metagWGS: a nextflow workflow to analyze metagenomic data

Joanna Fourquet, Céline Noirot, Pierre Martin, Jean Mainguy, Géraldine Pascal et **Claire Hoede**

Journées du PEPI IBIS – 16/18 novembre 2021

The samples we want to analyze

https://oceans.taraexpeditions.org/

Marine sample

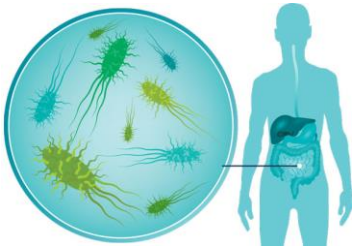


Image: newannyart/Thinkstock
https://www.health.harvard.edu

Gut sample

https://www6.inrae.fr/isite-agroecologie-bfc/

Soil sample

Environmental samples

The biological questions we want to answer

https://oceans.taraexpeditions.org/

Marine sample

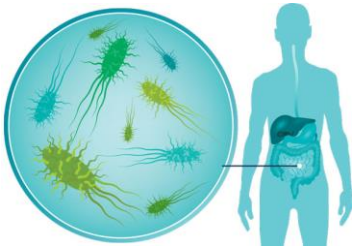


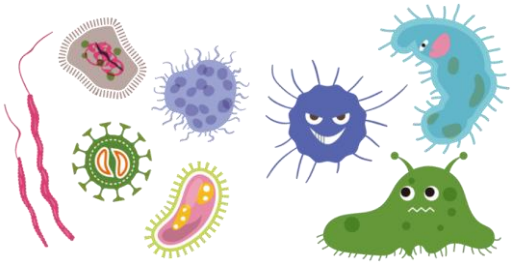
Image: newannyart/Thinkstock
https://www.health.harvard.edu

Gut sample

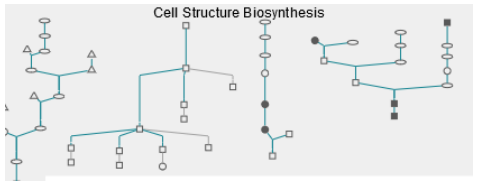
https://www6.inrae.fr/isite-agroecologie-bfc/

Soil sample

Who is there?



What genes are present in our samples?
What are their functions?



Metagenomics data: from sample to files

Samples

DNA extraction
+
Fragmentation
+
Library preparation

Whole genome DNA samples

Sequencing

Files of sequences = raw data

Whole genome shotgun sequencing data 2x150bp (Illumina)

? ? ? Data analysis ? ? ? ? ? ?

?

Metagenomics data: from sample to files

Samples

DNA extraction
+
Fragmentation
+
Library preparation

Whole genome DNA samples

Sequencing

Files of sequences = raw data

Whole genome shotgun sequencing data 2x150bp (Illumina)

Bioinformatics

Data analysis

Statistics

Bioinformatics analysis:
- Output files
- Output graphs

Biostatistical analysis

Metagenomics data: from sample to files

Samples

DNA extraction
+
Fragmentation
+
Library preparation

Whole genome DNA samples

Sequencing

Files of sequences = raw data

Whole genome shotgun sequencing data 2x150bp (Illumina)

Bioinformatics

Data analysis

Statistics

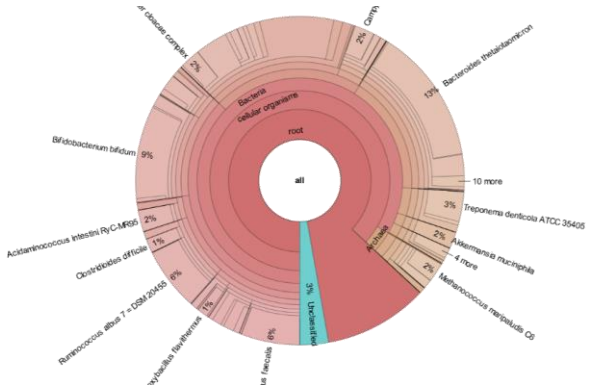
Bioinformatics analysis:
- Output files
- Output graphs

Biostatistical analysis

Definition of bioinformatics output data

Who is there?

Taxonomic affiliation of reads



Taxonomic affiliation (gene, contig)

Sequence id	consensus taxid	consensus lineage
Sequence 1	210	cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter; Helicobacter pylori
Sequence 2	1681	cellular organisms; Bacteria; Terrabacteria group; Actinobacteria; Actinobacteria; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; Bifidobacterium bifidum
Sequence 3	1358	cellular organisms; Bacteria; Terrabacteria group; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae; Lactococcus; Lactococcus lactis

What genes are present in our samples?
What are their functions?

