Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Department of Animal Breeding and Genetics

Novel workflow for Metagenomics and transcriptomics analysis of Anaerobic Digestive systems.

Renaud Van Damme

Independent project in Biology • 30 credits
Uppsala 2020

Novel workflow for Metagenomics and transcriptomics analysis of A.D. systems.

Renaud Van Damme

Supervisor:	Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, section Bioinformatics
Assistant supervisor:	Bettina Müller, Swedish University of Agricultural Sciences, Department of Molecular Sciences
Assistant supervisor:	Christian Brandt, Swedish University of Agricultural Sciences, Department of Molecular Sciences
Examiner:	Johan Dicksved, Swedish University of Agricultural Sciences, Department of Animal Nutrition and Management

Credits:	30 credits
Level:	Second cycle, A2E
Course title:	Independent project in Biology, A2E
Course code:	EX0871
Course coordinating department:	Department of Animal Breeding and Genetics
Place of publication:	Uppsala
Year of publication:	2020
Online publication:	https://stud.epsilon.slu.se
Keywords:	Metagenomics, transcriptomics, pipeline, A.D systems, Biogas, Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Abstract

The A.D. systems (anaerobic digestion), when used in biogas reactors, are an advanced ecological way to produce energy while treating waste. The majority of the microbial community of the reactor remains unknown to this day, due to the impossibility to culture most of the bacteria individually. Metagenomics and transcriptomics aim to discover those bacteria and understand the interactions within the community. HTS (high throughput sequencing) technology opens new possibilities in terms of length of the reads sequenced and accuracy. Sequencing done by Oxford Nanopore machines can produce long reads while having a slightly worse accuracy than other machines, where Illumina sequencing machines have a higher accuracy to the detriment of lengths. The two sequencing methods complement each other, and the hybrid assembly uses both long and short reads to create longer and more accurate contigs that can then be further analysed.

Here is presented a metagenomics pipeline (MUFFIN) based on the hybrid assembly of short and long reads followed by multiple differential binning methods and refinement to produce high-quality bins and their annotations. The pipeline is written by using Nextflow to achieve high reproducibility and fast and straightforward use of the pipeline. This pipeline also produces the taxonomic classification of the bins as well as a transcription, quantification and annotation of RNAseq data. The pipeline was tested using one biogas reactor as an example to assess the capacity of MUFFIN to process and output relevant files needed to analyse the microbial community and their function. A parsing script was developed to analyse and summarise the annotations files. The script outputs a quantification file of the transcripts annotated, an HTML file summarising the pathways across the bins and transcripts, and an HTML file for each bin summarising the annotation.

Keywords: Metagenomics, transcriptomics, pipeline, A.D systems, Biogas

Table of contents

List of tables

List of figures

Abbreviations

16s	16s rRNA sequencing
A.D. system	Anaerobic digestion systems
Cloacimonetes	*candidatus* cloacimonetes
HTS	High Throughput Sequencing
MAGs	Metagenome-assembled genomes
mRNA	messenger RNA
NGS	Next-generation sequencing
RNAseq	RNA sequencing
SNVs	Single nucleotide variants

1 Introduction

To treat any kind of organic wastestreams, different methods are available where each of them has pros and cons (Eriksson, Strid, and Hansson 2015; Arafat, Jijakli, and Ahsan 2015). Amongst those, one stands out in terms of low environmental impact as it goes along with sustainable energy production: This method is called engineered anaerobic digestion (A.D.), that use bacteria and archaea to degrade different kinds of organic waste while producing, e.g. methane (biogas) in so-called biogas reactor (Wellinger, Murphy, and Baxter 2013; Atelge et al. 2018).

The biogas plants are used to produce biomethane (methane from a biological source) and are implanted in various countries of the world as a sustainable alternative for energy production. These countries include Germany, Italy, Sweden, Finland, France, Belgium (Torrijos 2016) but also China, India, Canada, and other countries (Raboni and Urbini 2014). The global use of A.D. systems, make the study of the microorganisms and their interactions in those worldwide systems a potential key to understanding the function of lesser known bacteria. Retrieving the genome and functions of those bacteria could lead to an increase of the production as well as some new critical discoveries. Biogas can be generated by using different organic resources such as agricultural waste, sewage sludge, manure, industrial food waste, organic household waste and crops. Methane can be either upgraded to biofuel or used to produce electricity or heat.

The production of methane by microorganisms is called methanogenesis and is realised by methanogenic archaea in strictly anaerobic conditions. However, the whole anaerobic degradation process into methane and carbon dioxide is more complex and requires the harmonised and combined activities of a vast number of different microorganisms. It involves multiple trophic levels, responsible for depolymerization, primary and secondary fermentation, acidogenesis, acetogenesis, and methanogenesis (Pelletier et al. 2008). The microbial community of the reactor should be complementary and depends strongly on syntrophic interaction in order to complete the entire degradation (Solli et al. 2014). Thus, knowing the composition of the microbial communities of the reactors helps to understand the metabolic mechanism and interactions and also helps in the optimisation of biogas production.

The analysis of a microbial community relies on the use of metagenomics analyses as most of the microorganisms cannot be cultivated for individual analysis. The use of metagenomics already much helped in the discoveries of new bacteria, belonging even to new, undescribed phyla.

For example, one of the new phyla discovered through metagenomics analysis is the “Candidatus Cloacimonetes” phylum, which has been deduced from the genome reconstruction of “Candidatus Cloacamonas acidaminovorans” (Pelletier et al. 2008). Candidatus Cloacimonetes has been found at significant abundances in different biogas reactor samples; ranging from 10% to 15% (Botello Suárez et al. 2018; Lee et al. 2018; Solli et al. 2014; Pelletier et al. 2008). The use of metagenomics and high throughput sequencing are prerequisites as most of the unknown bacteria are unculturable as they can be profoundly complex to be cultured and might require the presence of other microorganisms (Steen et al. 2019). The phylum ‘Candidatus Cloacimonetes” might be involved in the degradation of organic waste and also involved in the methanogenesis step: a hypothesis endorsed by the increase of population through time in biogas reactor (Solli et al. 2014). Investigating the potential role of such bacteria in this complex degradation can be crucial to the optimisation of the production of biogas.

Summarised, metagenomics and transcriptomics analyses are critical elements in research when it comes to unculturable bacteria and their functions and interactions within complex microbial consortia (Parks et al. 2017; Sunagawa et al. 2015). In that sense, reproducibility and convenient handling of such bioinformatics analyses are of crucial importance for scientific research since it lightens the bioinformatics workload put on the researcher.

Metagenomics analysis is the sequencing of a microbiome without distinction/selection of a specific organism. Using tools specific for metagenomics, we can reconstruct and polish “potential” organisms each of those is then compared to known organisms to assess their existence, potential existence (through similarities) or if they are error due to the process of creation. The potential obtention of information about uncultured and unknown bacteria as well as about the functional potential of known bacteria, make metagenomics a suitable analysis method in this specific study.

2 Background

2.1 Biogas reactor

A biogas reactor is a fermentation chamber with a controlled environment used to produce biogas. This consists of the main chamber equipped with different sensors to control the environment, a heating system to maintain the optimal temperature, a gas exit to harvest the biogas produced and a matter entry to input the organic matter. Biogas reactors can range from household-scale (China) to large-scale as typically found in Europe. For research purposes, lab-scale reactors with a volume ranging from one to five L can be used to mimic large-scale processes in order to explore the relationship between microbial community, function and process performance. A variety of feedstocks can be used ranging from agricultural waste, industrial waste from food production, organic household waste, to sewage sludge. The gas can then be stored for external usage or used directly to produce electricity and heat or upgraded to biofuel.

2.2 The anaerobic food chain in biogas processes

The methane production requires the collaboration of diverse trophic levels, including de-polymerization, primary and secondary fermentation, acido-genesis, aceto-genesis, and methanogenesis (Pelletier et al. 2008).

