MetagWGS

A nextflow workflow to analyse whole genome shotgun metagenomics data (Illumina short reads or Pacbio HiFi long reads)

GenPhySE
Génétique Physiologie et Systèmes d'Elevage

Equipe NED
Géraldine Pascal
Sylvie Combes

Claire Hoede

MIA TOULOUSE

Plateforme Bioinfo Occitanie Toulouse

Contents

- The workflow features and steps
- Brief comparison with the most popular workflows
- Focus on binning






Joanna Fourquet
Céline Noirot
Pierre Martin
Jean Mainguy
Maïna Vienne
Vincent Darbot

Projet cofinancé par le Fonds Européen de Développement Régional

Plateforme Bioinfo Occitanie Toulouse

Genotoul Bioinfo

Workflow features

metaWGS

Type of NGS data:
whole genome shotgun sequencing (Illumina HiSeq3000 or NovaSeq, paired, 2*150bp ; PacBio HiFi reads, single-end)

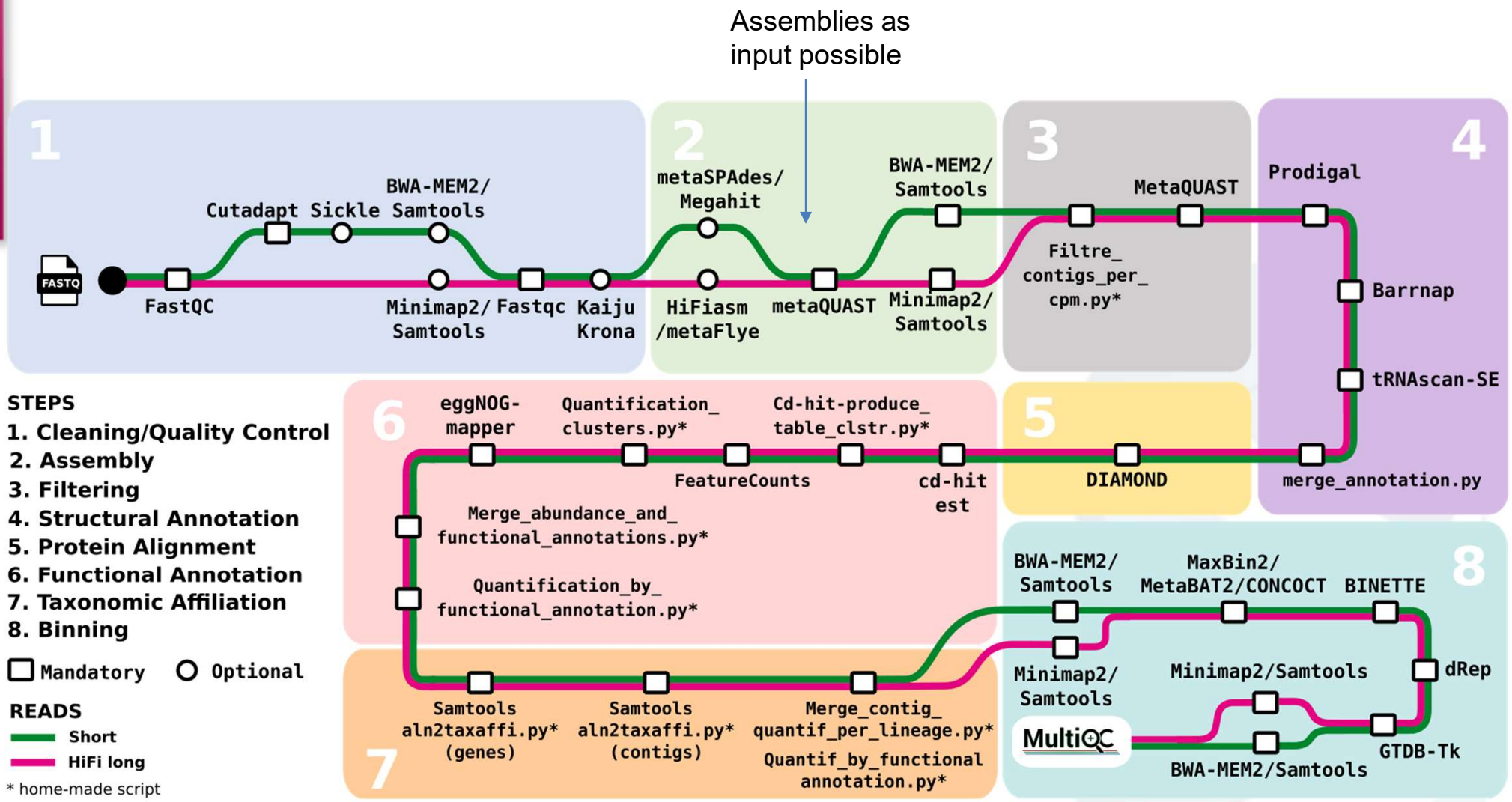
Workflow:
a scalable and reproducible metagenomic analysis with a nextflow pipeline using singularity containers