Read quantification by gene

Gene	Sample 1	Sample 2
Gene 1	844	887
Gene 2	847	891
Gene 3	4092	4389
Gene 4	5279	3702
Gene 5	584	611

Read quantification by function

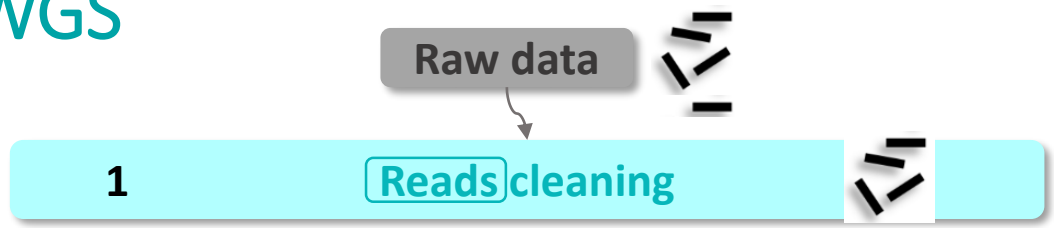
Function	Sample 1	Sample 2
Function 1	1840	5857
Function 2	4010	4506
Function 3	6005	7052
Function 4	9500	4506
Function 5	1234	5678

metagWGS

Raw data

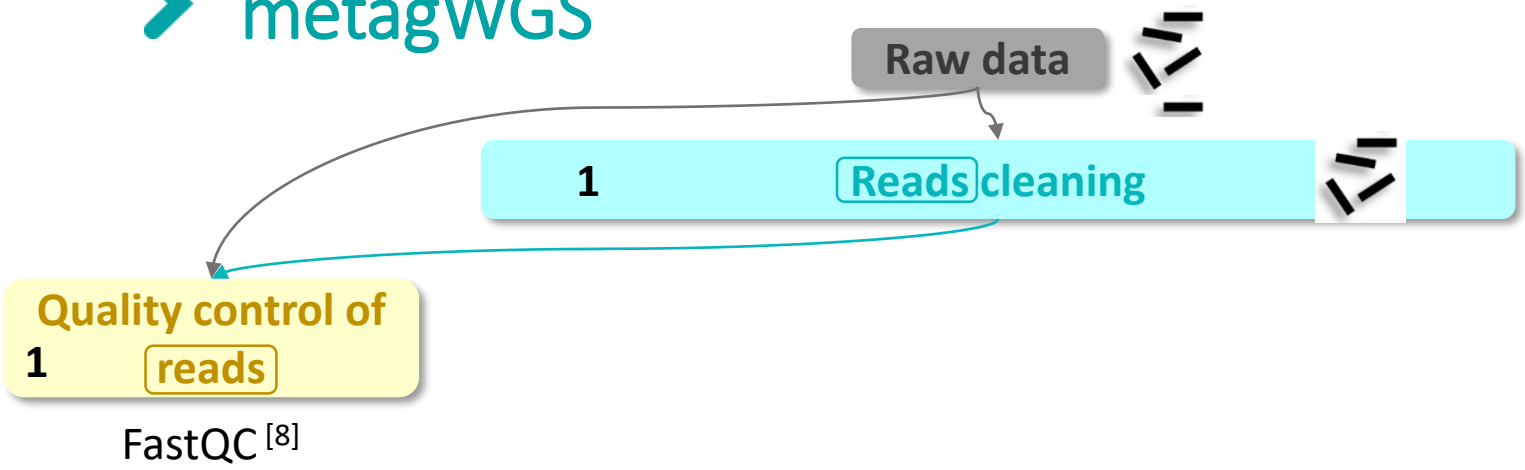
INRAE

metagWGS



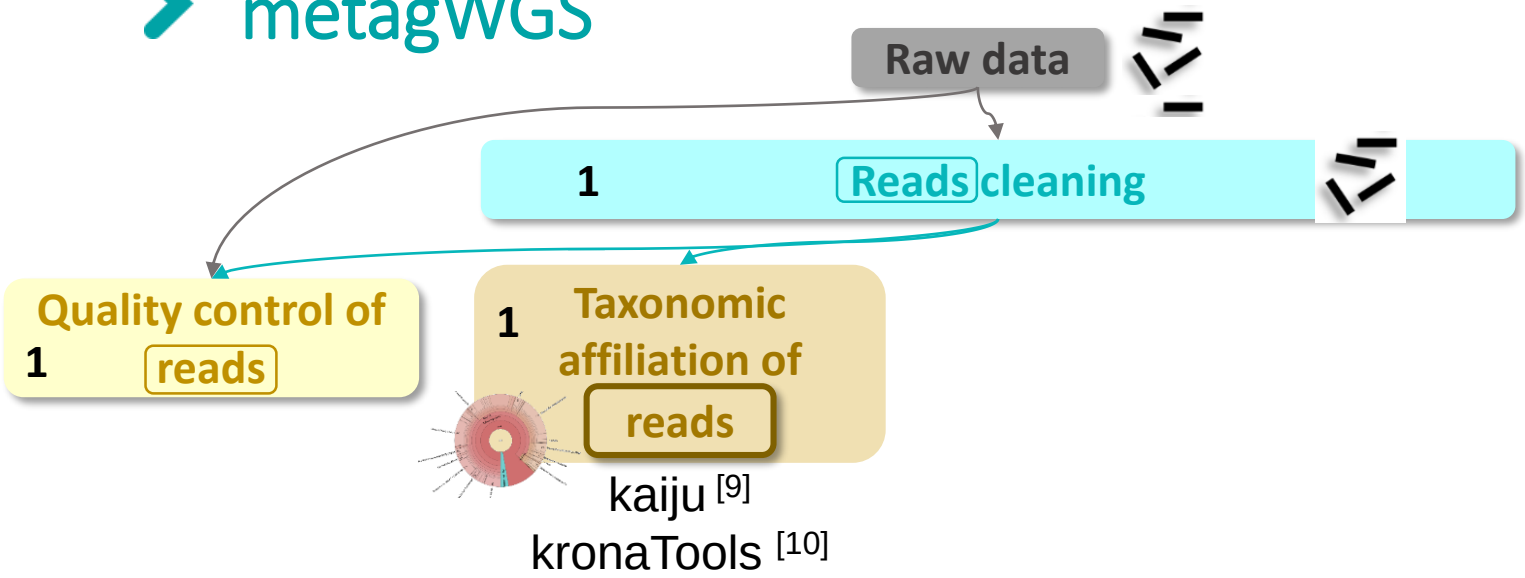
cutadapt [5], sickle [6], bwa [7] (remove host reads)

metagWGS

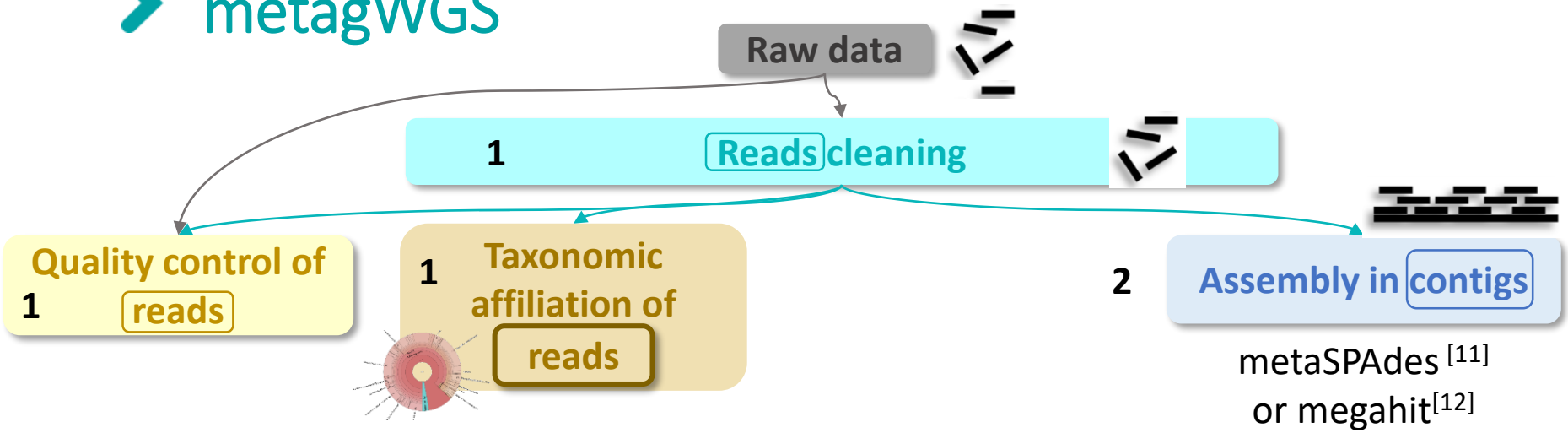


FastQC [8]

