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

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

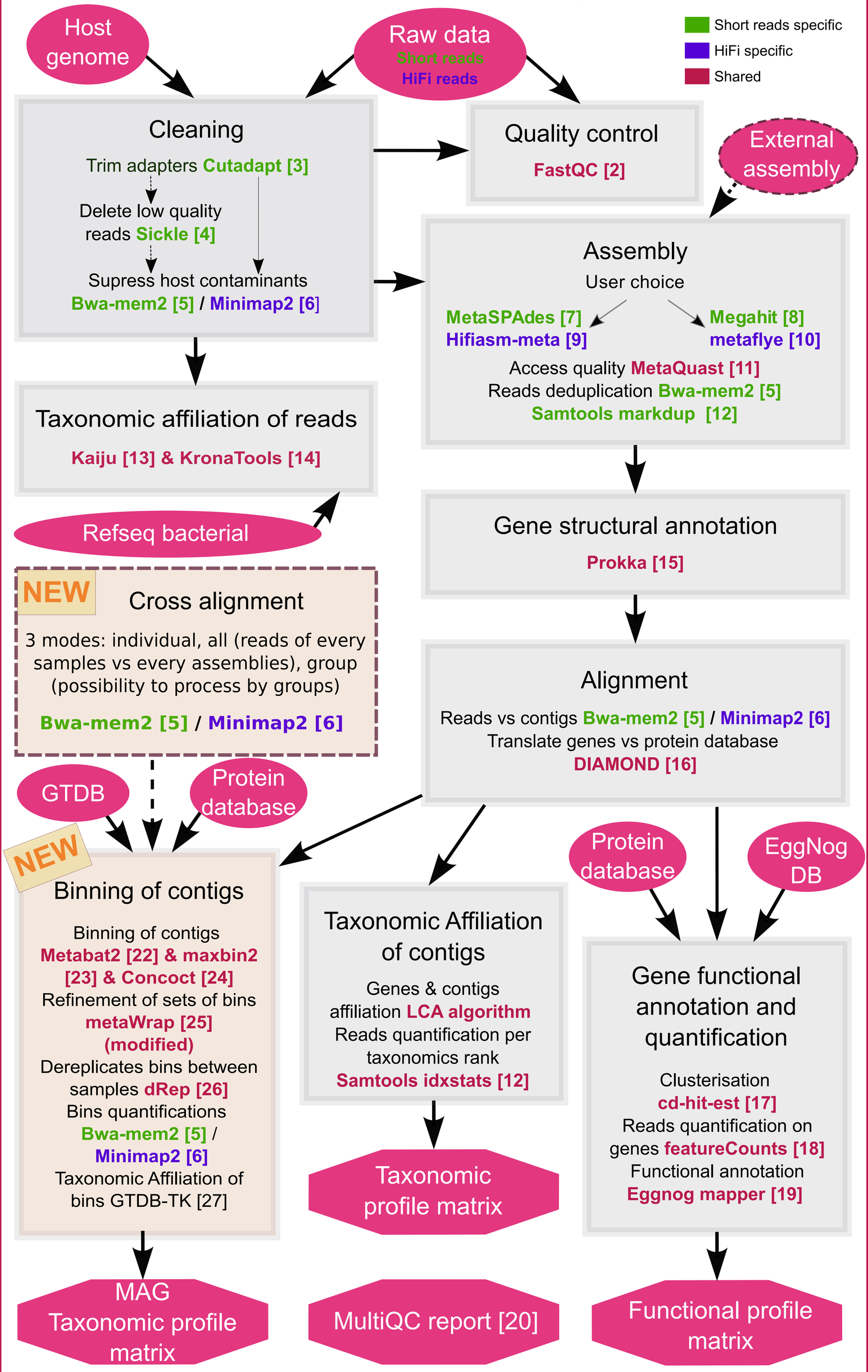
metagWGS: a workflow to analyse short and long HiFi metagenomic reads