Fully documented

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

Plateforme Bioinfo Occitanie Toulouse

metagWGS



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

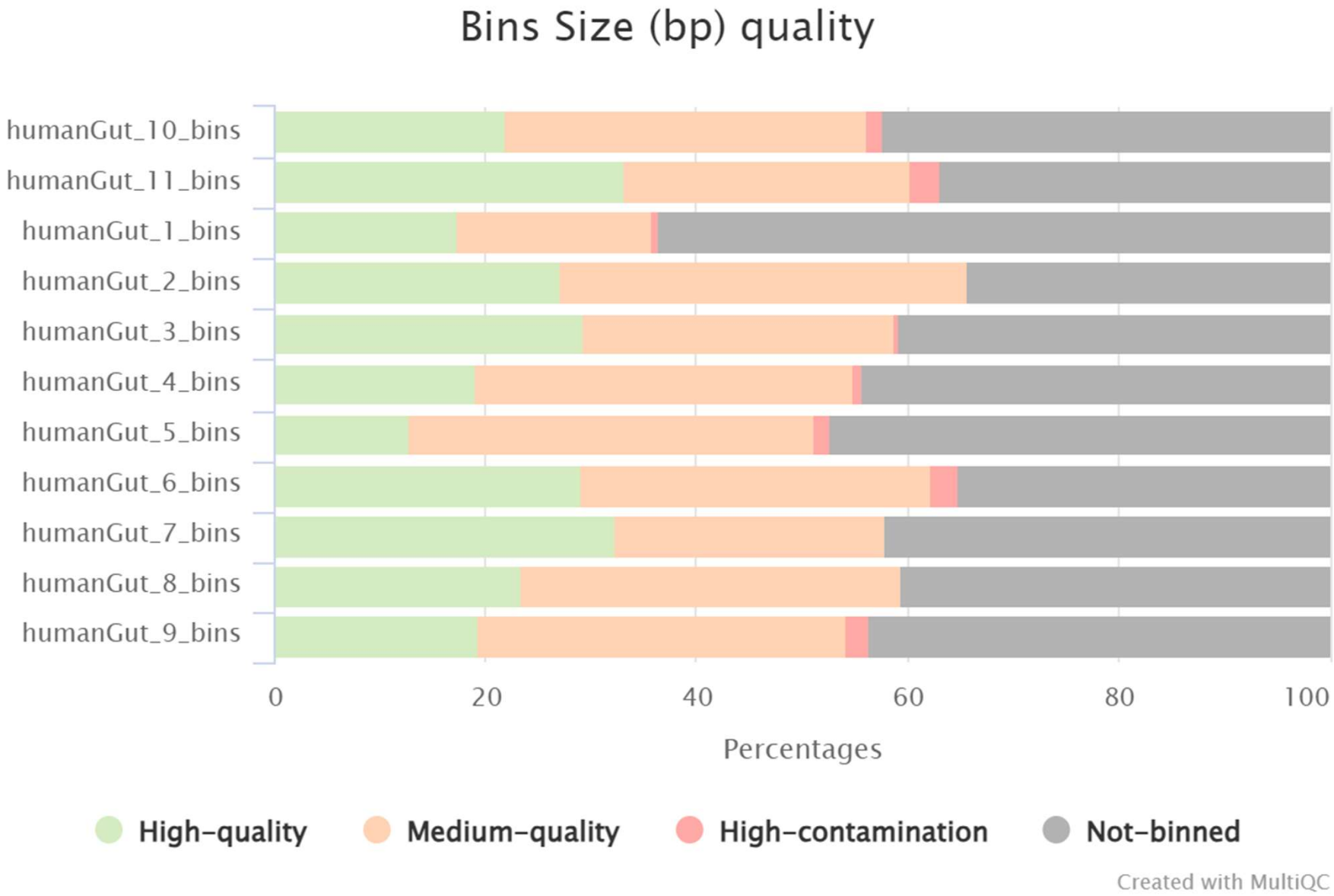
Plateforme Bioinfo Occitanie Toulouse

Type of results

Level	Reads		Genes in all contigs	All contigs		MAGS / bins	
Type of annotation / tools	**taxo**	**function**	**taxo**	**taxo**	**genes function**	**taxo**	**genes function**
MAG (nf-core)	YES	NO	NO	NO	NO	YES	YES (but not rRNA and tRNA)
Metawrap	YES	NO	NO	YES	NO	YES	YES
VEBA	NO	NO	NO	NO	NO	YES	YES
Atlas	NO	NO	NO	NO	YES	YES	YES
metagWGS (HiFi reads possible)	YES	NO	YES	YES	YES	YES	YES (via contigs)
HiFi-MAGs-pipeline (HiFi reads)	NO	NO	NO	NO	NO	YES	NO

Plateforme Bioinfo Occitanie Toulouse

Around 50% of contigs are not found in any output bins



Binning from assemblies HiFi

- Comparaison between metagWGS 2.4.2 and PacBio HiFi-MAGS-pipeline 2.0.2 binnings on the same assemblies.

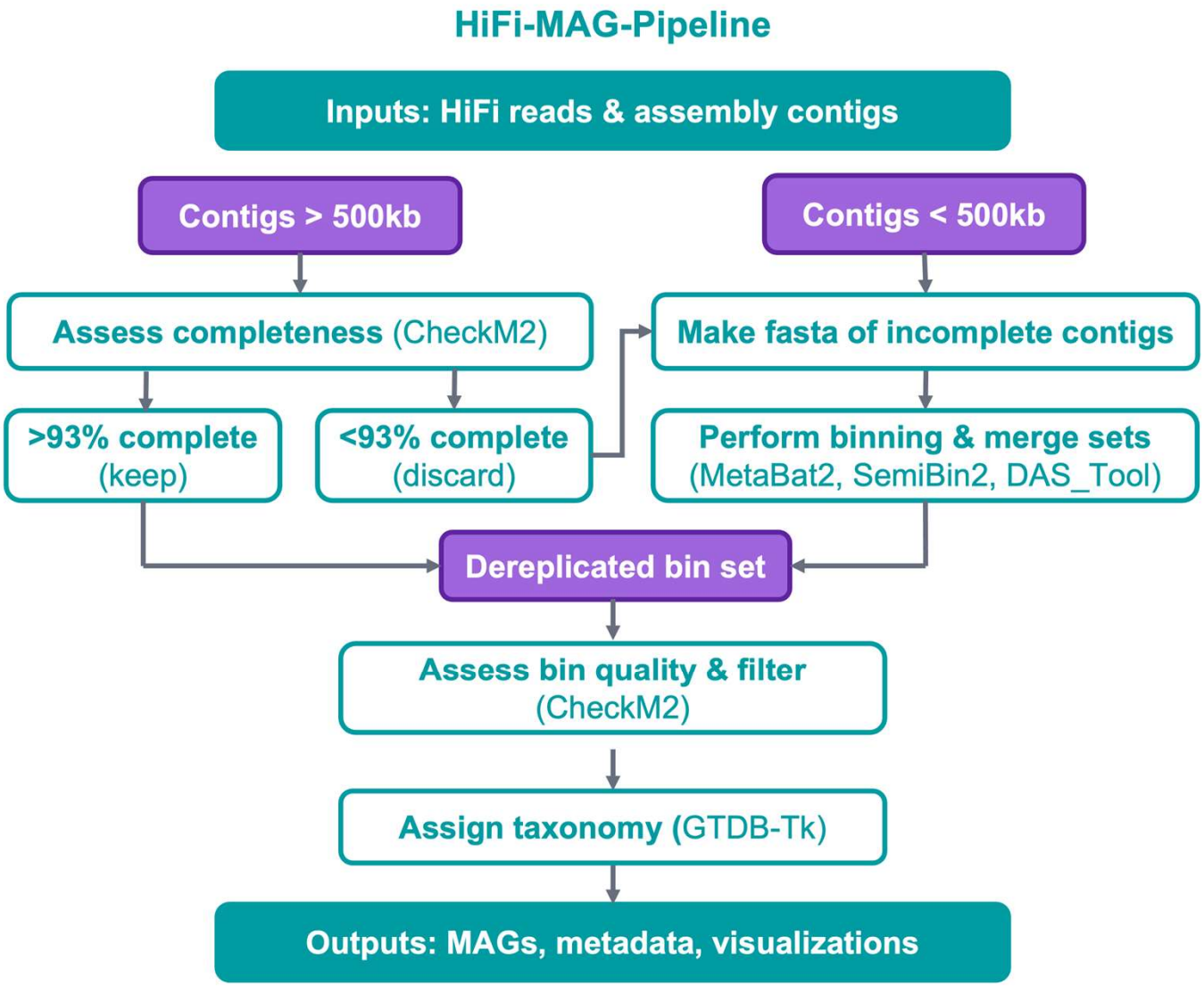
- 11 metagenomic samples from human gut public data sequenced with PacBio Sequel II (HiFi reads) from Accession: PRJNA754443 (Gehrig et al., 2022)