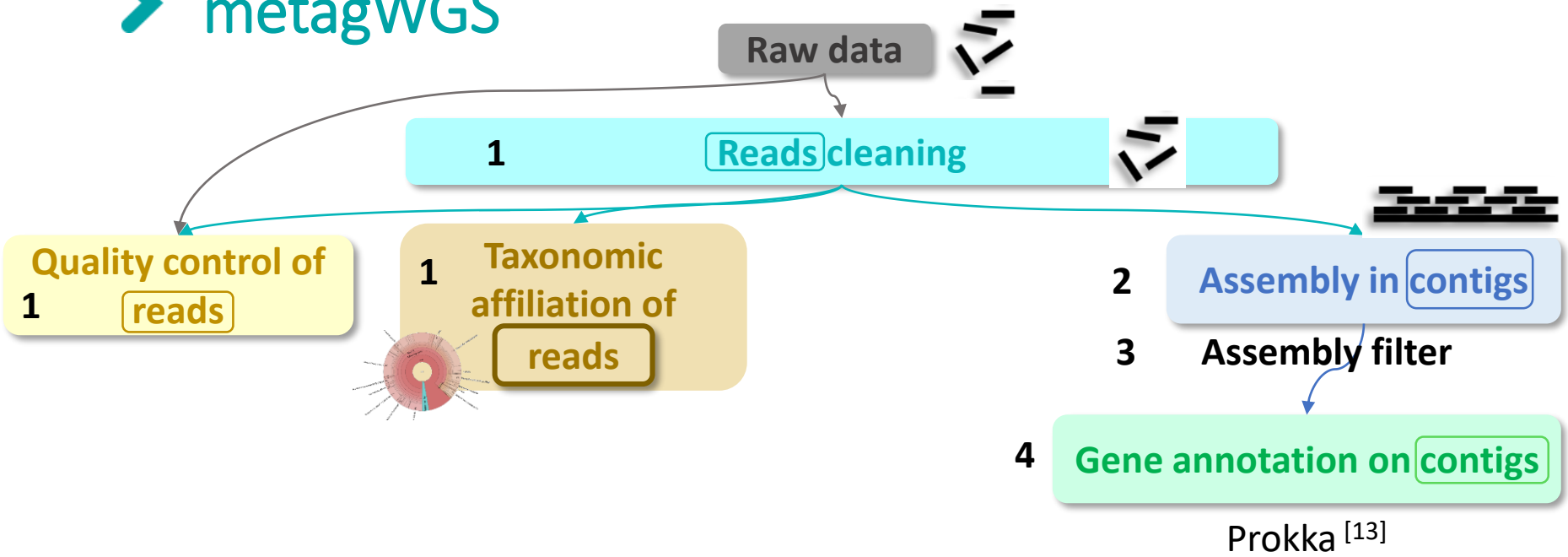
metagWGS


Raw data
1 Reads cleaning
Quality control of
1 reads
1 Taxonomic affiliation of
reads
kaiju [9]
kronaTools [10]

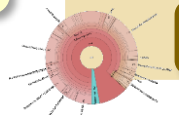
metagWGS


Raw data
1 Reads cleaning
Quality control of
1 reads
1 Taxonomic affiliation of reads
2 Assembly in contigs
metaSPAdes [11]
or megahit[12]

metagWGS


Raw data
1 Reads cleaning
Quality control of
1 reads
1 Taxonomic affiliation of reads
2 Assembly in contigs
3 Assembly filter
4 Gene annotation on contigs
Prokka [13]

metagWGS



Raw data

1 Reads cleaning

Quality control of 1 reads

1 Taxonomic affiliation of reads

2 Assembly in contigs

3 Assembly filter

4 Gene annotation on contigs

DIAMOND [14] 5 Alignment vs protein database

Python script 7 Taxonomic affiliation

genes contigs

H. pylori
Helicobacter

Gene A Gene B Gene C Gene D

→ Contig: H. pylori

Sequence id	consensus taxid	consensus lineage
Sequence 1	210	cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter; Helicobacter pylori
Sequence 2	1681	cellular organisms; Bacteria; Terrabacteria group; Actinobacteria; Actinobacteria; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; Bifidobacterium bifidum

metagWGS

Raw data

1 Reads cleaning

Quality control of 1 reads

1 Taxonomic affiliation of reads

2 Assembly in contigs

3 Assembly filter

4 Gene annotation on contigs

5 Alignment vs protein database

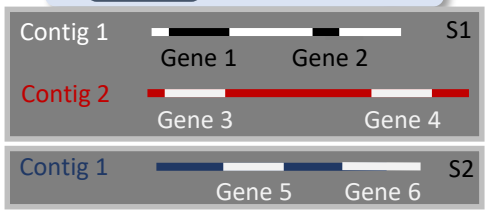
7 Taxonomic affiliation

genes | contigs

OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
Bifidobacterium bifidum	...	3467585	3

metagWGS

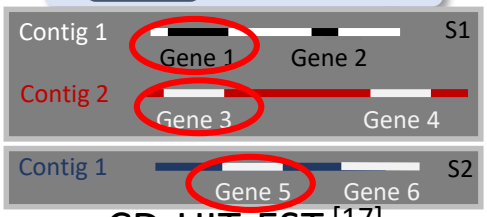
Gene	Sample 1	Sample 2
Gene 1	844	887
Gene 2	847	891
...	...	...



OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
Bifidobacterium bifidum	...	3467585	3

metagWGS

Gene	Sample 1	Sample 2
Gene 1	844	887
Gene 2	847	891
...	...	...



OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
Bifidobacterium bifidum	...	3467585	3

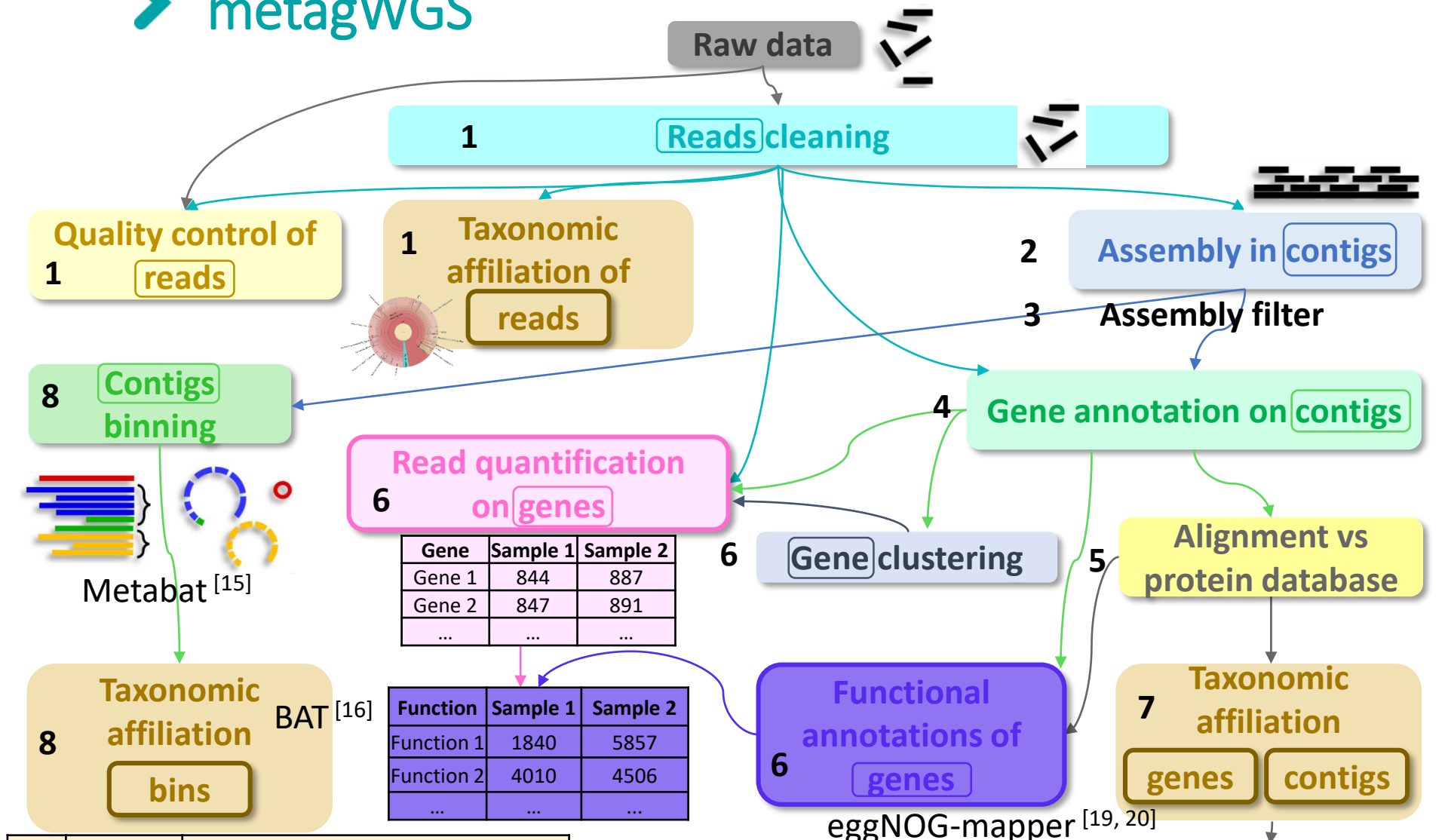
metagWGS

Gene	Sample 1	Sample 2
Gene 1	844	887
Gene 2	847	891
...	...	...

Function	Sample 1	Sample 2
Function 1	1840	5857
Function 2	4010	4506
...	...	...

OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
Bifidobacterium bifidum	...	3467585	3

metagWGS



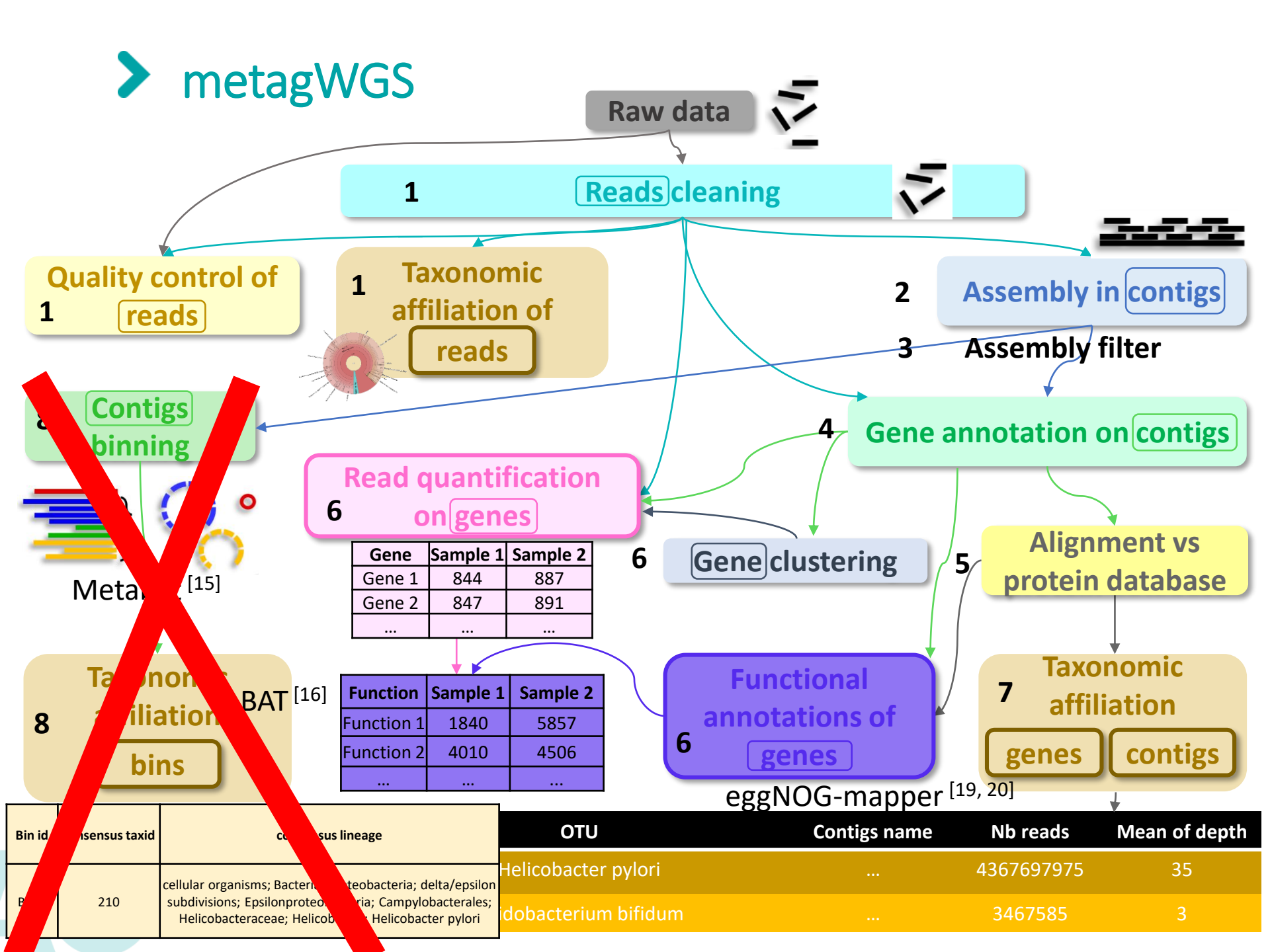
Gene	Sample 1	Sample 2
Gene 1	844	887
Gene 2	847	891
...	...	...

Function	Sample 1	Sample 2
Function 1	1840	5857
Function 2	4010	4506
...	...	...