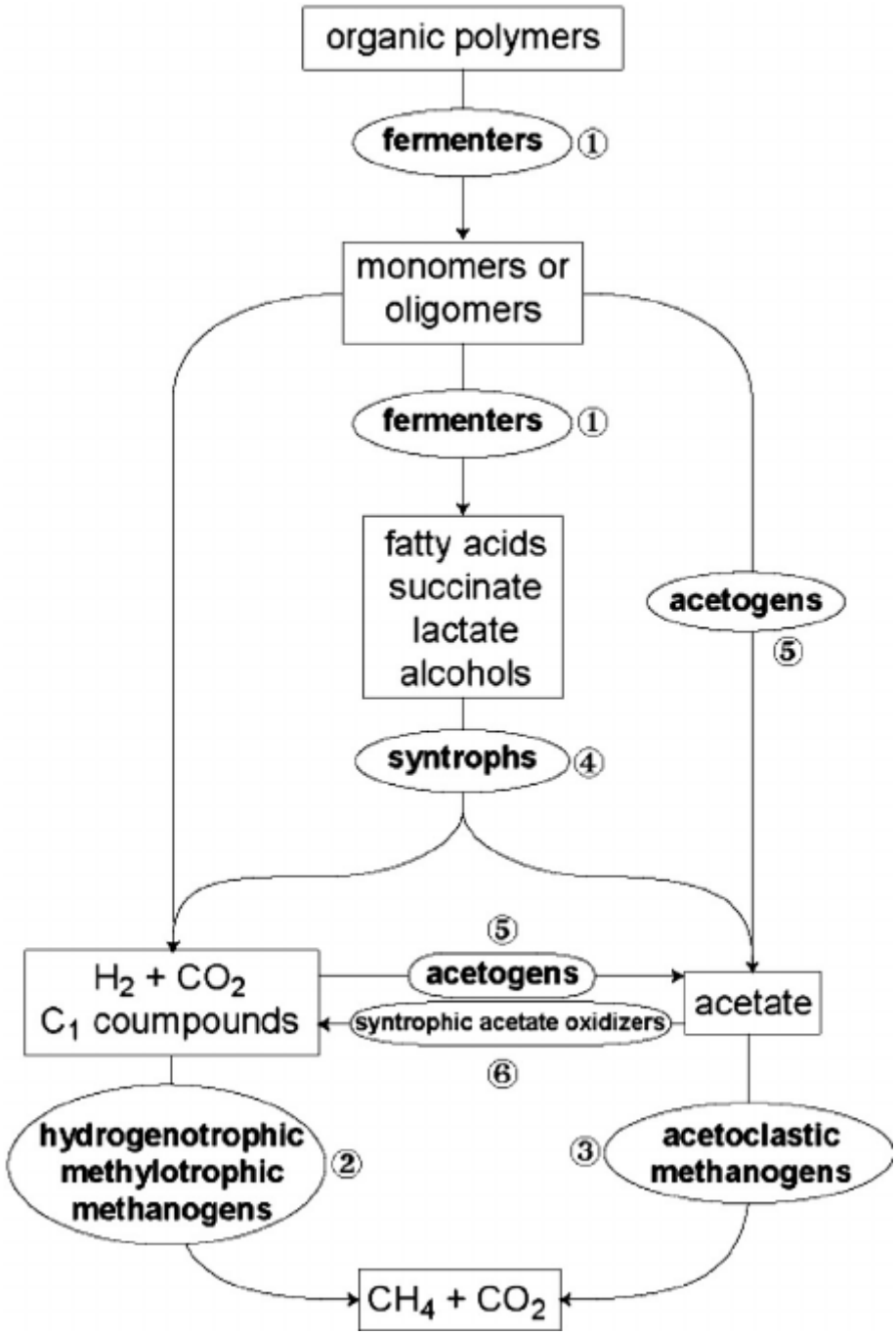
The de-polymerization, also called hydrolysis, is involved in the reduction of complex and large organic compounds into smaller and simpler compounds (such as peptides, amino acids, fatty acid, sugars), which will be further digested in the following steps (Angelidaki et al. 2011).

The acidogenesis is the transformation of amino acids and sugar into hydrogen and diverse organic acid , required by the acetogenesis to produce the next intermediate product.

The acetogenesis is both the synthesis of acetate from the reduction of carbon dioxide or further oxidation of organic acids produced in the previous steps (Ragsdale and Pierce 2008).

The methanogenesis is the last step in the production of methane from organic waste. It uses either acetate or hydrogen and carbon dioxide to produce methane and carbon dioxide. The two different pathways occur depending on the present Archaea, which produces methane. The first is the use of carbon dioxide with hydrogen to produce methane. It can be produced by different groups of Archaea but is under low ammonia conditions not the most productive. The second is the cleavage of acetate into carbon dioxide and methane. It is estimated that two-thirds of the methane produced globally comes from this reaction; only a few genera are known to use this pathway (Liu and Whitman 2008).

Figure 1 - Schematic of the anaerobic degradation of organic matter into methane. Source https://www.researchgate.net/figure/Schematic-anaerobic-food-chain-for-the-conversion-of-complex-organic-matter-to-methane-in_fig1_250924004 *(Mesle, Dromart, and Oger 2013)*



2.3 Cloacimonetes

The “Candidatus Cloacimonetes” phylum is present in different anaerobic environment partly up to 10% to 15% of the environmental bacterial population according to some articles and has been found at partly high abundance in WWTP and biogas plants (Solli et al. 2014; Botello Suárez et al. 2018; Lee et al. 2018). It could represent a new bacterial division that is up to 10% of the bacterial community (Pelletier et al. 2008). The “Candidatus Cloacamonas acidaminovorans” genome reconstructed in 2008 provided the first evidence of this potential new division. As this bacterium is not culturable and did not receive much interest in the past, there is little information about it. However, this phylum might be of great importance to the biogas processes. There are indications that they are involved in syntrophic interactions and it was found to be present in many anaerobic degradation systems that revolve around the fermentation of amino acids (Pelletier et al. 2008).

2.4 Sequencing approaches

In this study, we sequenced metagenomes using three different methods. An Oxford Nanopore MinIon sequencing, Illumina DNA Miseq sequencing, and RNA sequencing using Illumina Miseq.

2.4.1 Oxford Nanopore MinIon

The MinIon sequencing is a long-read sequencing method that consists of the library preparation of the sample, followed by the direct sequencing of the sample. The sequencing is not synthesis based, but it sequences by passing through the DNA strands through pores where all bases are read in real-time by the pore mentioned.

This method allows the sequencing of longer reads since there is no limitation due to a synthesis of the read. After the sequencing comes the basecalling execution that converts the signal received from the sequencing machine to nucleotides, this method eliminates the PCR bias as there is no PCR amplification. The use of long reads also circumvents issues in the reconstruction of genomes since the reads are longer and issues like repeats, gaps or contamination have less probability of influencing the assembly of the reads. However, this sequencing method also has disadvantages, such as a lower precision on the base level. While it is easier to assemble and map long sequences of DNA together, the lack of precision on the base level makes the MinIon a lower argument when it comes for instance, to single nucleotide variants (SNVs).

2.4.2 Illumina Sequencing

The Illumina sequencing is a short-reads sequencing method that consists of the library preparation (including shearing, PCR, adding adapters). Then in the sequencing machine, clusters of strands are created, followed by the synthesis of the strand clusters start. To each nucleotide-binding event, fluorescent light is emitted, and the reading of this sequential colour emission creates the reads. This method has many restrictions in terms of reads size as well as speed. However, it is more accurate on the base level than the long read NGS, which enables analysis of SNVs on a better and more accurate level.