- very good assemblies with
  - N50 between 92.2 Kpb and 529.8 Kpb,
  - % of mapped reads on the assemblies between 95.2% and 97.8%

- Medium quality bins : completeness > 50% and contamination < 10%

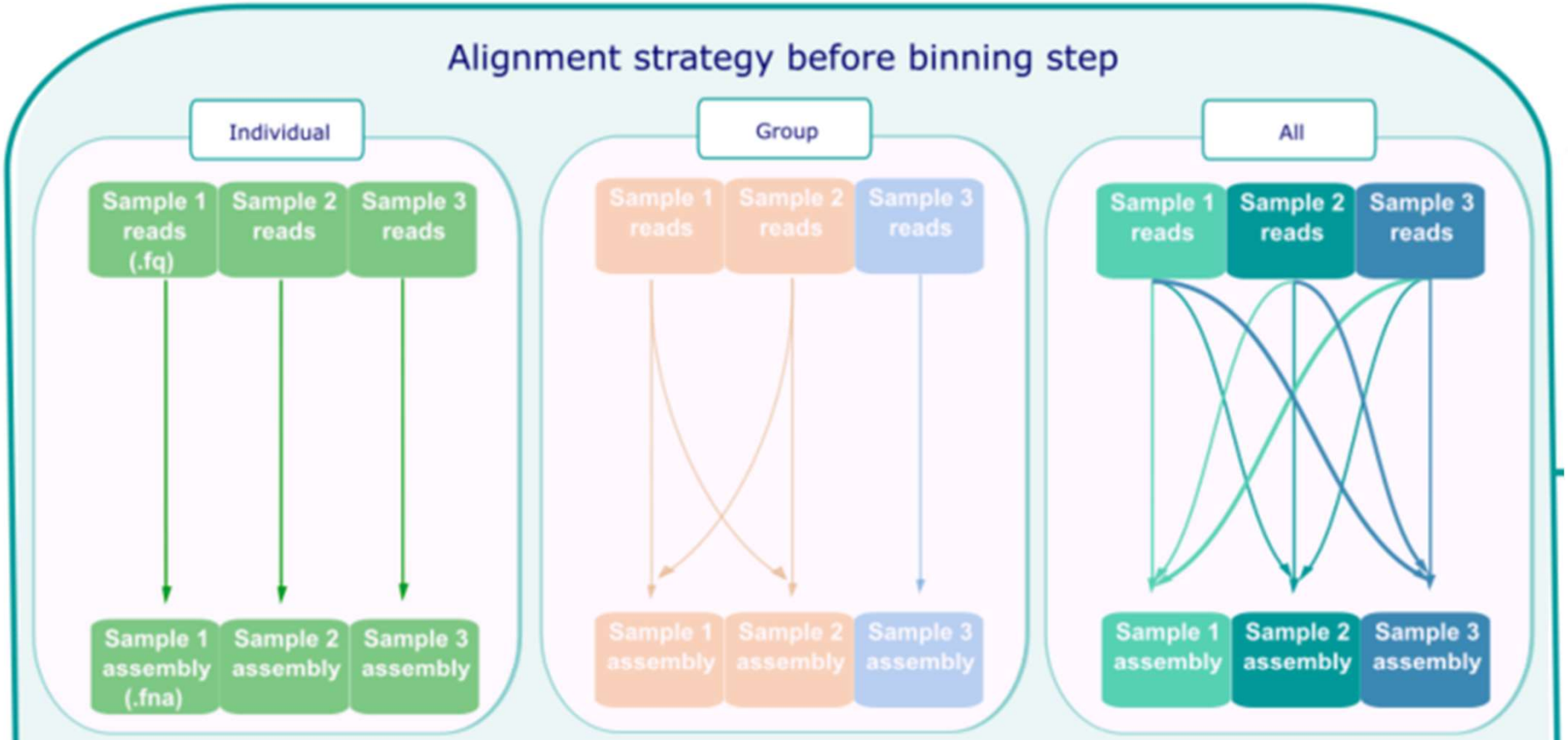
- High-quality bins : completeness > 90% and contamination < 5%

Pacbio HiFi-MAGs-pipeline



https://github.com/PacificBiosciences/pb-metagenomics-tools/blob/master/docs/Tutorial-HiFi-MAG-Pipeline.md