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

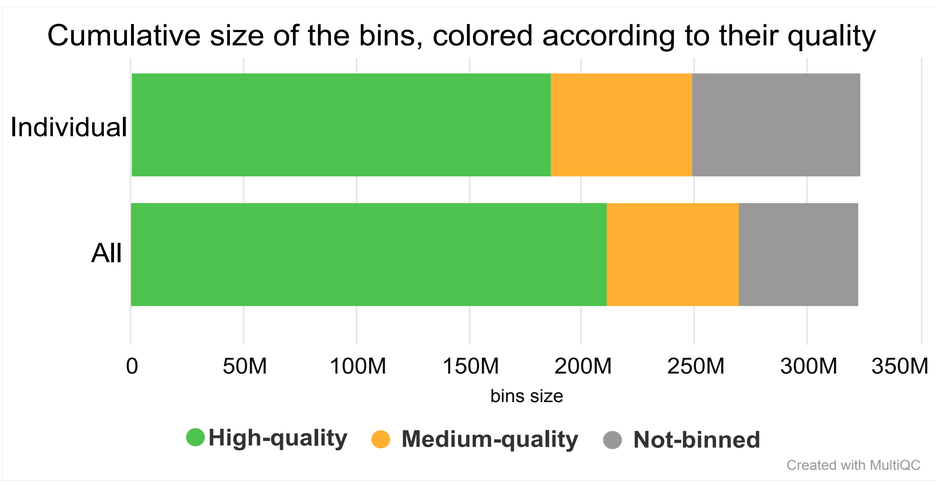


Workflow features

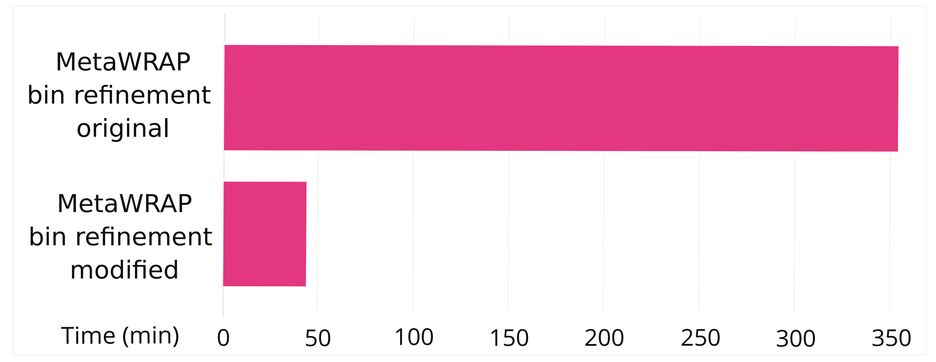
Type of NGS data:
whole genome shotgun sequencing (Illumina HiSeq3000 or NovaSeq, paired, 2*150bp ; PacBio HiFi reads, single-end)

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).



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

[1] Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. Nat Biotechnol 35, 316–319 (2017).
[2] Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. (2010).
[3] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12 (2011).
[4] Joshi, N. A. & Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (2011).
[5] Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS) 314–324 (2019).
[6] Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
[7] Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834 (2017).
[8] Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676 (2015).
[9] Feng, X., Cheng, H., Portik, D. & Li, H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. Nat Methods 19, 671–674 (2022).
[10] Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17, 1103–1110 (2020).
[11] Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 32, 1088–1090 (2016).
[12] Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
[13] Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7, 11257 (2016).
[14] Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12, 385 (2011).
[15] Seemann, T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069 (2014).
[16] Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12, 59–60 (2015).
[17] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152 (2012).
[18] Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930 (2014).
[19] Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38, 5825–5829 (2021).
[20] Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016).
[21] Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. PLoS One 12, e0177459 (2017).
[22] Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359 (2019).
[23] Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607 (2016).
[24] Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat Methods 11, 1144–1146 (2014).
[25] Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6, 158 (2018).
[26] Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11, 2864–2868 (2017).
[27] Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927 (2020).
[28] Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods 20, 1203–1212 (2023).
[29] Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055 (2015).

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