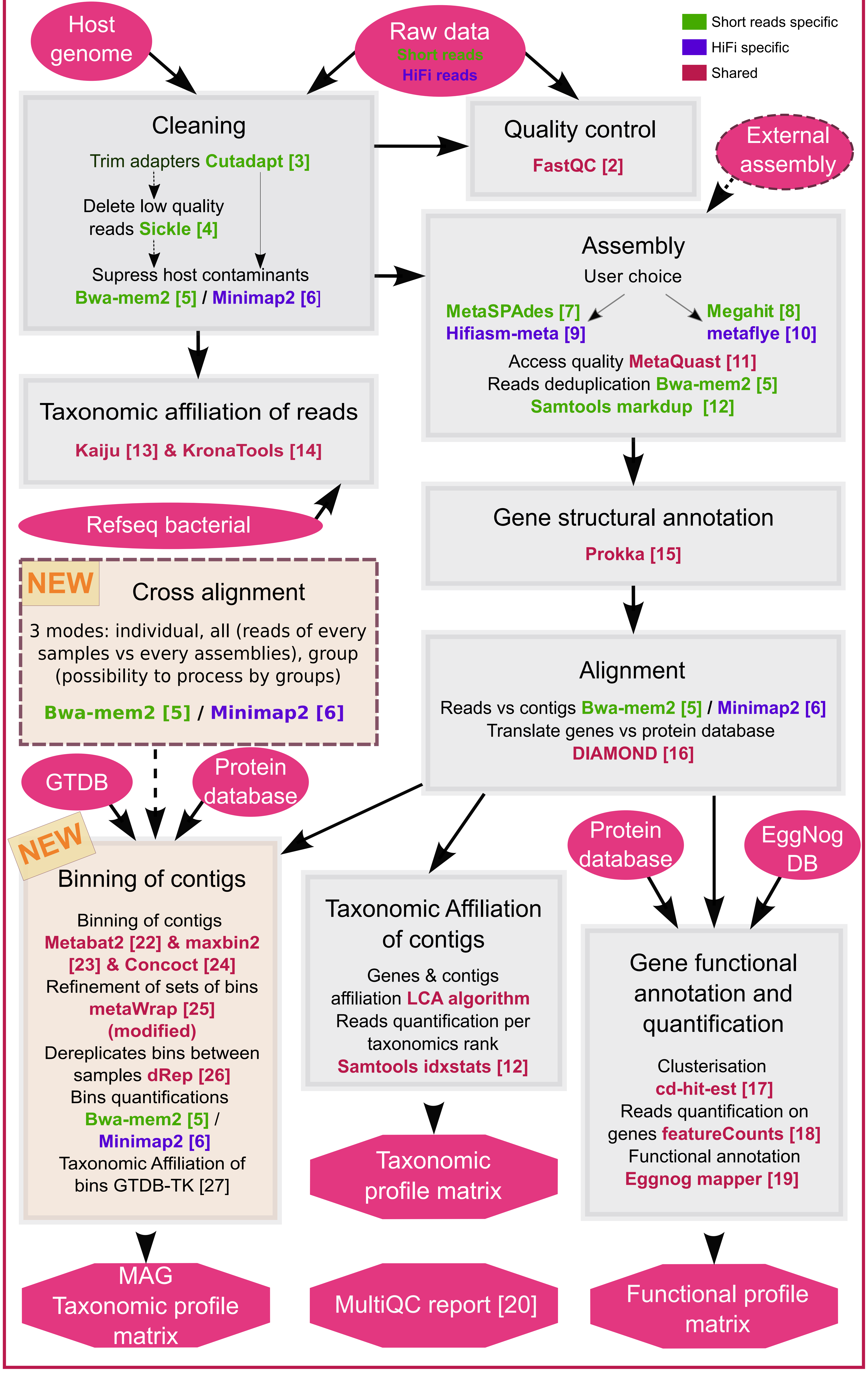
metagWGS: a workflow to analyse short and long HiFi metagenomic reads