The binning : alignment



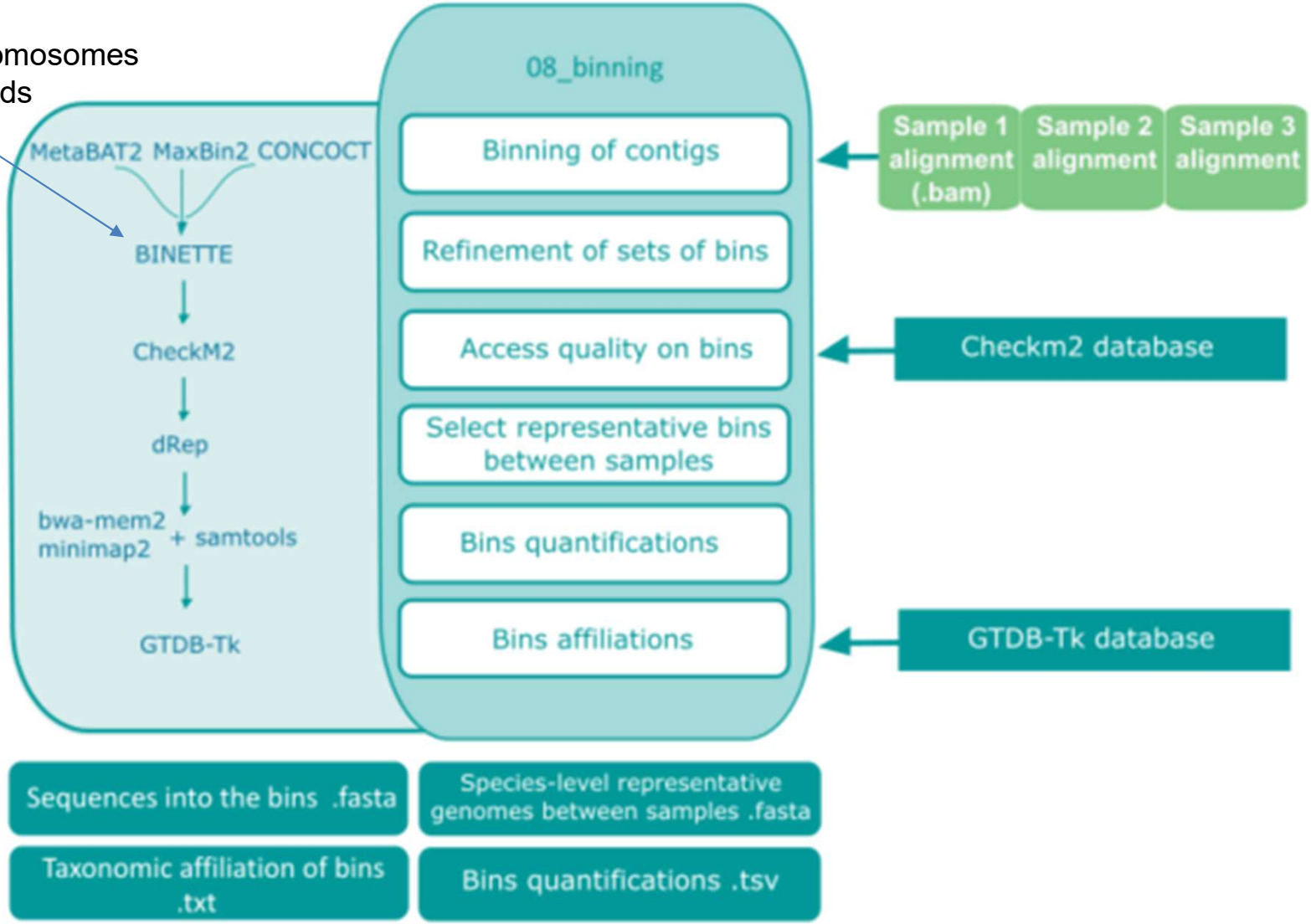
Individual : The reads of each metagenomic sample are aligned to their own assembly.

Group : The reads of metagenomics samples that belong to the same group (defined in the Sample Sheet) are aligned against each sample assembly within the group.

All : The reads of every metagenomics samples are aligned against every sample assembly.

The binning

+ circular chromosomes
when HiFi reads



Binette

Jean Mainguy

- https://github.com/genotoul-bioinfo/Binette

- Inspired by metawrap but better:
  - faster (~ 7 x)
  - not limited to 3 sets of bins (to add the circular chromosomes)
  - selects the best bins in a more elegant way (considers more possible solutions)

- From the sets of bins it is given as input, Binette builds new hybrid bins. A bin = set of contigs. When two bins overlap (share at least one contig), Binette creates new bins:
  - The intersection bin: contigs shared by the bins.
  - The difference bin: contigs found only in one bins and not in the other.
  - The union bin: all contigs contained in the overlapping bins.

- We then use checkm2 to estimate the quality of the bins and choose the best possible one.

Uritskiy, *et al*. *Microbiome* **6**, 158 (2018)

Projet cofinancé par le Fonds Européen de Développement Régional

Plateforme Bioinfo Occitanie Toulouse

Results

	Medium quality		High quality	
sample	**metagWGS**	**HiFi-MAGS-pipeline**	**metagWGS**	**HiFi-MAGS-pipeline**
humanGut_1 (SRR15489020)	46	50	20	20
humanGut_2 (SRR15489019)	53	53	19	17
humanGut_3 (SRR15489018)	105	93	45	43
humanGut_4 (SRR15489017)	90	89	25	19
humanGut_5 (SRR15489016)	30	28	6	5
humanGut_6 (SRR15489015)	65	50	21	17
humanGut_7 (SRR15489014)	38	34	18	15
humanGut_8 (SRR15489013)	43	38	15	14
humanGut_9 (SRR15489011)	79	68	24	23
humanGut_10 (SRR15489010)	84	80	29	25
humanGut_11 (SRR15489009)	75	76	33	33
Total	**708**	**659**	**255**	**231**

Table 1 : number of bins produced for each sample with more than 50% completeness and less than 10% contamination by the two workflows (medium quality) and number of bins produced for each sample with more than 90% completeness and less than 5% contamination by the two workflows (high quality).

From all these 708 bins, metagWGS obtains 246 MAGs dereplicated to 95% ANI (see suppl. Table 3)

Pacbio workflow is faster

Thanks for your attention

Do you have any questions ?

Plateforme Bioinfo Occitanie Toulouse

Contents

- The benefits of HiFi reads on the assembly






Joanna Fourquet Céline Noirot Pierre Martin Jean Mainguy Maïna Vienne Vincent Darbot

Projet cofinancé par le Fonds Européen de Développement Régional

Two datasets sequenced in Pacbio HiFi

- **The mock in 2 versions:**
  - **Mock ADN: mixture of genomic dna from 10 species**
  - **Mock bact: mix of cells of 10 species whose DNA has been extracted by A. Castinel**