2.4.3 RNA sequencing

RNA represents the active functions of the cell, where the DNA represents all the information of the organism (all structure and function the organism contains). There are mainly three essential types of RNA involved in the creation of the proteins that serves the activity of the cell. The mRNA that encodes the sequence of amino acids translated in proteins, the rRNA when combined with ribosomal proteins, forms the ribosomes which translate the mRNA into proteins and the tRNA that transport amino acids to the ribosome during the translation (Bastide and David 2018).

The more traditional RNA sequencing consists of the sampling of genetic material followed by isolation of the total RNA and removal of any residual DNA by DNase digestion. According to the RNA targeted (mRNA, rRNA), the use of specific beads can be executed. This allows keeping only the targeted RNA, for instance, the mRNA. The RNA is then reverse transcripted to obtain cDNA(Sessitsch et al. 2002) that will be sheared, amplified and then sequenced in a short-read sequencing machine.

The use of long-read sequencing machine can significantly reduce the library preparation, as it does not require any PCR amplification. When the sequencing of RNA is done in a Nanopore sequencing machine, the use of cDNA is not mandatory, and the use of native RNA is possible.

2.4.4 Metagenomics

Metagenomics in a broad term, includes two different methods to analyse the population of a microbiome. One is the whole metagenomics shotgun sequencing, and the other is the 16s rRNA gene amplicon sequencing (Ghosh, Mehta, and Khan 2019).

Whole metagenomics shotgun sequencing consists of the sequencing of all DNA information of a microbiome sample without any isolation or culture of a specific organism.

The 16s rRNA gene sequencing is not a metagenomics method since the purpose is not to retrieve genomes of organisms in the microbiome. Nevertheless, it aims to identify organisms present in a microbiome by relying on the taxonomic information obtained from partial sequences of the 16s rRNA genes. The 16s method is based on the sequencing of amplicons retrieved by PCR of the 16s rRNA gene of all the organisms present in the sample. The procedure includes DNA extraction followed by a 16s rRNA gene amplification (Nurul et al. 2019) and sequencing (long or short reads).

2.5 Hybrid Assembly

A hybrid assembly is an assembly approach that uses both long and short reads. The assembly of long-reads alone is useful to avoid repeats and gaps in the reconstruction of the genomes, but it also has flaws like the higher error rate on a base level, ranging from 15% to 40% (Ma et al. 2019). The short reads assembly does not possess such an error rate and thus is useful for a base level analysis. In the case of the short reads assembly, the flaws are the gaps and repeats.

The hybrid assembly is tentative to combine the advantages of both sequencing methods to produce assemblies/genomes of a higher quality while trying to avoid their respective disadvantages.

2.6 Reproducibility

The reproducibility and ease to analyse are critical to scientific research. Automated pipelines are developed to lighten the charge of informatic work put on the searcher. Various pipelines already exist to automate the research (e.g., the nf-core collection of pipelines (Ewels et al. 2019)). They are based on workflow management systems such as nextflow or snakemake, and those management systems allow to create from scratch a pipeline but also make through the use of software containers that have everything ready for the use of the pipeline. Another advantage is the possibility to parallelise the work to speed up the process but also to use the workflow on high performance computing clusters and clouds. Making those pipelines highly portable, adaptable, powerful and easy to use.

2.7 Nextflow

Nextflow (Tommaso et al. 2017) is a workflow management system allowing high reproducibility through the use of software containers (such as docker or singularity). It is also oriented to the portable optimisation, and the pipeline is separated from the configuration of the system at use. Nextflow is developed to work on most HPC and server executors (SGE, SLURM,…) and also on cloud computing (Google Life Science, Amazon AWS). Nextflow is an efficient workflow management system with simplified utilisation both as developers and users. Indeed, it uses a global DSL regarding the construction of the pipeline while at the same time allowing the use of various programming languages (Python, Perl, R, Ruby) and scripting language (Bash script). It also provides tools to abstract and manages file naming in global variables to reduce the ambiguity (Leipzig 2017).

Part of the development of Nextflow is to create a new syntax that aims to simplify the conception and use of pipelines by changing the creation of a unique process for each task to the invocation of the said process from modules in a specific order. A module is simply a function or task saved in a different file that can be called in the main script. The creation of “modules” that can be used multiple times and use in different pipelines without the need to rewrite everything is the key to simplification.

3 Aim

This work aimed at the creation of a metagenomics and transcriptomics pipeline for microbial analysis. Moreover, it will be tested on the analysis of an anaerobic digestion system. The pipeline shall be able to be run by anyone that has access to a computer with basic Linux knowledge and biological data of interest.

It shall produce helpful and informative result files for the microbial analysis of an environmental sample or specific bacteria of interest. To achieve this, different objectives were decided:

- Find or create an ergonomic, automated, and reproducible analysing pipeline that would be able to combine the information of both the Illumina and MinIon sequencing.
- Obtain Metagenome-assembled genomes (MAGs) of good quality from this pipeline, as well as useful taxonomic classifications and functional annotations.
- Through the use of RNAseq, data obtain good quality and complementary transcriptomes.
- Access quickly results browsing summary files from the different objectives mentioned above.

4 Material and method:

4.1 Sample used

Out of all the samples prepared by Christian Brandt and Bettina Müller, only one DNA sample (Nanopore and Illumina) and one RNA sample (Illumina) from the same reactor was used in the Pipeline and will have the result display in “Results”. The samples are from a biogas reactor present in Uppsala. The following chapters (4.2 to 4.4) describe the extraction, library preparation and sequencing of different samples processed at the same time. In total for the DNA (nanopore and Illumina), 20 samples from 20 different biogas reactor (10 Swedish and 10 Germans) where sequenced. For the RNA, six samples from five different reactors where sequenced.

4.2 Nanopore DNA extraction, library preparation, and sequencing

DNA extraction, library preparation, and sequencing were done by Christian Brandt (postdoc at SLU) for Nanopore sequencing.

This protocol for DNA extraction and Nanopore sequencing can be found in the submitted manuscript of Christian Brandt article 10.21203/rs.2.17734/v1 (Abundance Tracking by Long-Read Nanopore Sequencing of Complex Microbial Communities in Samples from 20 Different Biogas/Wastewater Plants).

All samples were sequenced using a MinION Sequencer for 72 hours or until no sequencing activity was observed, using either an R.4.9.1 or R.4.9 flow cell (FLO-MIN106) for each sample. The MinKNOW software was used with active channel selection enabled and basecalling deactivated. A ‘flow cell-refuel’ step after approx. 18-20 hours of runtime by adding 75 µL of a 1:1 water-SQB-Buffer mixture to the flow cell via the SpotON port. SQB-Buffer is part of the Oxford-Nanopore SQK-LSK109 Kit.

4.2.1 Basecalling

Basecalling was performed using the GPU accelerated guppy basecaller with the high accuracy model and adapter trimming (available at https://nanoporetech.com).

4.3 Illumina DNA extraction, library preparation, and sequencing

Illumina sequencing was performed using the same DNA material (purified by Christian Brand) as it was used for nanopore sequencing.

4.4 mRNA extraction, library preparation, and Illumina sequencing

mRNA preparation was done by Bettina Muller (associate Professor at SLU) as described in (Manzoor et al. 2016). 200 mg fresh digester sludge has been used as starting material.

Library preparation and sequencing were performed by Scilifelab using the TruSeq stranded mRNA library preparation kit (Illumina Inc.). In total, six mRNA pools were sequenced on one Miseq lane. After the cluster generation, the sequencing was done and was a 75 cycles paired-end sequencing in one run.

4.5 Bioinformatic Workflow:

Having a functional and reproducible workflow to analyse the sample is essential as the complexity of the metagenomics tools and the interconnections between them is not always straightforward. Various workflows are already available such as MetaWRAP(Uritskiy, DiRuggiero, and Taylor 2018), Anvi'o (Eren et al. 2015), SAMSA2 (Westreich et al. 2018), Humann (Abubucker et al. 2012) or MG-RAST (Meyer et al. 2008) but none of them uses a hybrid approach. Creating a pipeline was the solution.