Joanna Fourquet[1,2,3], Maïna Vienne[1,2*], Jean Mainguy[1,2*], Vincent Darbot[3*], Pierre Martin[1,2], Olivier Bouchez[4], Adrien Castinel[4], Sylvie Combes[3], Carole Iampietro[4], Christine Gaspin[1,2], Denis Milan[4], Cécile Donnadieu[4], Céline Noirot[1,2], Geraldine Pascal[3] and Claire Hoede[1,2*]

[1] Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France
[2] Université Fédérale de Toulouse, INRAE, MIAT, 31326, Castanet-Tolosan, France
[3] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France
[4] INRAE, GeT-PlaGe, Genotoul – INRAE – 31326 Castanet-Tolosan, France
* Present at ECCB 2022

Corresponding author: claire.hoede@inrae.fr

Production of whole metagenome assembly, functional and taxonomic profile



Host genome

Raw data
Short reads
HiFi reads

🟩 Short reads specific
🟪 HiFi specific
🟥 Shared

Cleaning
Trim adapters **Cutadapt [3]**
Delete low quality reads **Sickle [4]**
Supress host contaminants
Bwa-mem2 [5] / **Minimap2 [6]**

Quality control
FastQC [2]

External assembly

Assembly
User choice
MetaSPAdes [7] **Megahit [8]**
Hifiasm-meta [9] **metaflye [10]**
Access quality **MetaQuast [11]**
Reads deduplication **Bwa-mem2 [5]**
Samtools markdup [12]

Taxonomic affiliation of reads
Kaiju [13] & KronaTools [14]

Refseq bacterial

NEW
Cross alignment
3 modes: individual, all (reads of every samples vs every assemblies), group (possibility to process by groups)
Bwa-mem2 [5] / **Minimap2 [6]**

Gene structural annotation
Prokka [15]

Alignment
Reads vs contigs **Bwa-mem2 [5]** / **Minimap2 [6]**
Translate genes vs protein database
DIAMOND [16]

GTDB

Protein database

NEW
Binning of contigs
Binning of contigs
Metabat2 [22] & maxbin2 [23] & Concoct [24]
Refinement of sets of bins
metaWrap [25] (modified)
Dereplicates bins between samples **dRep [26]**
Bins quantifications
Bwa-mem2 [5] / **Minimap2 [6]**
Taxonomic Affiliation of bins GTDB-TK [27]

Taxonomic Affiliation of contigs
Genes & contigs affiliation **LCA algorithm**
Reads quantification per taxonomics rank
Samtools idxstats [12]

Protein database

EggNog DB

Gene functional annotation and quantification
Clusterisation
cd-hit-est [17]
Reads quantification on genes **featureCounts [18]**
Functional annotation
Eggnog mapper [19]

Taxonomic profile matrix

MAG Taxonomic profile matrix

MultiQC report [20]

