Title	YAMP : Yet Another Metagenomic Pipeline
Author	*Alessia Visconti*, Tiphaine C Martin, Mario Falchi
Affiliation	Department of Twins Research & Genetic Epidemiology, King's College London
Contact	alessia.visconti@kcl.ac.uk
URL	https://github.com/alesssia/YAMP
License	GNU GPL 3

Thanks to the increased cost-effectiveness of high-throughput technologies, the number of studies focusing on microorganisms (bacteria, archaea, microbial eukaryotes, fungi, and viruses) and on their connections with human health and diseases has surged, and, consequently, a plethora of approaches and software has been made available for their study, making it difficult to select the best methods and tools.

Here we present *Yet Another Metagenomic Pipeline* (YAMP) that, starting from the raw sequencing data and having a strong focus on quality control (QC), allows, within hours, the data processing up to the functional annotation. Specifically, the QC (performed by means of several tools from the BBmap suite [1]), allows de-duplication, trimming, and decontamination of metagenomics (and metatranscriptomics) sequences, and each of these steps is accompanied by the visualisation of the data quality. The QC is followed by multiple steps aiming at characterising the taxonomic and functional diversity of the microbial community. Namely, taxonomic binning and profiling is performed by means of MetaPhlAn2 [2], which uses clade-specific markers to both detect the organisms present in a microbiome sample and to estimate their relative abundance. The functional capabilities of the microbiome community are currently assessed by the HUMAnN2 pipeline [3] which first stratifies the community in known and unclassified organisms using the MetaPhlAn2 results and the ChocoPhlAn pan-genome database, and then combines these results with those obtained through an organism-agnostic search on the UniRef proteomic database. The next YAMP release, currently under development, will also support MOCAT2 [4] and an optimised version of the HUMAnN2 pipeline. QIIME [5] is used to evaluate multiple diversity measures.

YAMP is constructed on Nextflow [6], a framework based on the dataflow programming model, which allows writing workflows that are highly parallel, easily portable (including on distributed systems), and very flexible and customisable, characteristics which have been inherited by YAMP. Users can decide the flow of their analyses, for instance limiting them to the QC or using already QC-ed sequences. New modules can be added easily and the existing ones can be customised – even though we have already provided default parameters deriving from our own experience. While YAMP is developed to be routinely used in clinical research, the expert bioinformaticians will appreciate its flexibility and modularisation. YAMP is accompanied by a Docker container [7], that saves the users from the hassle of installing the required software, increasing, at the same time, the reproducibility of the YAMP results.

References

1. https://sourceforge.net/projects/bbmap
2. Truong, D.T., et al. *Metaphlan2 for enhanced metagenomic taxonomic profiling.* Nature methods 12(10), 902–903 (2015)
3. https://bitbucket.org/biobakery/humann2
4. Kultima, J.R., et al. *MOCAT2: A metagenomic assembly, annotation and profiling framework.* Bioinformatics 32(16), 2520–2523 (2016).
5. Caporaso, J.G., et al. *QIIME allows analysis of high-throughput community sequencing data.* Nature Methods 7(5), 335–336 (2010)
6. Di Tommaso, P., et al. *Nextflow enables reproducible computational workflows.* Nature Biotechnology 35, 316–319 (2017)
7. https://www.docker.com