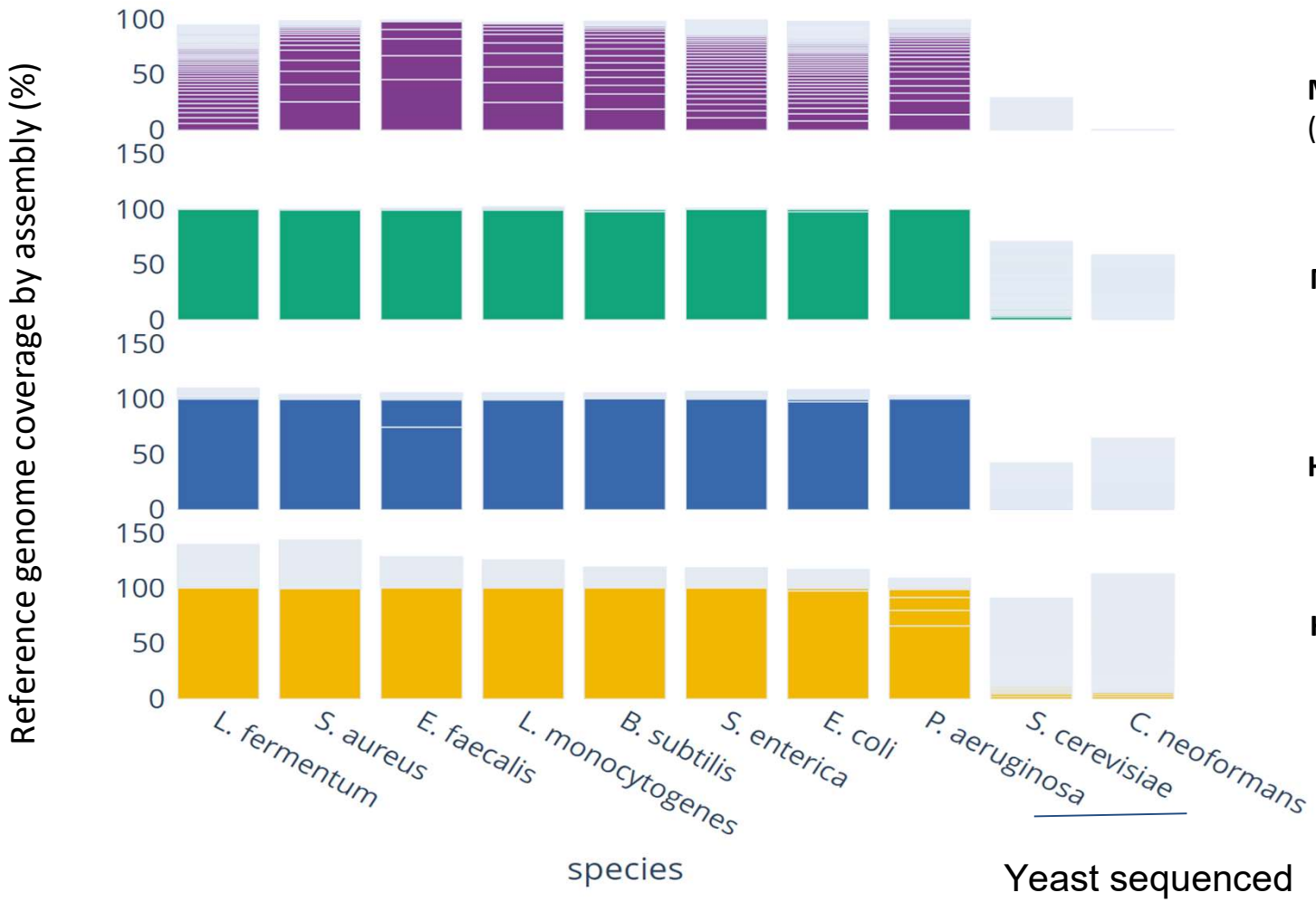
Sample Name	Mean read length (pbs)	Nb seq (Millions)
Mock ADN	5095	1.1
Mock bact	4880	0.9

- **The sample 8 of expomicopig:**
  - **Several depth**

Projet cofinancé par le Fonds Européen de Développement Régiona

Mocks assembly are very good with HiFi reads

Mock bact



MetaSPAdes Illumina
(Maxime Manno, get-plage, get-l

Metaflye

HiCanu

Hifiasm-meta

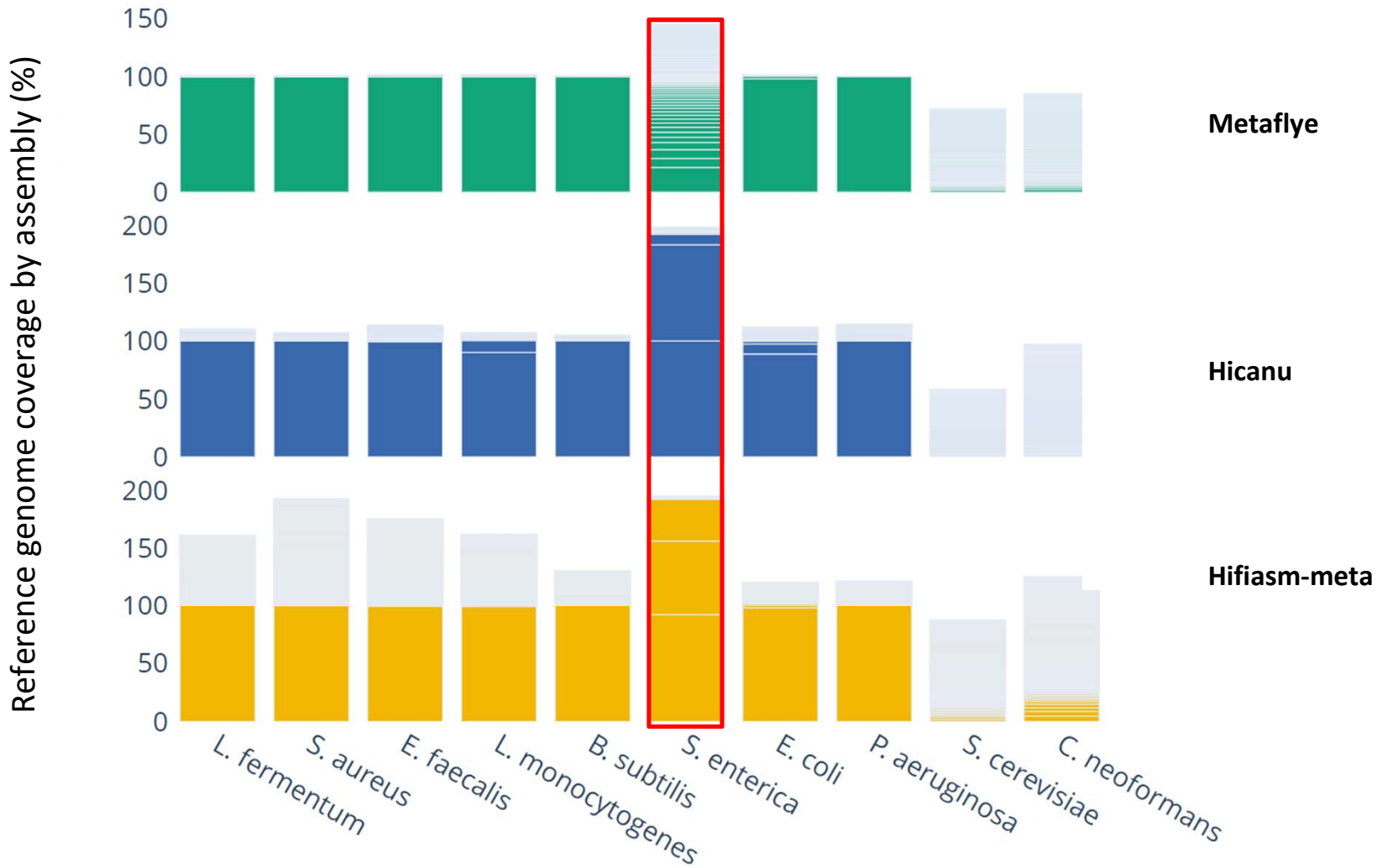
Illumina : 68M reads pairs (150pbs each)