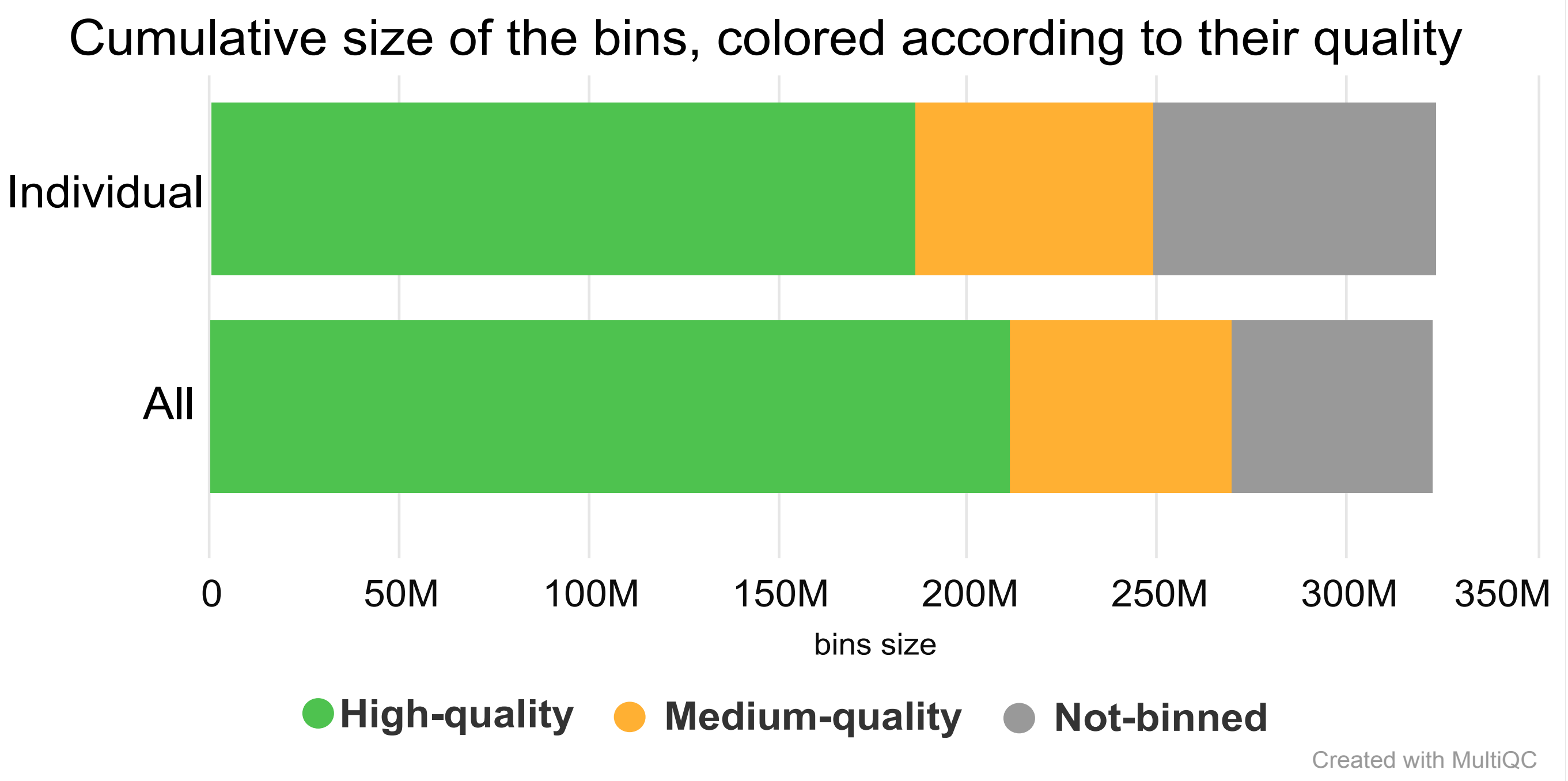
Functional profile matrix

Workflow features

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

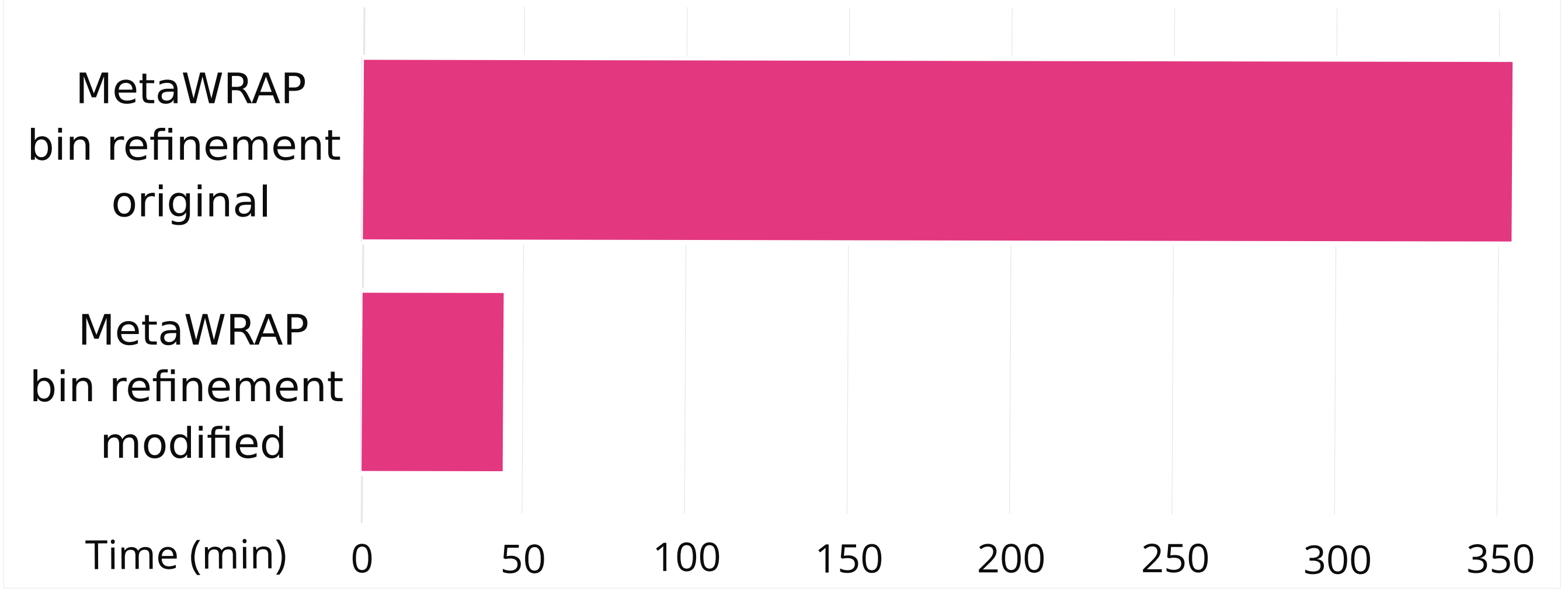
Workflow:
a scalable and reproducible metagenomic analysis with a nextflow [1] pipeline using Singularity [21] containers

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



Cumulative size of the bins, colored according to their quality

The strategy of aligning reads of every samples (Cross-alignment: All) against each assembly improved the quality of binning. Tests were done on a synthetic mock composed of 142 bacteria and archea genomes with 3 samples of 66.651.100 Illumina paired-end reads (2x150bp).



Execution time in minutes of the original MetaWRAP bin refinement module [25] compare to the improved version implemented in metagGWS, on the synthetic mock. The improved version uses Checkm2 [28] instead of Checkm1 [29] and takes avantage of a custom resume parameter. The modified version gives very similar results.

Conclusion

The new version of **metagWGS (2.3)** allows the **analysis of Illumina short reads or PacBio HiFi long-reads sequencing data** and brings as a **major new feature the binning of contigs.**

The workflow proposes to use the abundance information contained in nearby samples to **improve the binning by implementing cross-alignment per sample set.**

We have also improved the performance of the **bins refinement step by dividing the execution time by 7**.

References

Acknowledgements

Projet cofinancé par le Fonds Européen de Développement Régional

ATB_Biofilm funded by PNREST Anses, 2020/01/142
Funded by France Génomique (ANR-10-INBS-09-08)
Funded by Antiselfish Project (Labex Ecofect)
Funded by ExpoMicoPig project (France Futur elevage)
Vincent Darbot's work has been supported by RESALAB OUEST
SeqOccIn financed by FEDER funds
(Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020)

Perspectives

Better assembly of the minority species when the sequencing depth is not sufficient: implementation of co-assembly (by giving the possibility to normalize data first).

Improve the performance of the workflow: replacing Prokka [15] with other tools.

Long term perspectives: enable the annotation of antibiotic resistance genes and of the mobilome