4.5.1 Making the pipeline

The pipeline should have a hybrid approach of the assembly of the reads, should be versatile (run on different Unix systems and configuration), easy to use, parallelised and should not require multiple additional installations.

To address the versatility and the parallelisation, the use of nextflow (Tommaso et al. 2017) as the workflow manager system appeared to be the best choice. It provides an abstraction layer making the pipeline an unspecific script with the configuration related to the platform used independently.

The use of Docker for the software containers makes the pipeline reproducible and not sensitive to the machine or software versions. Docker loads the required container, executes the software, output the result and closes itself, with no version

control or compatibility issues. In addition to Docker, we use conda as an environment manager. It creates a dedicated environment for the software to run, installs it, runs and done.

4.5.2 The pipeline (MUFFIN)

The pipeline called MUFFIN consists of three different steps that can be run together or independently; the steps are “assemble,” “classify” and “annotate.” For the paper about MUFFIN, see Appendix n°3.

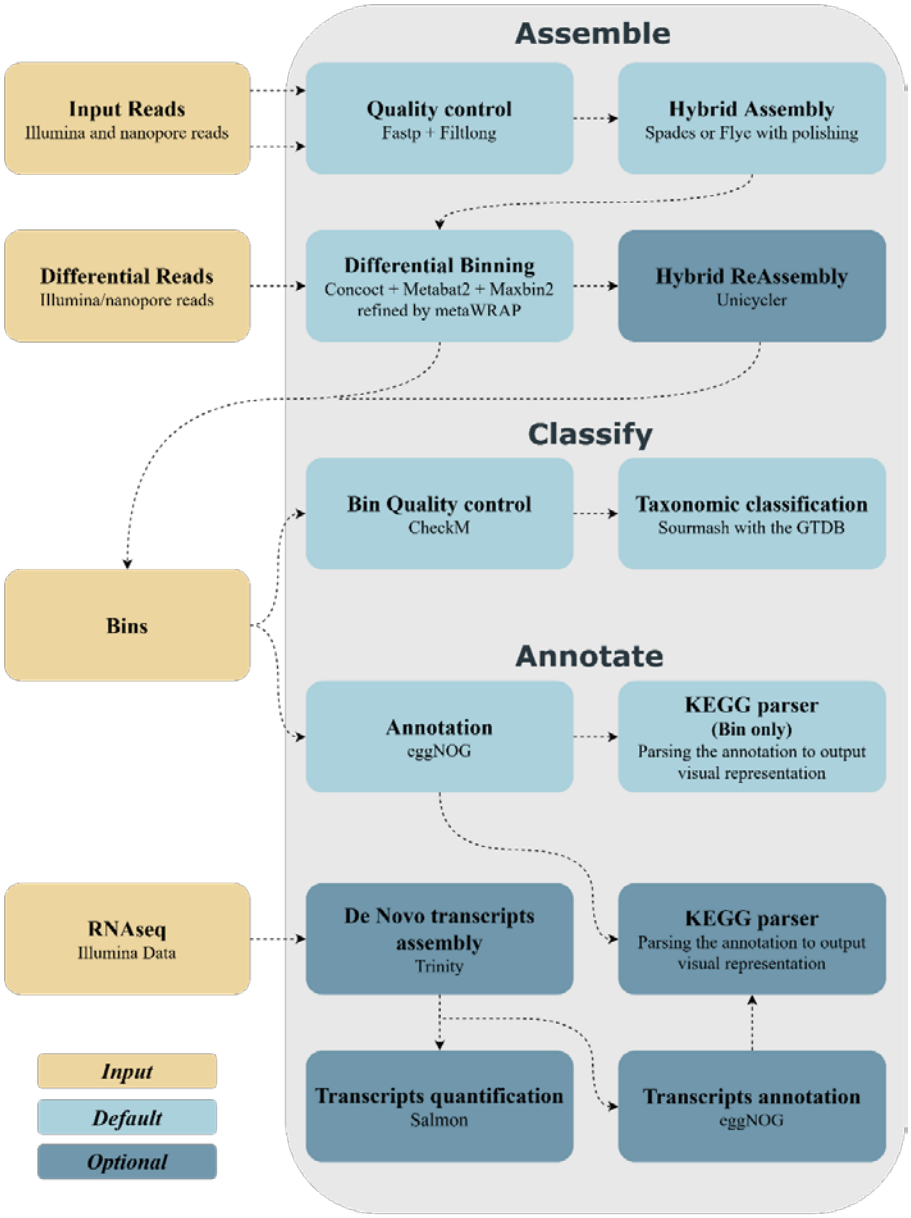


***Figure 2**- The chart of the MUFFIN pipeline.*

4.5.3 Assemble

Assemble is the first step; it requires the Illumina and nanopore reads as input.

The first substep is the quality control, for the Illumina data by fastp (Chen et al. 2018), and for the nanopore, the default is a discard of the shortest reads (under 2000bp) and the use of Filtlong (https://github.com/rrwick/Filtlong) as an option.

The second substep is the assembly of the reads. Two different assembly methods are available. The one used in the example is the metagenomic and hybrid approach of SPAdes (Bankevich et al. 2012; Nurk et al. 2017). The other method available is the long read assembly using Flye (Kolmogorov et al. 2019) metagenomics approach. Flye is followed by a polishing of the contigs with, Racon (Vaser et al. 2017) , medaka (https://github.com/nanoporetech/medaka) and Pilon (Walker et al. 2014).

The third substep is the binning of the contigs obtained. Three different binning methods followed by a refining of the bins compose the substep. The binning methods are CONCOCT (Alneberg et al. 2014), a binning method based on nucleotide composition – kmer frequencies and coverage data, MaxBin2 (Wu et al. 2014) using depth-of-coverage, nucleotide composition, and marker genes and MetaBAT2 (Kang et al. 2015) an adaptive binning algorithm. CONCOCT and MetaBAT2 can if provided, accept additional reads set to improve the binning through the use of differential binning. The result of those three binning is then inputted in the refining step of the MetaWRAP pipeline (Uritskiy, DiRuggiero, and Taylor 2018).

Once those bins obtain, an optional re-assembly substep remains. This substep consists of the mapping of the reads against the bins using SAMtools (Li et al. 2009) Minimap2 (Li 2018) and BWA (Li and Durbin 2009). Followed by the retrieval of the reads maps to each bin with seqtk (https://github.com/lh3/seqtk). Those retrieved reads (Illumina and Nanopore) are then re-assembled using the Unicycler hybrid approach (Wick et al. 2017).

4.5.4 Classify

The classify step, requires either the bins or reassembled bins from the assemble step or bins submitted by the user.

The first substep is the quality assessment of the bins done by CheckM (Parks et al. 2015) using the CheckM database.

The second substep is the taxonomic classification of the bins by sourmash (Brown and Irber 2016) using the GT-DataBase (Parks et al. 2018).

The result of both is then put in a comma-separated file.

4.5.5 Annotate

The last step requires the bins or reassembled bins from the assemble step or submitted bins. It also can accept in addition to the bins, RNAseq data.

The annotation of the bins is done by eggNOG (Huerta-Cepas et al. 2017) using the eggNOG database version 5 (Huerta-Cepas et al. 2019). EggNOG is a powerful tool providing in the output KEGG (pathway, ko, module, reaction), Gene Ontology terms, EC numbers, COG, and other information.

The RNAseq data is quality controlled using fastp (Chen et al. 2018) followed by a *de novo* transcriptome assembly using Trinity (Haas et al. 2013) and Salmon (Patro et al. 2017) the transcripts are then annotated by eggNOG (Huerta-Cepas et al. 2017) using the eggNOG database version 5 (Huerta-Cepas et al. 2019).