Yeast sequenced between 2X and 5X

Projet cofinancé par le Fonds Européen de Développement Régiona

Mocks assembly are very good with HiFi reads

Mock ADN



Projet cofinancé par le Fonds Européen de Développement Régiona

Genotoul Bioinfo

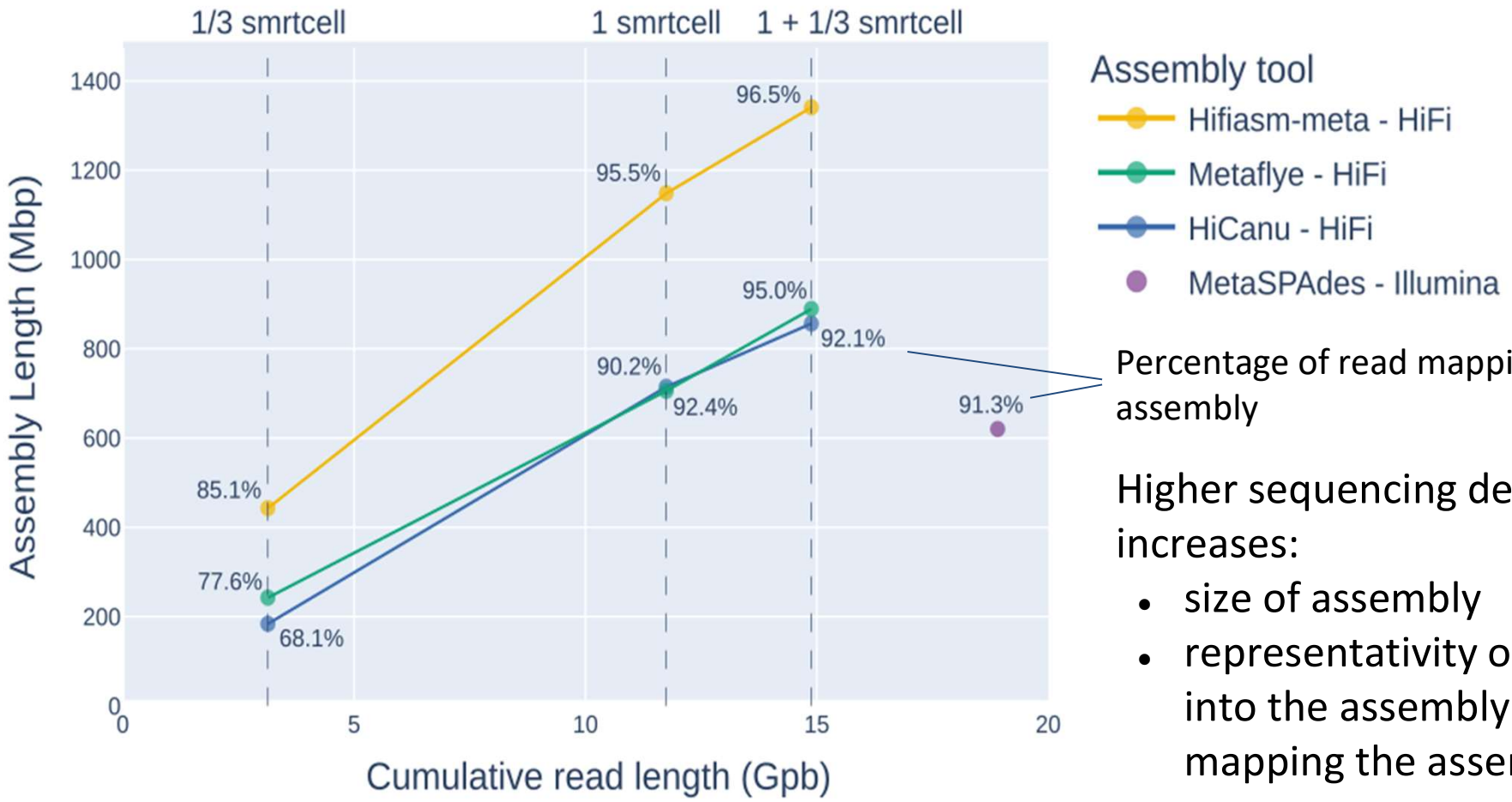
The mock community is very simple and therefore easy to assemble.
How does the analysis behave with more complex sample ?

Sample 8, pig feces dataset

Genotoul GeT
SeqOccIn
L'EUROPE S'ENGAGE L'OCCITANIE AGIT
UNION EUROPÉENNE
La Région Occitanie Pyrénées · Méditerranée
Projet cofinancé par le Fonds Européen de Développement Régional

Assembly size increases with higher sequencing depth

Feces sample 8



Percentage of read mapping the assembly

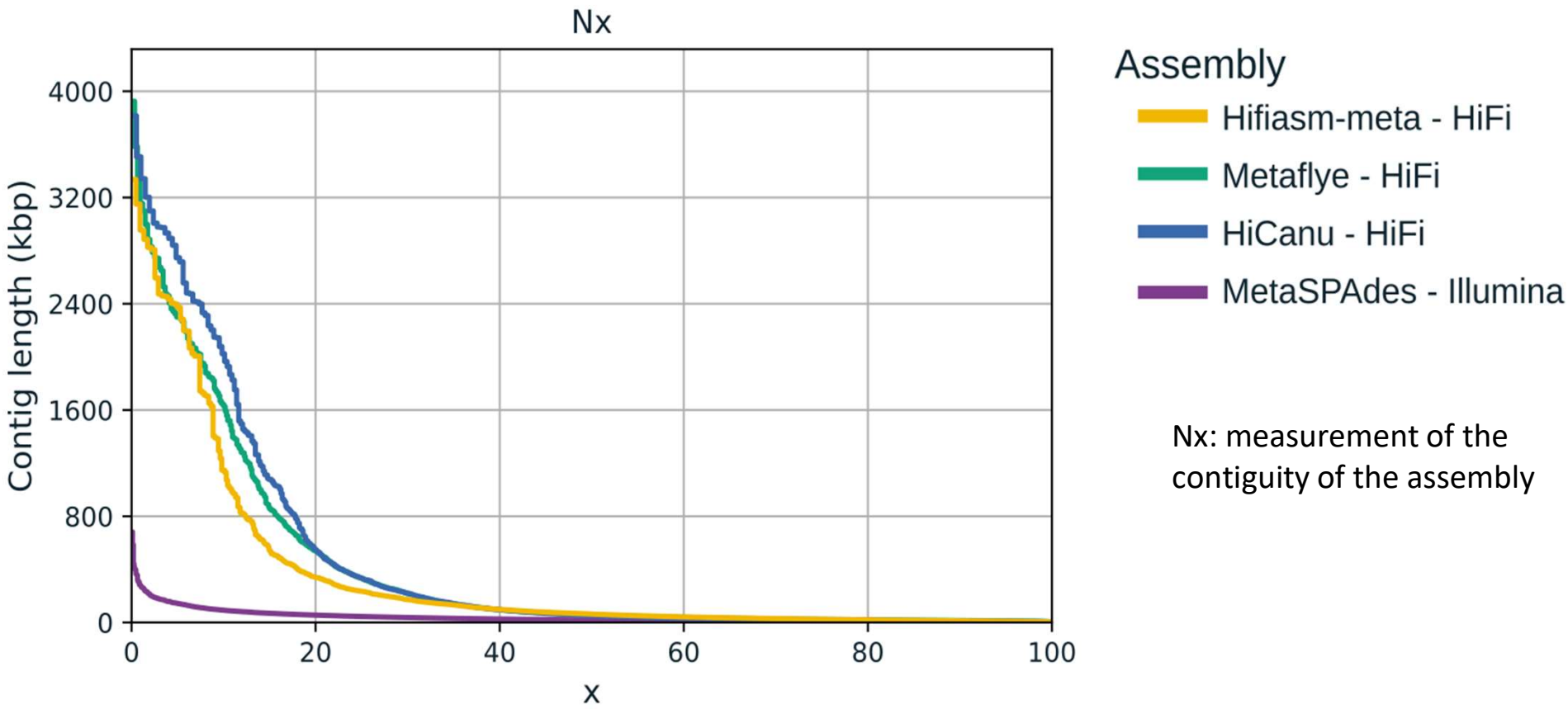
Higher sequencing depth increases:

