



**HAL**  
open science

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

Joanna Fourquet, Maina Vienne, Jean Mainguy, Vincent Darbot, Pierre Martin, Olivier Bouchez, Adrien Castinel, Sylvie Combes, Carole Iampietro, Christine Gaspin, et al.

### ► To cite this version:

Joanna Fourquet, Maina Vienne, Jean Mainguy, Vincent Darbot, Pierre Martin, et al.. metagWGS: a workflow to analyse short and long HiFi metagenomic reads. ECCB2022, Sep 2022, Sitges, Spain. hal-03788263

**HAL Id: hal-03788263**

**<https://hal.inrae.fr/hal-03788263v1>**

Submitted on 26 Sep 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

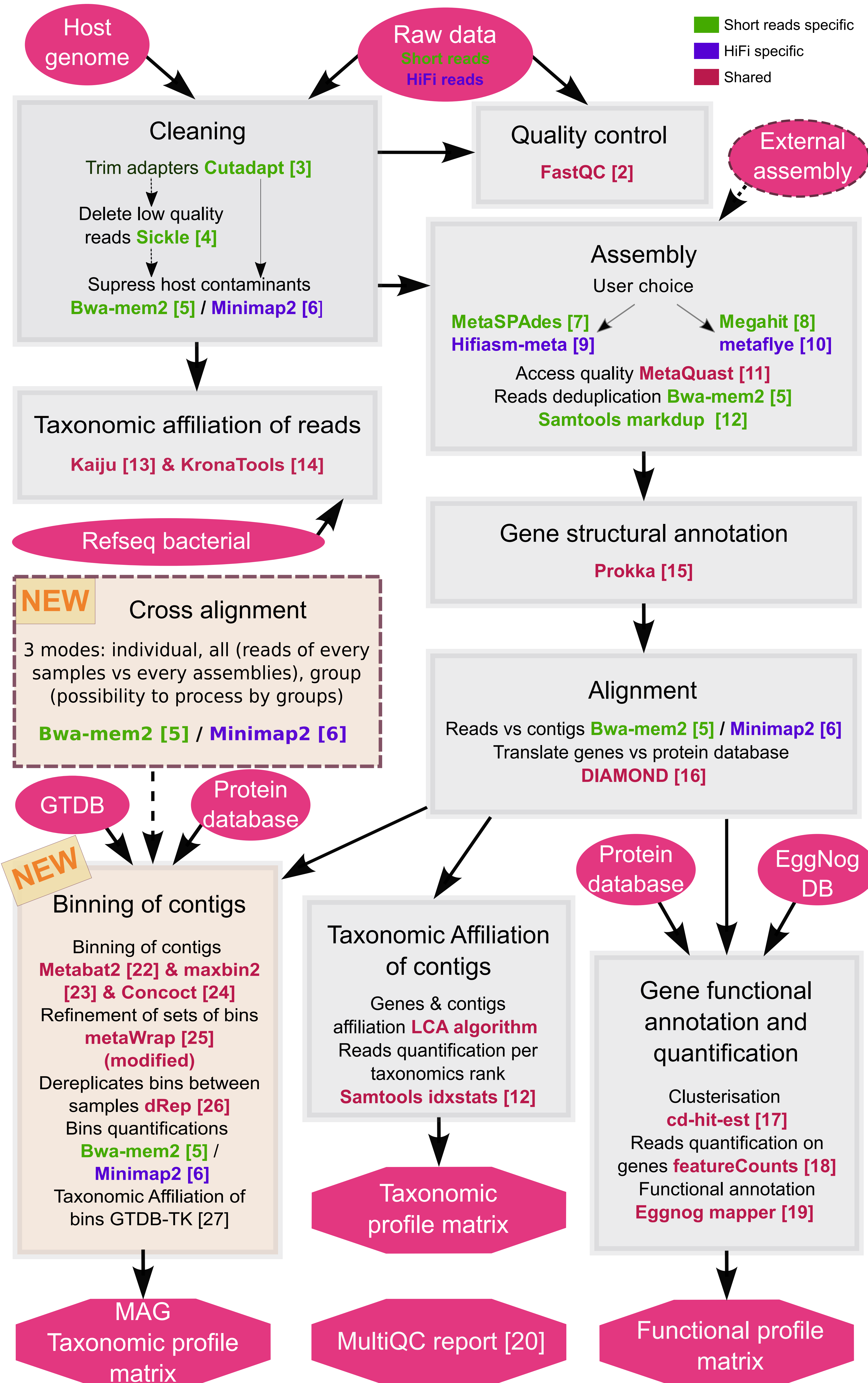
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Joanna Fourquet<sup>1,2,3</sup>, Maïna Vienne<sup>1,2\*</sup>, Jean Mainguy<sup>1,2\*</sup>, Vincent Darbot<sup>3\*</sup>, Pierre Martin<sup>1,2</sup>, Olivier Bouchez<sup>4</sup>, Adrien Castinel<sup>4</sup>, Sylvie Combes<sup>3</sup>, Carole Iampietro<sup>4</sup>, Christine Gaspin<sup>1,2</sup>, Denis Milan<sup>4</sup>, Cécile Donnadiou<sup>4</sup>, Céline Noïrot<sup>1,2</sup>, Geraldine Pascal<sup>3</sup> and Claire Hoede<sup>1,2\*</sup>