The final substep is the execution of a parser for the annotation files that will create HTML files regrouping in an easily readable way the pathways present in the bins as well as the genes using the KEGG ID outputted in the annotation file with the KEGG PATHWAY database (see Figure n°3).

***Table 1**- Software used in MUFFIN*

Task	Software	Version	References
QC illumina	fastp	0.20.0	(Chen et al. 2018)
QC ont	Filtlong	0.2.0	https://github.com/rrwick/Filtlong
metagenomic composition of ont	sourmash	2.0.0a10	(Brown and Irber 2016)
Hybrid assembly	metaSPAdes	3.13.1	(Nurk et al. 2017)
	Unicycler	0.4.8	(Wick et al. 2017)
Long read assembly	MetaFlye	2.6	(Kolmogorov et al. 2019)
polishing	Racon	1.4.7	(Vaser et al. 2017)
	medaka	0.11.0	https://github.com/nanoporetech/medaka
	Pilon	1.23	(Walker et al. 2014)
mapping	minimap2	2.17	(Li 2018)
	BWA	0.7.17	(Li and Durbin 2009)
	SAMtools	1.9	(Li et al. 2009)
retrieve reads mapped to contig	seqtk	1.3	https://github.com/lh3/seqtk
Binning	MetaBAT2	2.14	(Kang et al. 2015)
	MaxBin2	2.2.4	(Wu et al. 2014)

Task	Software	Version	References
	CONCOCT	1.0.0	(Alneberg et al. 2014)
	MetaWRAP	1.2.1	(Uritskiy, DiRuggiero, and Taylor 2018)
QC binning	CheckM	1.0.18	(Parks et al. 2015)
Taxonomic Classification	sourmash using the GT-DataBase	Sourmash:2.0.0a10 GTDB is version R89	(Brown and Irber 2016) (Parks et al. 2018)
Annotations (bin and RNA)	eggNOG	eggNOG db v5.0 eggNOG mapper v2.0.1	(Huerta-Cepas et al. 2017) (Huerta-Cepas et al. 2019)
De novo transcript and quantification	Trinity	2.8.5	(Haas et al. 2013)
	Salmon	0.15.0	(Patro et al. 2017)

Sample overview
Summary of the pathways and orthologs on a sample level

Pathway highlight 1	...	Pathway highlight 5	Bins Compositions
Pathway name + link to the highlight 1	...	Pathway name + link to the highlight 5	Bin 1 [X orthologs identical to RNAseq, Y orthologs not found in RNAseq]; Bin 2 [X , Y]; ..... Bin n° [X , Y]

Bin overview
Summary of the pathways and orthologs for each bin

Pathway highlight 1 to 4	List of orthologs present in both RNAseq and the bin	List of orthologs only present in the bin
Pathway name + link to the highlight	Ortholog name +link Ortholog name +link ..... Ortholog name +link	Ortholog name +link Ortholog name +link .... Ortholog name +link

Pathway figure
contains a highlight of the orthologs present.

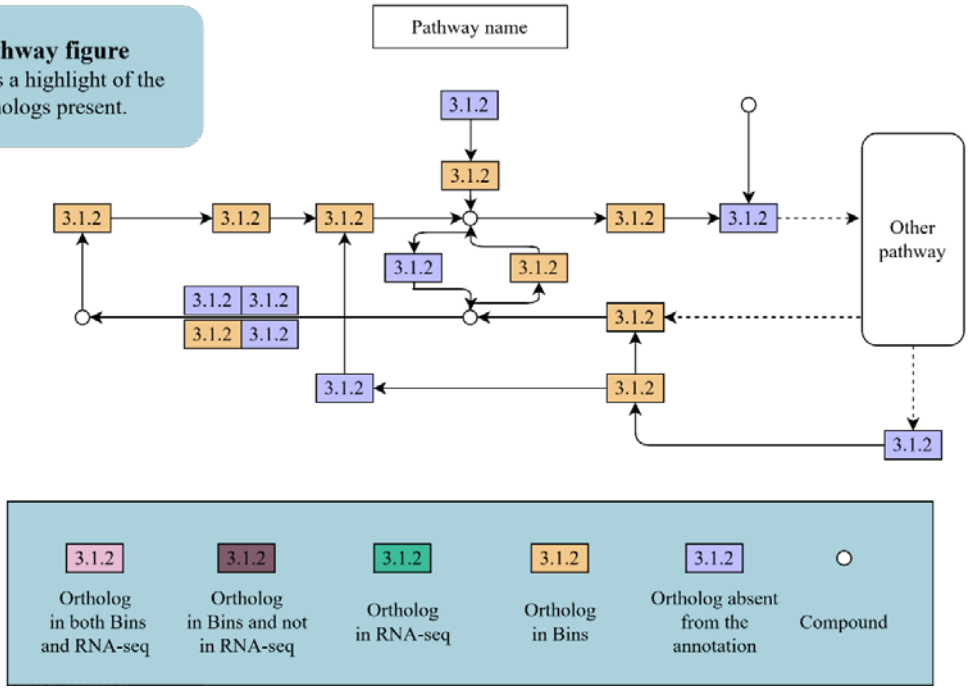


Figure 3 *- Example of the parser output*

4.5.6 The databases

The CheckM database is a collection of precalculated data used by CheckM to assign taxonomy and check the completeness and contamination of a bin. It is composed of markers genes grouped into lineage-specific collocated markers sets. Those markers sets are the critical element of CheckM to assess the completeness and contamination of a bin(Parks et al. 2015). CheckM database is limited to known markers

from known lineages and should be used as an indicator of the quality of the bins. If the work is on oriented lesser-known or unknown bacteria, CheckM results might not reflect the actual quality of the bin but only a grade of similarities between the bin and a potentially close lineage.

GT database (GTDB) or genome taxonomy database is a standardised microbial taxonomy database based on phylogeny. This database constructs its phylogeny using genomes from RefSeq (O'Leary et al. 2016) and GenBank (Clark et al. 2016) but increasingly also using draft genomes of metagenomics and single-cell uncultured organisms trying to improve the genomic representation of the microbial world(source: https://gtdb.ecogenomic.org/about).

eggNOG 5.0 database is a database of ortholog relationships, functional annotation, and gene evolutionary histories(Huerta-Cepas et al. 2019). Used by the eggNOG annotation tool (eggNOG mapper V2)(Huerta-Cepas et al. 2017), it forms both for the eggNOG service. EggNOG is a system for automated construction and annotation of orthologous groups of genes, using phylogenetic resolution, automatically updated, and contains a hierarchy of orthologous groups to balance phylogenetic coverage and resolution(Jensen et al. 2008).

The KEGG PATHWAY database "is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies." (source: https://www.kegg.jp/). It contains the different pathways as well as modules of those pathways, reaction, enzyme, gene, genomes.

5 Results

The samples analysed are labelled “02-SW” biogas sample for the DNA and “BM03” sample for the RNA. 02-SW is a thermophilic (52°C) biogas reactor using slaughter and food waste.

5.1 Quality Control

5.1.1 Nanopore

The quality of the nanopore sequencing was good with over 3.6 million reads produced, 20 gigabases called, and 81.4% of reads passing the QC filter. The mean read length is 5.786bp, an N50 of 8.604, and the mean read quality (QV) is 10.3.
The quality report was produced by Nanoplot (De Coster et al. 2018).