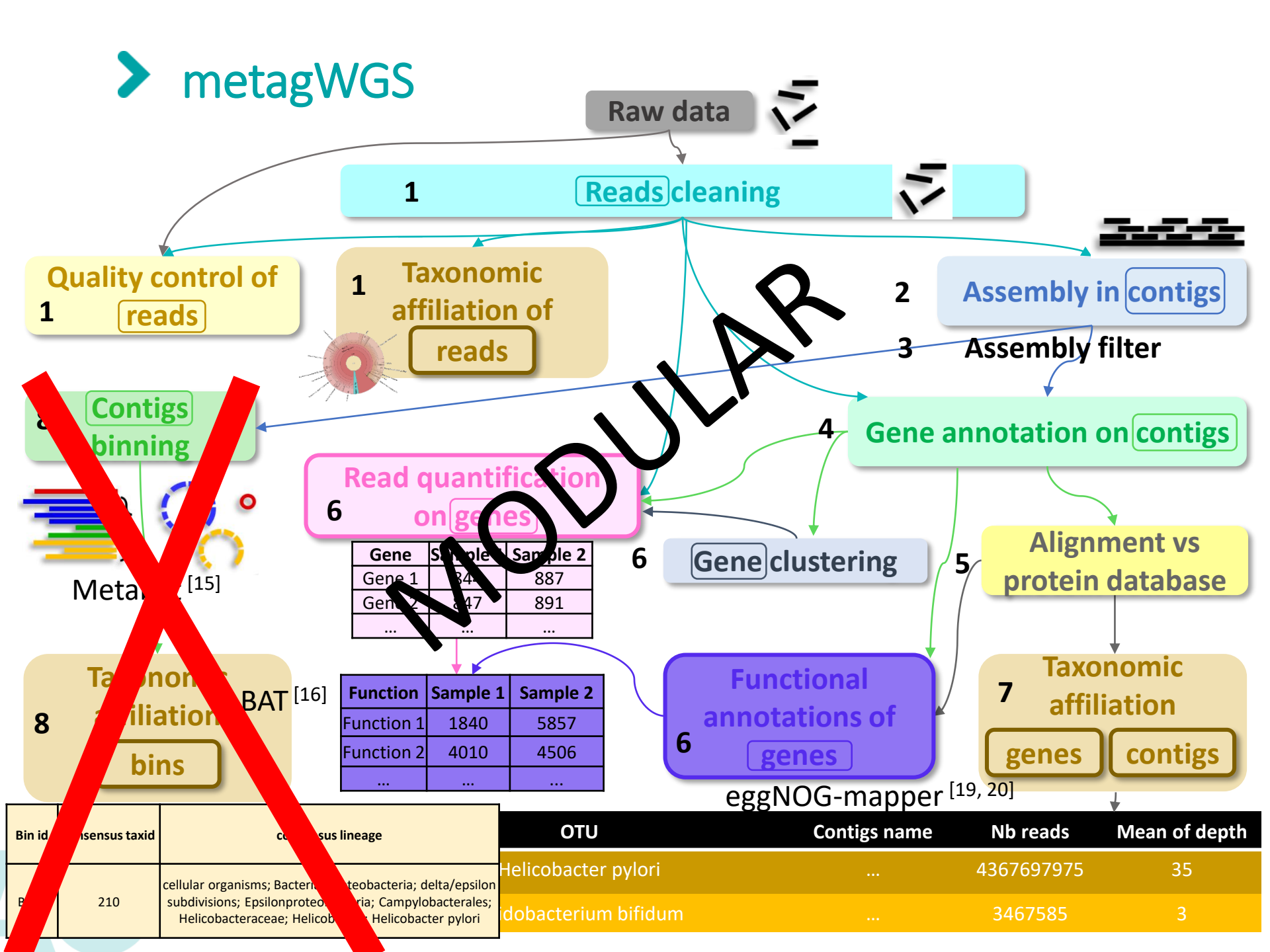
Bin id	consensus taxid	consensus lineage
Bin 1	210	cellular organisms; Bacteria; Proteobacteria; delta/epsilon subdivisions; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter; Helicobacter pylori

OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
idobacterium bifidum	...	3467585	3

metagWGS


Raw data
1 Reads cleaning
Quality control of reads
1
1 Taxonomic affiliation of reads
2 Assembly in contigs
3 Assembly filter
Contigs binning
4 Gene annotation on contigs
Read quantification on genes
6
Gene clustering
6
5 Alignment vs protein database
Gene | Sample 1 | Sample 2
Gene 1 | 844 | 887
Gene 2 | 847 | 891
... | ... | ...
Function | Sample 1 | Sample 2
Function 1 | 1840 | 5857
Function 2 | 4010 | 4506
... | ... | ...
Functional annotations of genes
6
eggNOG-mapper [19, 20]
7 Taxonomic affiliation
genes
contigs
8 Taxonomic affiliation
bins
BAT [16]
Bin id | sensus taxid | cus lineage
210 | cellular organisms; Bacteri teobacteria; delta/epsilon subdivisions; Epsilonproteo ria; Campylobacterales; Helicobacteraceae; Helicob Helicobacter pylori
OTU | Contigs name | Nb reads | Mean of depth
Helicobacter pylori | ... | 4367697975 | 35
idobacterium bifidum | ... | 3467585 | 3

metagWGS



Gene	Sample 1	Sample 2
Gene 1	344	887
Gene 2	647	891
...	...	...