- size of assembly
- representativity of the data into the assembly (% read mapping the assembly)

Projet cofinancé par le Fonds Européen de Développement Régional

Excellent contiguity of HiFi assemblies

Feces sample 8



Nx: measurement of the contiguity of the assembly

Projet cofinancé par le Fonds Européen de Développement Régional

To conclude on assembly step

- HiFi assemblies are much less fragmented than Illumina reads, but require sufficient sequencing depth at a higher cost than Illumina.
- more expensive

Projet cofinancé par le Fonds Européen de Développement Régiona

Plateforme Bioinfo Occitanie Toulouse

Output tree

01_clean_qc	15/05/2023 18:02	Dossier de fichiers
02_assembly	15/05/2023 18:02	Dossier de fichiers
03_filtering	15/05/2023 18:03	Dossier de fichiers
04_structural_annot	15/05/2023 18:03	Dossier de fichiers
05_protein_alignment	15/05/2023 18:05	Dossier de fichiers
06_func_annot	15/05/2023 18:03	Dossier de fichiers
07_taxo_affi	15/05/2023 18:01	Dossier de fichiers
08_binning	15/05/2023 18:05	Dossier de fichiers
MultiQC	15/05/2023 18:01	Dossier de fichiers
pipeline_info	15/05/2023 18:05	Dossier de fichiers

CH1

MultiQC output

Kaiju

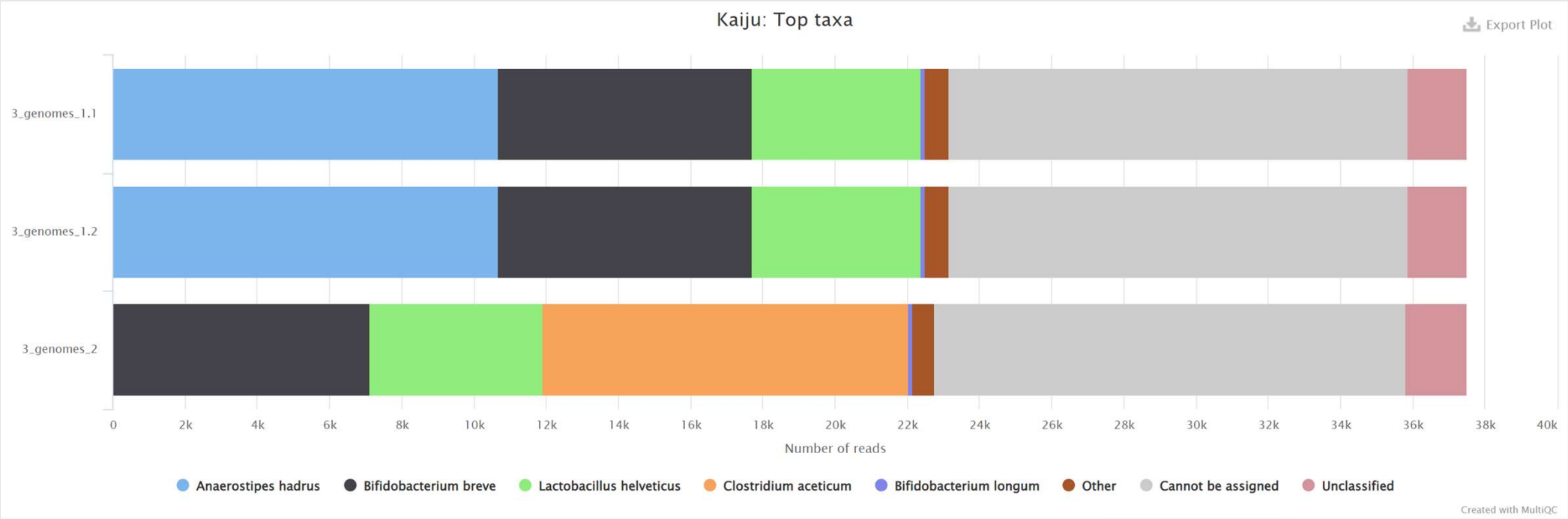
Kaiju a fast and sensitive taxonomic classification for metagenomics. DOI: 10.1038/ncomms11257.

Top taxa

Help

The number of reads falling into the top 5 taxa across different ranks.

Number of reads | Percentages | Species | Genus | Family | Order | Class | Phylum



Allons voir un exemple de multiQC….