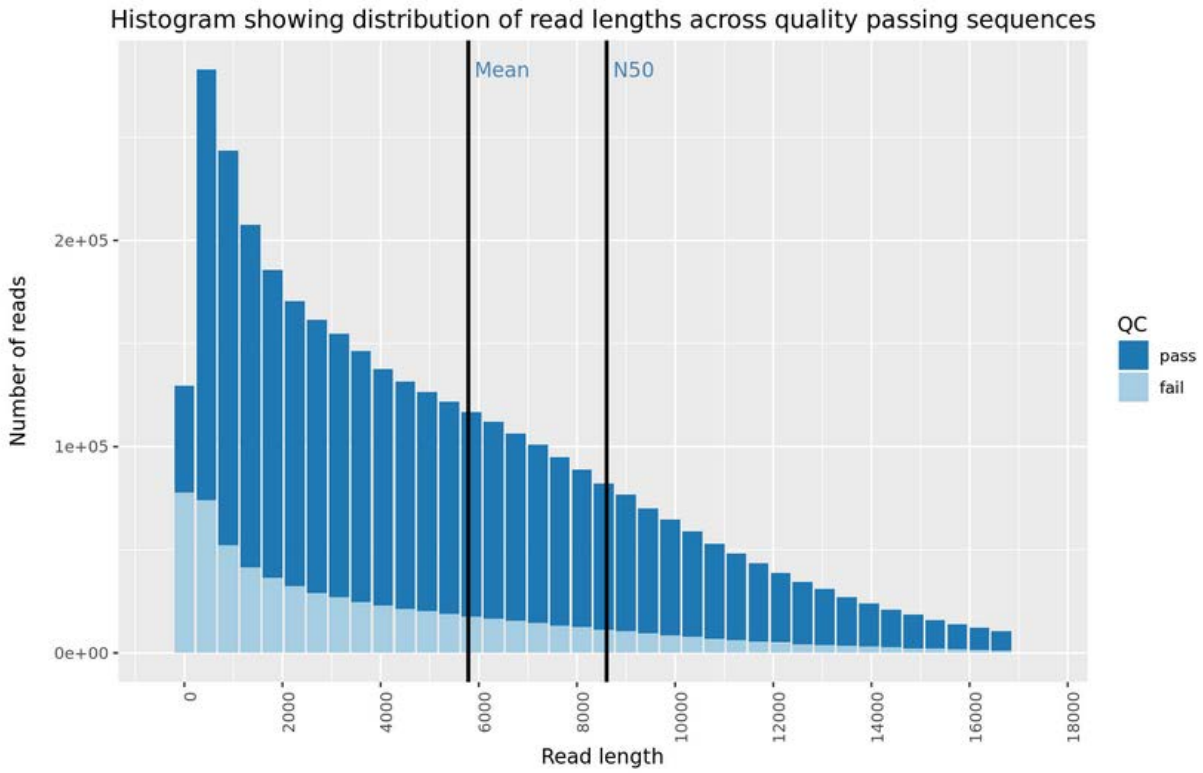


Figure 4 - The quality control output of nanopore sequencing after guppy basecalling.

The quality control of the pipeline was a strict removal of reads under 2000bp length.

5.1.2 Illumina

The quality of the raw Illumina reads was for R1 and R2, 18 946 658 reads with 46%GC, and a 151bp read length, the mean quality per read was 36 (Phred score).

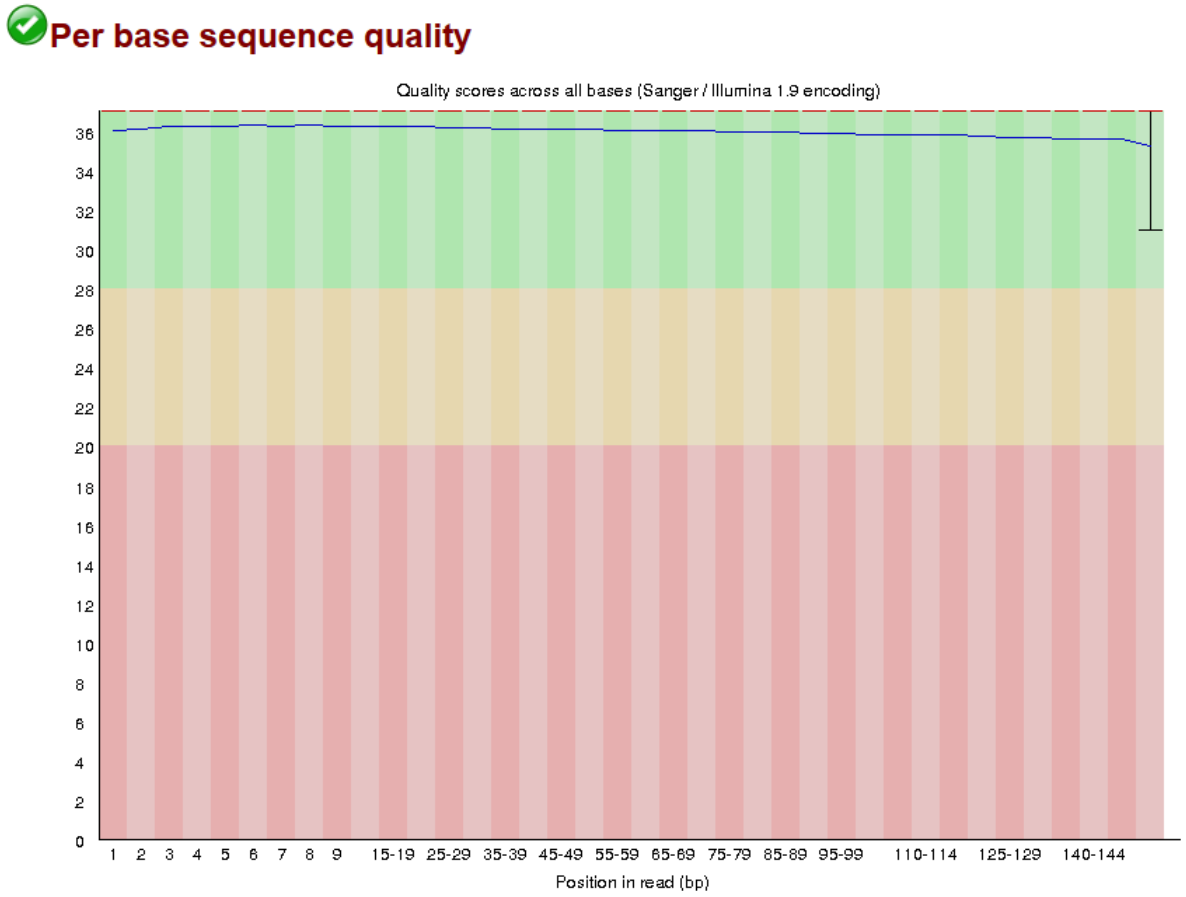


Figure 5 - The per sequence quality of the Illumina R1 read before any quality improvement with fastp.

The changes after fastp are the discard of the reads under 20 Phred score quality and the removal of an overrepresented sequence, a G nucleotide repetition present over 21 000 times.

The quality results were computed by FastQC (Andrews et al. 2012).

5.1.3 RNA

The quality of the raw Illumina reads was for R1 and R2, 4973956 reads with 48%GC, and a 76bp read length, the mean quality per read was 37 (Phred score).

