

Joanna Fourquet^{1*}, Jean Mainguy^{1*}, Maina Vienne¹, Céline Noirot¹, Pierre Martin¹, Vincent Darbot³, Olivier Bouchez², Adrien Castinel², Sylvie Combes³, Carole Iampietro², Christine Gaspin¹, Denis Milan², Cécile Donnadiou², Geraldine Pascal³ and Claire Hoede¹

¹ INRAE, Université de Toulouse, UR875 MIAT, Bioinformatics, PF GenoToul Bioinfo, F-31326, Castanet-Tolosan, France (doi: 10.15454/1.5572369328961167E12)

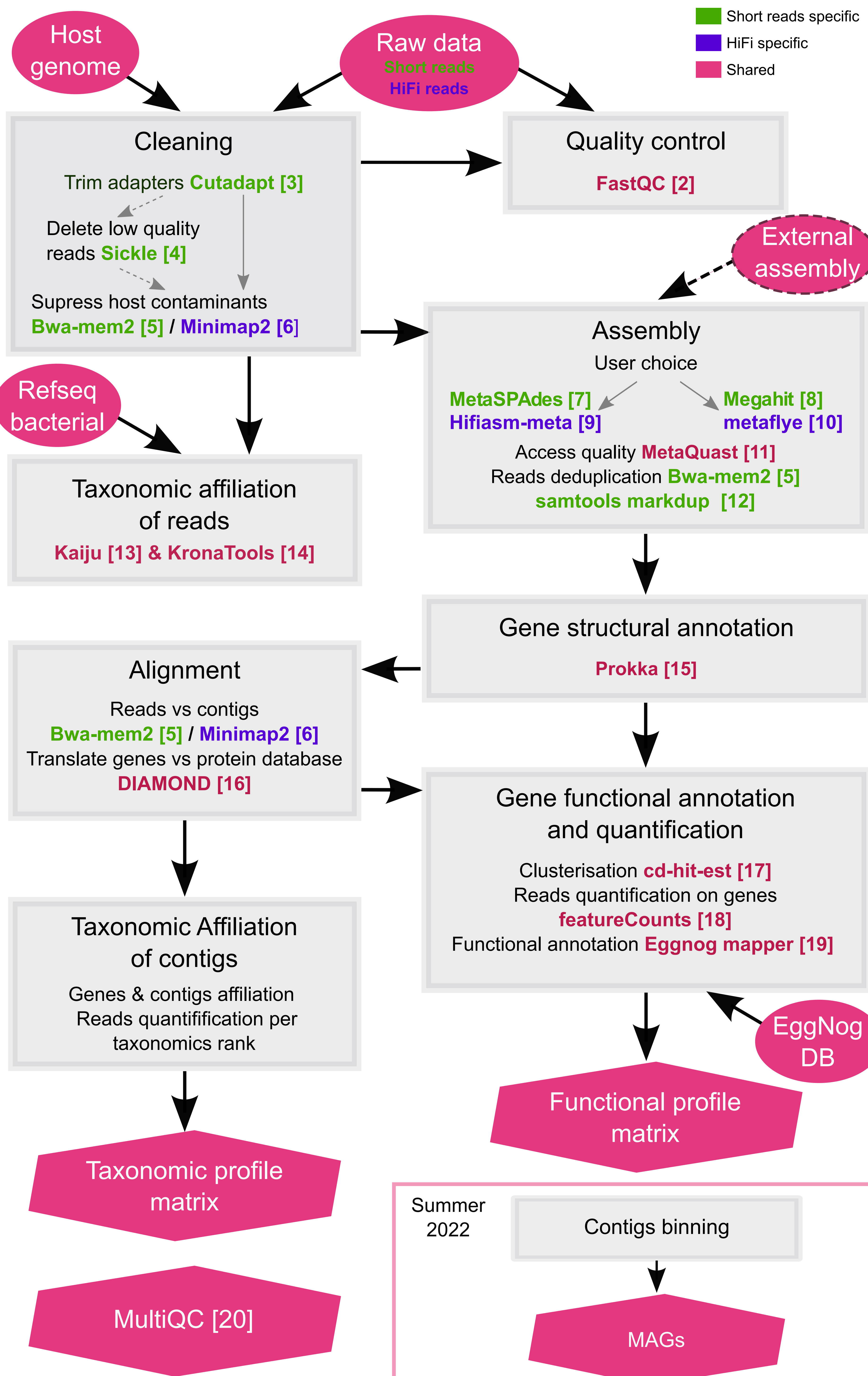
² INRAE, GeT-PlaGe, Genotoul – INRAE – 31326 Castanet-Tolosan, France (doi: 10.15454/1.5572370921303193E12)

³ GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

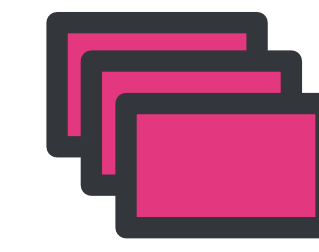
* These two authors contributed equally to this work

Corresponding author: claire.hoede@inrae.fr

Production of whole metagenome assembly, functional and taxonomic profile

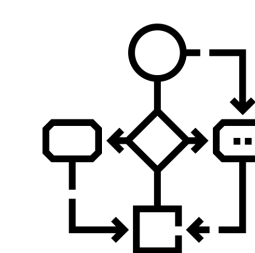


Workflow features



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 **S**ingularity [21] container

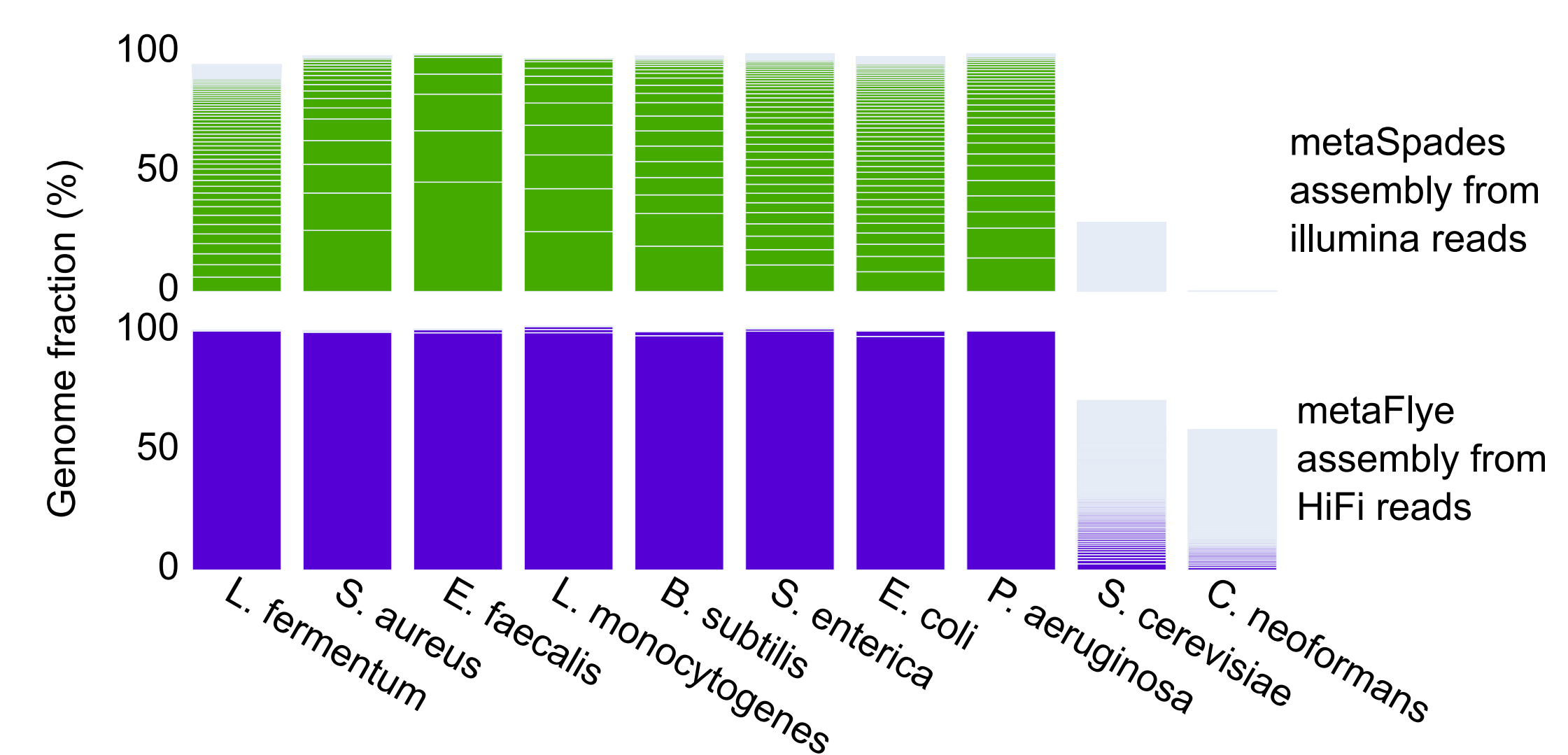


Fully documented

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

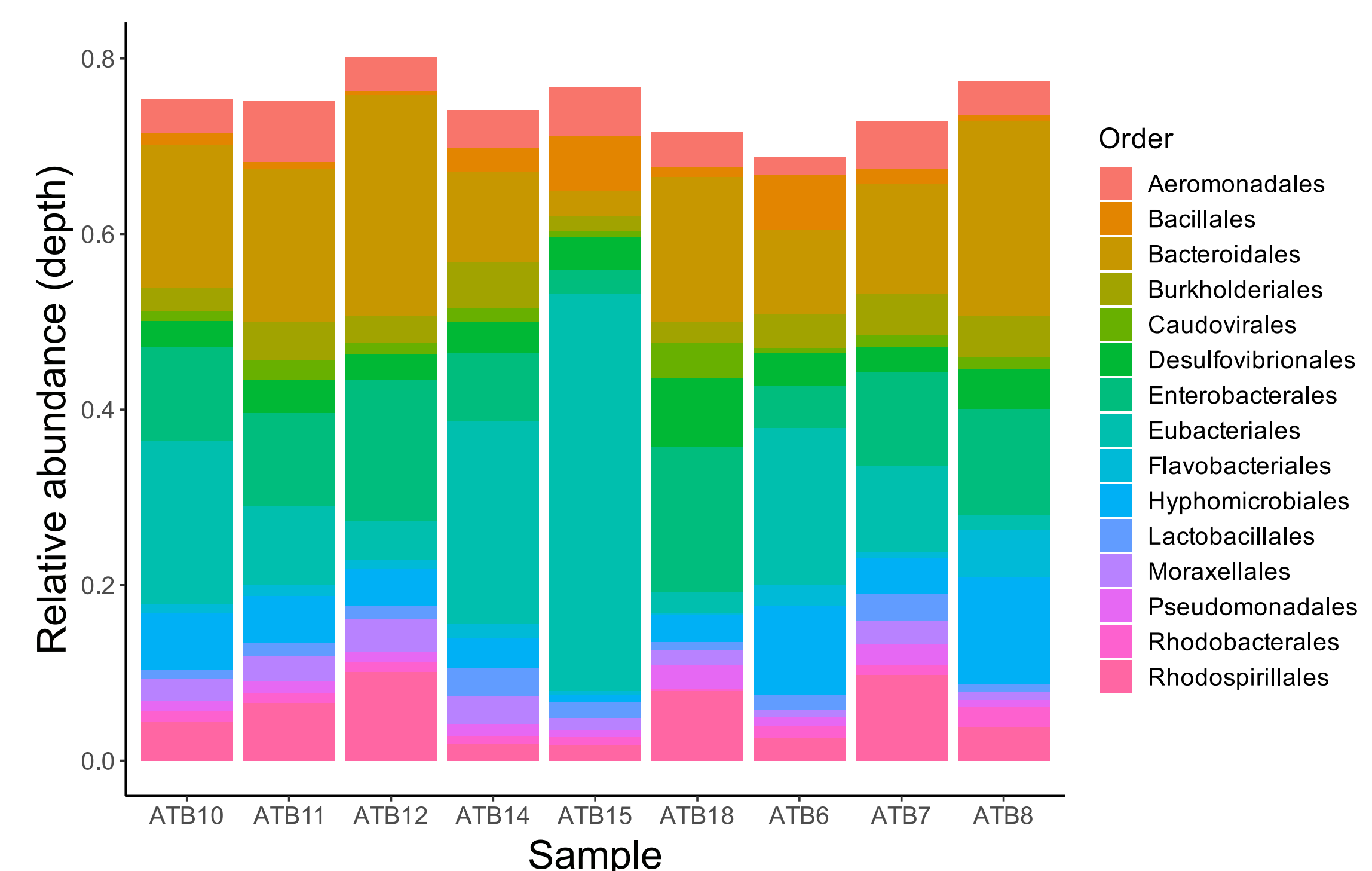
Results examples

HiFi vs Short reads assembly



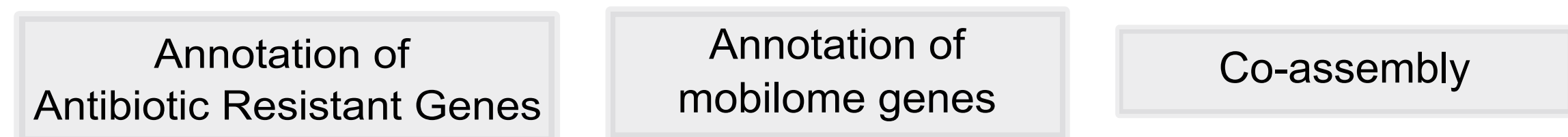
Bacterial genomes of mock Zymobiomics are assembled in one contig by HiFi reads. The genome fraction is the fraction of the reference genomes covers by the assembly. Each block in a bar represents a contig. Bacterial genomes are well covered by all assemblies but Illumina assembly is much more fragmented.

Taxonomic profile