Diapositive 23

CH1 Claire Hoede; 16/05/2023

pipeline_info outputs

- Db_versions.txt
- Software_versions.txt

```
DB (folder or file) Size    Last modification   Path
Accession2taxid prot.accession2taxid.FULL

Accession2taxid 360K    2021-12-16  /work/project/plateforme/metaG/functional_test/FT_banks/taxonomy_2021-12-7/prot.accession2taxid.FULL
Checkm2 uniref100.KO.1.dmnd

Checkm2 2.9G    2021-03-23  /work/project/plateforme/metaG/functional_test/FT_banks/checkm2DB/CheckM2_database/uniref100.KO.1.dmnd
Diamond refseq_bacteria_100000.dmnd

Diamond 48M 2021-10-25  /work/project/plateforme/metaG/functional_test/FT_banks/refseq_bacteria_2021-05-20/refseq_bacteria_100000.dmnd
Eggnog_Mapper data

Eggnog_Mapper   48G 2022-09-08  /work/project/plateforme/metaG/functional_test/FT_banks/eggnog-mapper-2.1.9/data
GTDBTK release207_v2

GTDBTK  66G 2022-05-09  /work/project/plateforme/metaG/databases/GTDBtk_data/release207_v2
Host_genome Homo_sapiens.GRCh38_chr21.fa

Host_genome 46M (1 seq) 2023-01-27  /work/project/plateforme/metaG/functional_test/metagwgs-test-datasets/small/input/host/Homo_sapiens.GRCh38_chr21.fa
Kaiju nodes.dmp

Kaiju   55G 2021-10-19  /work/project/plateforme/metaG/functional_test/FT_banks/kaijudb_refseq_2020-05-25
Taxdump new_taxdump

Taxdump 665M    2021-12-16  /work/project/plateforme/metaG/functional_test/FT_banks/taxonomy_2021-12-7/new_taxdump
```

Main outputs

- 05_protein_alignment :
  - M8 results of diamond on each gene

- 06_func_annot :

06_1_clustering	15/05/2023 18:04	Dossier de fichiers
06_2_quantification	15/05/2023 18:03	Dossier de fichiers
06_3_functional_annotation	15/05/2023 18:03	Dossier de fichiers

Functional annotation outputs

- 06_1_func_annot :
  - table_clst.txt

- 06_2_quantification :
  - Featurecount's results by gene for each sample

3_genomes_1.1_c796.CDS_17	3_genomes_1.1_c796.CDS_17
3_genomes_1.1_c796.CDS_17	3_genomes_1.2_c796.CDS_17
3_genomes_1.1_c796.CDS_17	3_genomes_2_c787.CDS_85
3_genomes_1.1_c866.CDS_175	3_genomes_1.1_c866.CDS_175
3_genomes_1.1_c866.CDS_175	3_genomes_1.2_c866.CDS_175
3_genomes_1.1_c866.CDS_175	3_genomes_2_c277.CDS_48
3_genomes_1.1_c730.CDS_27	3_genomes_1.1_c730.CDS_27
3_genomes_1.1_c730.CDS_27	3_genomes_1.2_c730.CDS_27
3_genomes_1.1_c479.CDS_92	3_genomes_1.1_c479.CDS_92
3_genomes_1.1_c479.CDS_92	3_genomes_1.2_c479.CDS_92
3_genomes_1.1_c826.CDS_26	3_genomes_1.1_c826.CDS_26
3_genomes_1.1_c826.CDS_26	3_genomes_1.2_c826.CDS_26
3_genomes_1.1_c826.CDS_26	3_genomes_2_c1566.CDS_55
3_genomes_1.1_c537.CDS_5	3_genomes_1.1_c537.CDS_5
3_genomes_1.1_c537.CDS_5	3_genomes_1.2_c537.CDS_5
3_genomes_1.1_c537.CDS_5	3_genomes_2_c2439.CDS_99
3_genomes_1.1_c360.CDS_43	3_genomes_1.1_c360.CDS_43
3_genomes_1.1_c360.CDS_43	3_genomes_1.2_c360.CDS_43
3_genomes_1.1_c23.CDS_61	3_genomes_1.1_c23.CDS_61
3_genomes_1.1_c23.CDS_61	3_genomes_1.2_c23.CDS_61
3_genomes_1.1_c730.CDS_53	3_genomes_1.1_c730.CDS_53
3_genomes_1.1_c730.CDS_53	3_genomes_1.2_c730.CDS_53
3_genomes_1.1_c898.CDS_20	3_genomes_1.1_c898.CDS_20
3_genomes_1.1_c898.CDS_20	3_genomes_1.2_c898.CDS_20
3_genomes_1.1_c898.CDS_20	3_genomes_2_c1599.CDS_153
3_genomes_1.1_c424.CDS_13	3_genomes_1.1_c424.CDS_13
3_genomes_1.1_c424.CDS_13	3_genomes_1.2_c424.CDS_13
3_genomes_1.1_c424.CDS_13	3_genomes_2_c2053.CDS_60
3_genomes_1.1_c813.CDS_11	3_genomes_1.1_c813.CDS_11
3_genomes_1.1_c813.CDS_11	3_genomes_1.2_c813.CDS_11
3_genomes_1.1_c813.CDS_11	3_genomes_2_c2439.CDS_63
3_genomes_2_c1036.CDS_1	3_genomes_1.1_c399.CDS_1
3_genomes_2_c1036.CDS_1	3_genomes_1.1_c613.CDS_1
3_genomes_2_c1036.CDS_1	3_genomes_1.2_c399.CDS_1
3_genomes_2_c1036.CDS_1	3_genomes_1.2_c613.CDS_1
3_genomes_2_c1036.CDS_1	3_genomes_2_c1036.CDS_1

Functional annotation outputs

- 06_3_func_annot :

GOs_abundance	15/05/2023 18:03	Fichier TSV	169 Ko
KEGG_ko_abundance	15/05/2023 18:03	Fichier TSV	45 Ko
KEGG_Module_abundance	15/05/2023 18:03	Fichier TSV	6 Ko
KEGG_Pathway_abundance	15/05/2023 18:03	Fichier TSV	10 Ko
PFAM_abundance	15/05/2023 18:03	Fichier TSV	57 Ko
Quantifications_and_functional_annotations	15/05/2023 18:03	Fichier TSV	3 914 Ko

Let's look at an example of these files ….

Taxonomic annotation outputs

- 07_2_affiliation_merged :
  - quantification_by_contig_lineage_all

- 07_3_plot :

abundance_per_rank	15/05/2023 18:01	Microsoft Edge HTM...	3 519 Ko
krona_mean_depth_abundance	15/05/2023 18:01	Microsoft Edge HTM...	229 Ko
krona_read_count_abundance	15/05/2023 18:01	Microsoft Edge HTM...	228 Ko
most_abundant_taxa	15/05/2023 18:01	Microsoft Edge HTM...	7 139 Ko

Let's look at an example of these files ….

Binning ouputs

- 08_binning :

08_1_binning_per_sample	15/05/2023 18:04	Dossier de fichiers
08_2_dereplicated_bins	15/05/2023 18:05	Dossier de fichiers
08_3_gtdbtk	15/05/2023 18:04	Dossier de fichiers
08_4_mapping_on_final_bins	15/05/2023 18:04	Dossier de fichiers
stats	15/05/2023 18:04	Dossier de fichiers
genomes_abundances	15/05/2023 18:04	Fichier TSV