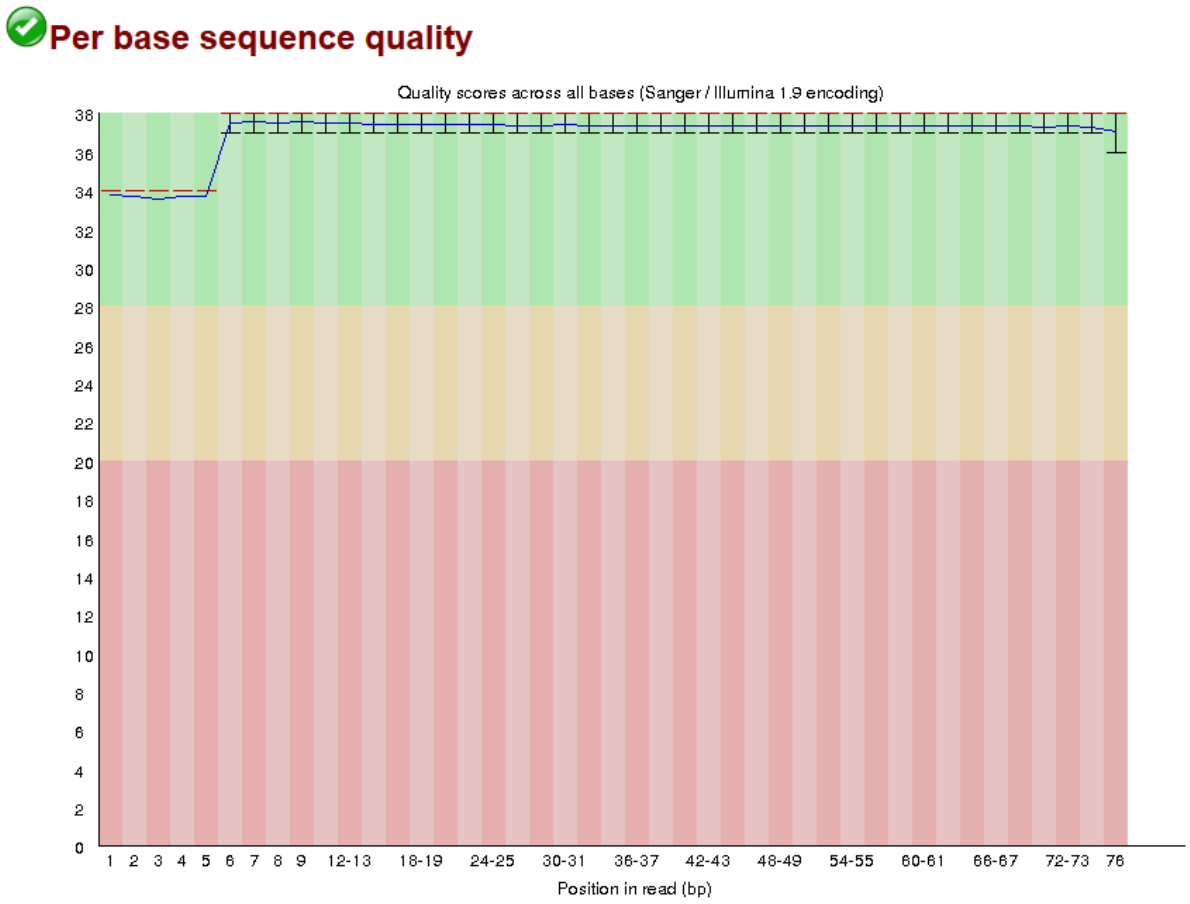


Figure 6 - The per sequence quality of the RNAseq R1 read before any quality improvement with fastp.

The changes after fastp are the discard of the reads under 19 Phred score quality and a slight diminution of the overrepresented sequences.

The quality results were computed by FastQC(Andrews et al. 2012).

5.2 Assembly

The assembly longest read is 266.071bp long, and there are 4.149 contigs longer than 10.000bp (26 306 contigs longer than 1000bp).

5.3 Binning

Two significant elements are used to assess the quality by CheckM: 1) The completeness. This is the % of gene markers sets of an organism in the CheckM database present in the bin. 2) The contamination. This is the % of gene markers sets of foreign organisms from the one attributed by CheckM in the bin.

The binning substep produced a total of 35 bins with completeness estimated by CheckM of 71.16% at lowest (bin 20) and 99.60% at highest (bin 16) and with contamination of 6.78% at the highest (bin 18) and lowest 0% contamination. The mean percentage of completeness is 90.99%, and for the contamination, it is 1.38%.

MaxBin2 produced 51 bins; MetaBAT2 produced 60 bins, and CONCOCT produced 138 bins. The binning and binning refinement only keeps the contigs they deem appropriate to the bins. MUFFIN can take the unbinned data and retrieve the reads from the sample that are not part of the bins. This can be very convenient as in the analysis of metagenomics, the genomic data of some organisms with a low population can be "hidden" in the analysis by organisms that represent a high proportion of the communities. MUFFIN allows the preservation of the unbinned data to rerun the analysis once the data of the highly present organism has been analysed. The second run of MUFFIN would then be more specific to lowly abundant organisms.

5.4 CheckM vs Sourmash (GTDB) classification

CheckM database is limited compared to the GTDB, but the comparison of the 2 showed that sourmash classify on a more specific level while also showing some disagreement between CheckM hit and sourmash (using the GTDB). GTDB also distinguishes the Firmicutes phylum in different phylum (e.g., Firmicutes_A, Firmicutes_b, Firmicutes_G). The complete table including CheckM quality control and Sourmash with the complete taxonomic resolution can be found in Appendix n°1

***Table 2**- Bins with their respective lineage from CheckM and sourmash (GTDB). Sourmash was limited to the class level, see Appendix n°1 for the complete taxonomic resolution.*

Bin ID	CheckM Marker lineage	Sourmash Status	Sourmash phylum	Sourmash Class
bin.01	c__Clostridia	found	p__Firmicutes_B	c__Syntrophomonadia
bin.02	k__Bacteria	found	p__Firmicutes	c__Bacilli
bin.03	p__Firmicutes	found	p__Firmicutes_G	c__UBA4882
bin.04	k__Bacteria	found	p__Thermotogota	c__Thermotogae

bin.05	o__Clostridiales	found	p__Firmicutes_A	c__Clostridia
bin.06	p__Firmicutes	found	p__Firmicutes_B	c__Syntropho-monadia
bin.07	c__Clostridia	found	p__Firmicutes_A	c__Clostridia
bin.08	o__Clostridiales	nomatch		
bin.09	p__Firmicutes	found	p__Firmicutes_G	c__SHA-98
bin.10	p__Euryarchae-ota	found	p__Halobacterota	c__Methanomi-crobia
bin.11	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.12	k__Bacteria	found	p__Thermotogota	c__Thermotogae
bin.13	k__Bacteria	disagree	p__Bacteroidota	c__Bacteroidia
bin.14	o__Clostridiales	found	p__Firmicutes_A	c__Clostridia
bin.15	p__Firmicutes	found	p__DTU030	c__DTU030
bin.16	p__Euryarchae-ota	found	p__Thermoplasma-tota	c__Thermoplas-mata
bin.17	p__Firmicutes	nomatch		
bin.18	k__Bacteria	found	p__Caldatribacteri-ota	c__Caldatribac-teriia
bin.19	p__Bacteroide-tes	found	p__Bacteroidota	c__Bacteroidia
bin.20	k__Bacteria	nomatch		
bin.21	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.22	p__Firmicutes	nomatch		
bin.23	k__Bacteria	found	p__Caldatribacteri-ota	c__Caldatribac-teriia
bin.24	k__Bacteria	found	p__Firmicutes	c__Bacilli
bin.25	p__Firmicutes	found	p__Firmicutes_G	c__SHA-98
bin.26	p__Firmicutes	disagree	p__Firmicutes_G	
bin.27	p__Firmicutes	found	p__Firmicutes_E	c__DTU015
bin.28	p__Firmicutes	nomatch		
bin.29	p__Firmicutes	found	p__Firmicutes_A	c__Thermove-nabulia
bin.30	p__Firmicutes	found	p__Firmicutes_G	c__Limnochordia
bin.31	p__Firmicutes	nomatch		
bin.32	p__Firmicutes	found	p__Firmicutes_F	c__Halanaero-biia
bin.33	p__Firmicutes	found	p__Firmicutes_D	c__Dethiobacte-ria
bin.34	p__Firmicutes	found	p__Firmicutes_G	c__DTU065
bin.35	k__Bacteria	nomatch		

5.5 RNA *de novo* Transcripts

The de novo transcript file produced by Trinity contained 48 283 transcripts and 43 426 “genes” with 44.26%GC. Based on all transcript contigs, the contig N50 is 899, the average contig length was 615.85, and the total assembled bases were 29 735 097 bases. Based on only the longest isoform per “gene,” the N50 is 640, the average contig length is 529.44, and the total of assembled bases is 22 991 484 bases.