Relative abundance of the 15 major orders of nine biofilms grown in bioreactors from sewage (example from ATB biofilm project). Graph made from the taxonomic profile matrix generated by metagWGS. ATB15 is a sample that was exposed to ciprofloxacin (antibiotic) during the 14 days of culture.

Perspectives



Acknowledgements

SeqOccln financed by FEDER funds (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020)

ATB_Biofilm funded by PNREST Anses, 2020/01/142

- [1] P. Di Tommaso et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol.*, 35(4):316-319, 2017.
- [2] S. Andrews. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, 2010.
- [3] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10-12, 2011.
- [4] NA Joshi, JN Fass. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software]. Available at <https://github.com/najoshi/sickle>, 2011.
- [5] H. Li et al. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754-60, 2009.
- [6] H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094-3100, 2018.
- [7] S. Nurk et al. MetaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, 27(5):824-834, 2017.
- [8] D. Li et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674-6, 2015.
- [9] H. Cheng. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, 18:170-175, 2021.
- [10] M. Kolmogorov. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, 17,1103-1110, 2020.
- [11] A. Mikheenko. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, 32: 1088-1090,2016
- [12] P. Danecek. Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008, 2021.
- [13] P. Menzel et al. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun.*, 7:11257, 2016.
- [14] BD Ondov et al. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385, 2011.
- [15] T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068-9, 2014.
- [16] B. Buchfink. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*, 18: 366-368, 2021.
- [17] L. Fu et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150-2, 2012.
- [18] Y. Liao et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30, 2014.
- [19] C. Cantalapiedra. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 38: 5825-5829, 2021
- [20] GM Kurtzer et al. Singularity: Scientific containers for mobility of compute. *PLoS One*, 12(5), 2017.
- [21] P. Ewels et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047-8, 2016.