑ Quick shared reminders
- ❑ Start cleaning your own data or data I provide if you do not have your own data (coffee break around 10h30).

- ❑ Lunch 12h00 – 13h00
- ❑ Consider the next steps in the analysis and adapt the configuration. Think to the best strategy to choose. (coffee break around 15h00)

- ❑ What's next ?

Genotoul Bioinfo

What do you remember from yesterday ?

Data provided: 2 choices

❑ 11 HiFi gut human samples publicly available

❑ Home made dataset (from migale) simulated

- 5 samples
- 1M or 2M reads per sample (best strategy : co-assembly)
- 50 Bacteria, 10 Archaea, 4 Viruses
- Illumina Hiseq 2x125 bp
- Uniform or log-normal abundance

TP3 – metagWGS

Statement:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP-3

Correction:
https://forge.inrae.fr/genotoul-bioinfo/metagwgs/-/wikis/TP_3-correction-tips

While the analyses are running, please remember to keep answering the questions in Practical 2

TP3 – metagWGS

See together the multiQC for simulated data

- % reads mapped
- % contigs in bins
- Virus are not in bins but in contigs
- etc

What's next ?

I propose to schedule 2 videoconference dates for 1h30 to answer your questions and help you if necessary. Please fill in the following survey:
https://evento.renater.fr/survey/visio-d-accompagnement-sur-le-traitement-de-vos-donnees-m8dme7lu
If possible, send me questions three working days in advance.

What's next ?

I need you to improve this training.
Please fill this satisfaction survey with a lot of comments:
https://sondages.inrae.fr/index.php/84236?lang=fr

Plateforme Bioinfo Occitanie Toulouse

Thank you for your participation !

You can contact me by e-mail:
claire.hoede[@]inrae.fr