Function	Sample 1	Sample 2
Function 1	1840	5857
Function 2	4010	4506
...	...	...

Bin id	...sensus taxid	c... ...sus lineage
B...	210	cellular organisms; Bacteri... ...teobacteria; delta/epsilon subdivisions; Epsilonproteo... ...ria; Campylobacterales; Helicobacteraceae; Helicob... Helicobacter pylori

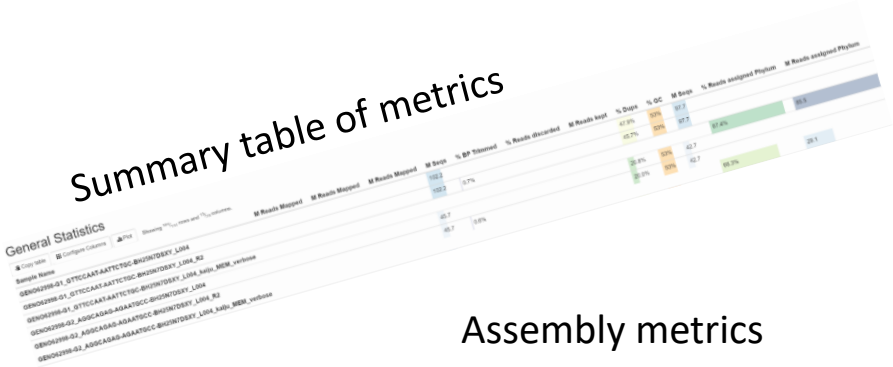
OTU	Contigs name	Nb reads	Mean of depth
Helicobacter pylori	...	4367697975	35
...idobacterium bifidum	...	3467585	3

metagWGS output files

Metrics reports with multiQC[23]

To complete the tabulated files of taxonomic and functional abundance matrices, other files are generated such as html reports of metrics on many steps of the workflow

Summary table of metrics

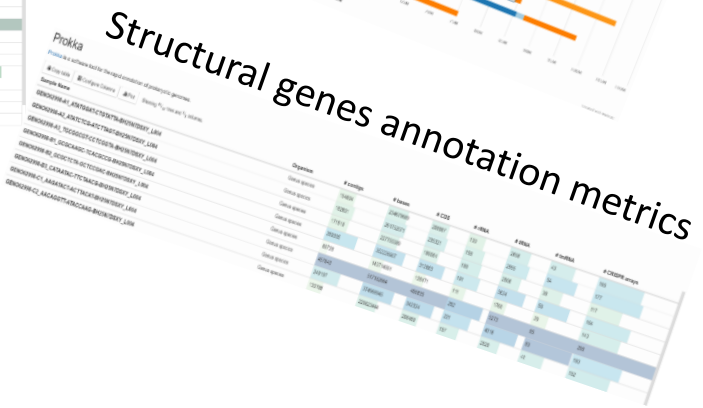


Taxonomic affiliation of reads metrics

Assembly metrics

Reads assignment to genes metrics

Structural genes annotation metrics



metagWGS is freely available

https://forgemia.inra.fr/genotoul-bioinfo/metagwgs

GPL Free Software
Free as in Freedom
GNU General Public License

- Master branch:
  - Version 2.1
  - Documentation:
    - Installation
    - Output
    - Usage
    - Use-case
  - Functional tests and associated documentation

Conclusion and future work

- Workflow easy to use, well documented, able to build taxonomic and functional profiles and many metrics.
- You can choose the step you want to run
- Next version 2.2 (main issues):
  - Binning strategies
  - Improve performances
  - …
  - Currently: upgrading to nextflow DSL2
- Add long reads (HiFi) component ➔ objective: an integrated and modular workflow able to deal with short or long reads metagenomic data

Tool reference

[1] Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.*, 2017.
[2] Kieser et al., ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data, *BMC Bioinformatics*, 2020.
[3] Köster et al., Snakemake - A scalable bioinformatics workflow engine. *Bioinformatics*, 2012.
[4] Available at *https://github.com/nf-core/mag*
[5] Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 2011.
[6] Joshi and Fass. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software]. Available at *https://github.com/najoshi/sickle*, 2011.
[7] Li and Durbin. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 2009.
[8] Andrews. FastQC. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc,* 2010.
[9] Menzel et al. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun.,* 2016.
[10] Ondov et al. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 2011.
[11] Nurk et al. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, 2017.
[12] Li et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 2015.
[13] Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 2014.
[14] Buchfink et al. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 2015.
[15] Kang et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 2019.
[16] von Meijenfeldt et al. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*, 2019.
[17] Fu et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 2012.
[18] Liao et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 2014.
[19] Huerta-Cepas et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*, 2017.
[20] Huerta-Cepas et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*, 2019

Questions ?

Merci de votre attention