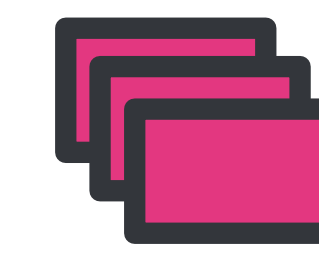
- 1 Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France  
2 Université Fédérale de Toulouse, INRAE, MIAT, 31326, Castanet-Tolosan, France  
3 GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France  
4 INRAE,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

Corresponding author: [claire.hoede@inrae.fr](mailto:claire.hoede@inrae.fr)

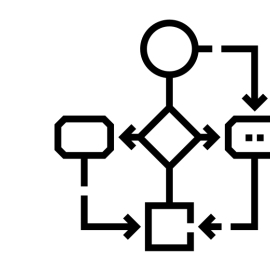
## Production of whole metagenome assembly, functional and taxonomic profile



## Workflow features



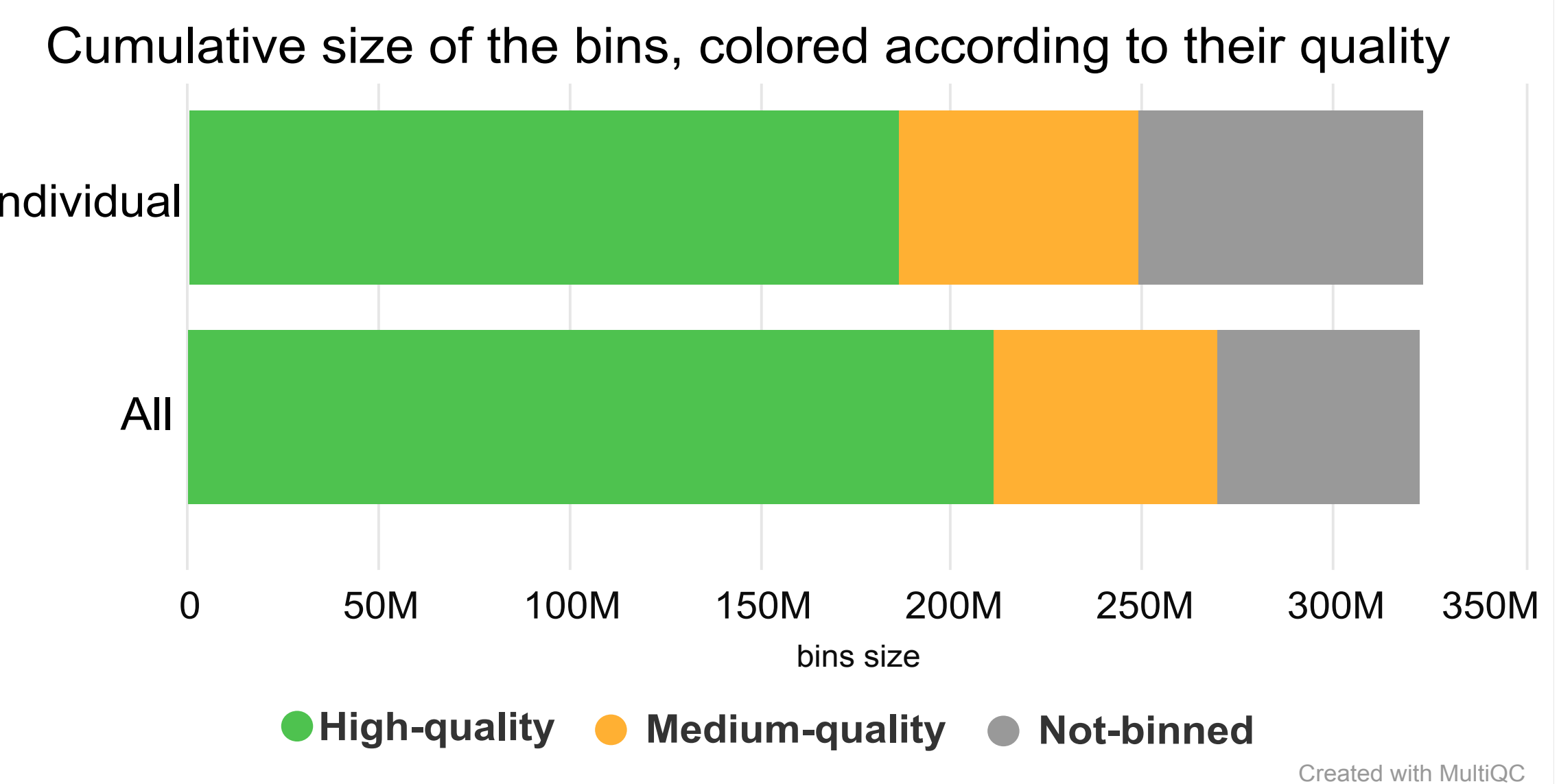
**Type of NGS data:**  
whole genome shotgun sequencing (Illumina HiSeq3000 or NovaSeq, paired, 2\*150bp ; PacBio HiFi reads, single-end)



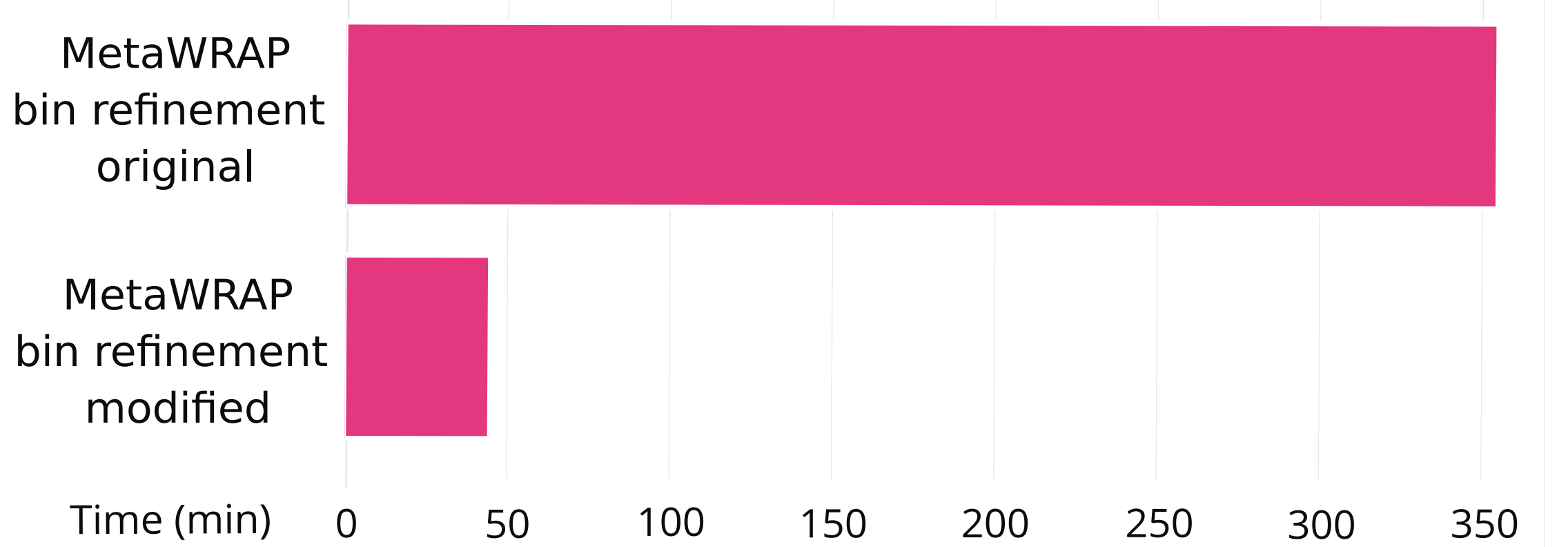
**Workflow:**  
a scalable and reproducible metagenomic analysis with a **nextflow** [1] pipeline using **Singularity** [21] containers



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



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).



The improved version uses Checkm2 [28] instead of Checkm1 [29] and takes advantage 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



Projet cofinancé par le Fonds Européen de Développement Régional

**Acknowledgements**  
ATB\_Biofilm funded by PIREST 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 élevage)  
Vincent Darbot's work has been supported by RESALAB OUEST  
SeqOcln 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 mobileome