Those results are from the TrinityStats.pl scripts of Trinity (Haas et al. 2013). The quantification of all the transcripts is normalised using the TPM methods and done by Salmon (Patro et al. 2017). We can deduce, from the total of assembled bases, that of the four million reads with a length of 75bp, most of them were used in the transcript assembly.

5.6 EggNOG annotation parsed.

The annotation of the RNAseq was done on all the transcripts produced, and there was no threshold of minimum quantification required. The annotation of the bins and RNAseq data by eggNOG gives as an output both ID of the KEGG pathway and ID of the KEGG orthology in the result files the said orthology ID (ko number) are called “genes” to simplify the explanations, each ko number represent one or multiple genes that are registered as different entries for different organisms. In the intent of making them more transparent and more straightforward for the HTML, the use of the ko number was chosen over the use of the gene name or entry of an arbitrary organism.

In the annotation of the bins, a total of 249 different pathways were found. In the annotation of the RNAseq data, 305 pathways were found. This difference of pathways found could be due to the use of only a majority of the initial genomic data as during the binning, some portion of the data was not associated with one of the 35 bins. There is also a difference in sequence depth. The depleted RNA was sequenced with more depth than the DNA.

Moreover, a part of the data was also not annotated. This could be due to a lack of information for those DNA sequences in the database or the fact that those sequences are incomplete as well as in regards to the annotation, a non-utility of those sequences. Only further research about those sequences could answer this. Sequences that did not contain annotation information about pathway while still being annotated for another database (GO term, BiGG Reaction, BRITE, CAZy).

The number of pathways being substantial, only a few will be shown here as an example(chapters 5.6.1 to 5.6.3). The chosen pathways are methane metabolism(ko00680), carbon metabolism(ko01200) that contains Acetogen module

(M00618), and glycolysis (ko00010). Those three pathways are involved in anaerobic degradation process to produce methane.

The point 5.6.1 to 5.6.3 are a rough representation of the presence of "gene" in the pathways. Unfortunately, the system put in place to give visual representation is limited to a total of 119 highlighted "genes" at a time in the pathway. This means that in the case of a number of genes superior to 119, the list must be reduced to 119 or lower to create the figure, and you can repeat the creation of the figure as much as needed with other subsets of the list. Here only 1 figure is shown so if the gene is highlighted it show the presence but if it is not highlighted, the creation of the other figure is required to assess graphically the absence of the said "gene."

5.6.1 The glycolysis

The glycolysis pathway is an example of a pathway highlighted by the parser.
In the sample tested, the glycolysis pathway contains a total of 136 "genes" in the RNAseq data. We can see from figure n°7 that the RNAseq data express most of the pathway genes. The number of "genes" present in both the bins and the RNAseq (denominated as "expressed") as well as the number of "genes" present in only in the bins but not in the RNAseq data (denominated as "non-expressed") are in the APPENDIX n° 2. In the APPENDIX n°2 is also present the figure of the "expressed genes" and the figure of the "non-expressed genes."

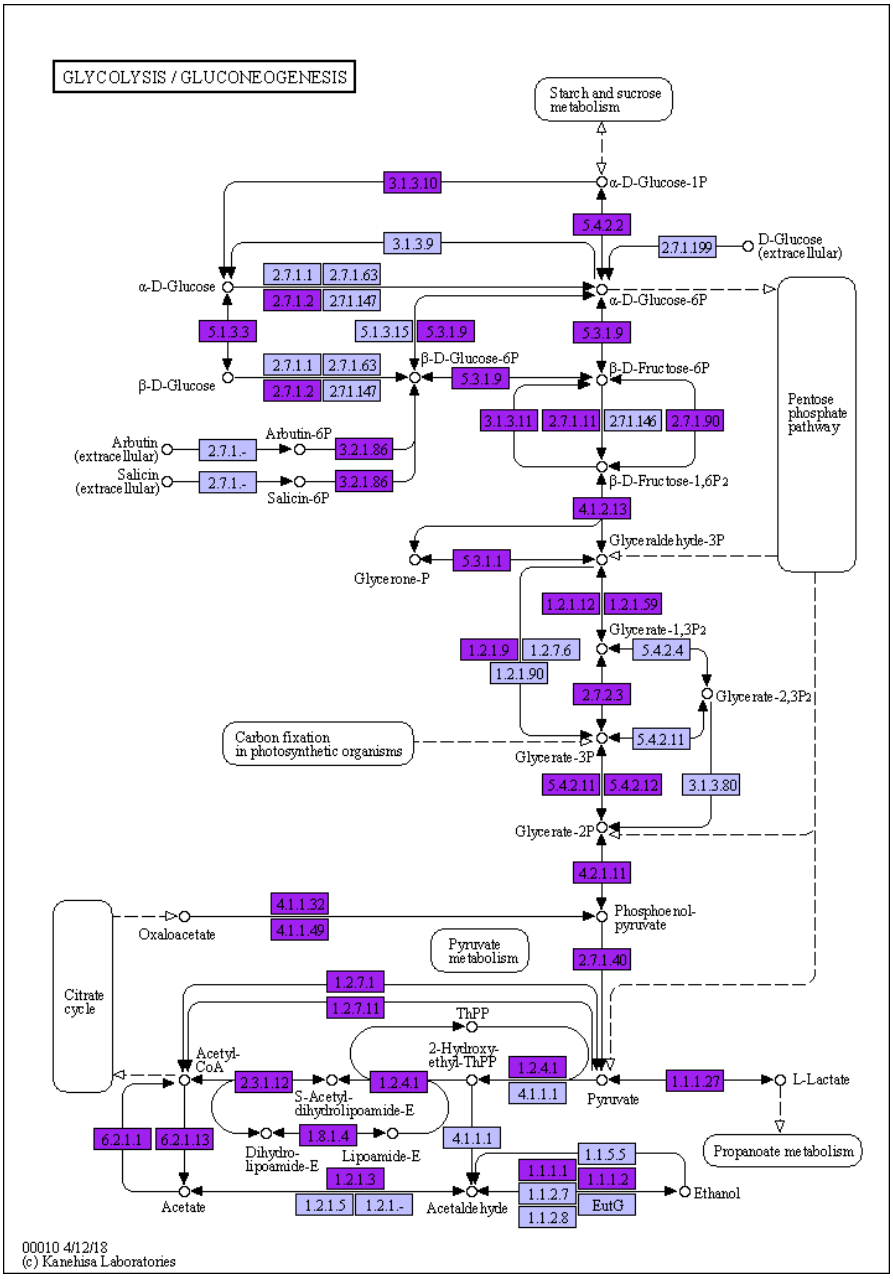


Figure 7- The glycolysis pathway with 119 out of 136 “genes” highlighted in purple

5.6.2 The methane metabolism

The methane metabolism pathway is an example of a pathway highlighted by the parser. In the sample tested, the methane metabolism contains a total of 206 “genes” in the RNAseq data. We can see from figure n°8 that a majority of the gene present in the bins are present in the RNAseq data (green). The number of “genes” present

in both the bins and the RNAseq data (denominated as “expressed”) as well as the number of “genes” present in only in the bins but not in the RNAseq data (denominated as “non-expressed”) are in the APPENDIX n°2. In the APPENDIX n°2 is also present the figure of the RNA “genes” and the figure of the bins “genes” without distinction by the presence of it in RNAseq data.

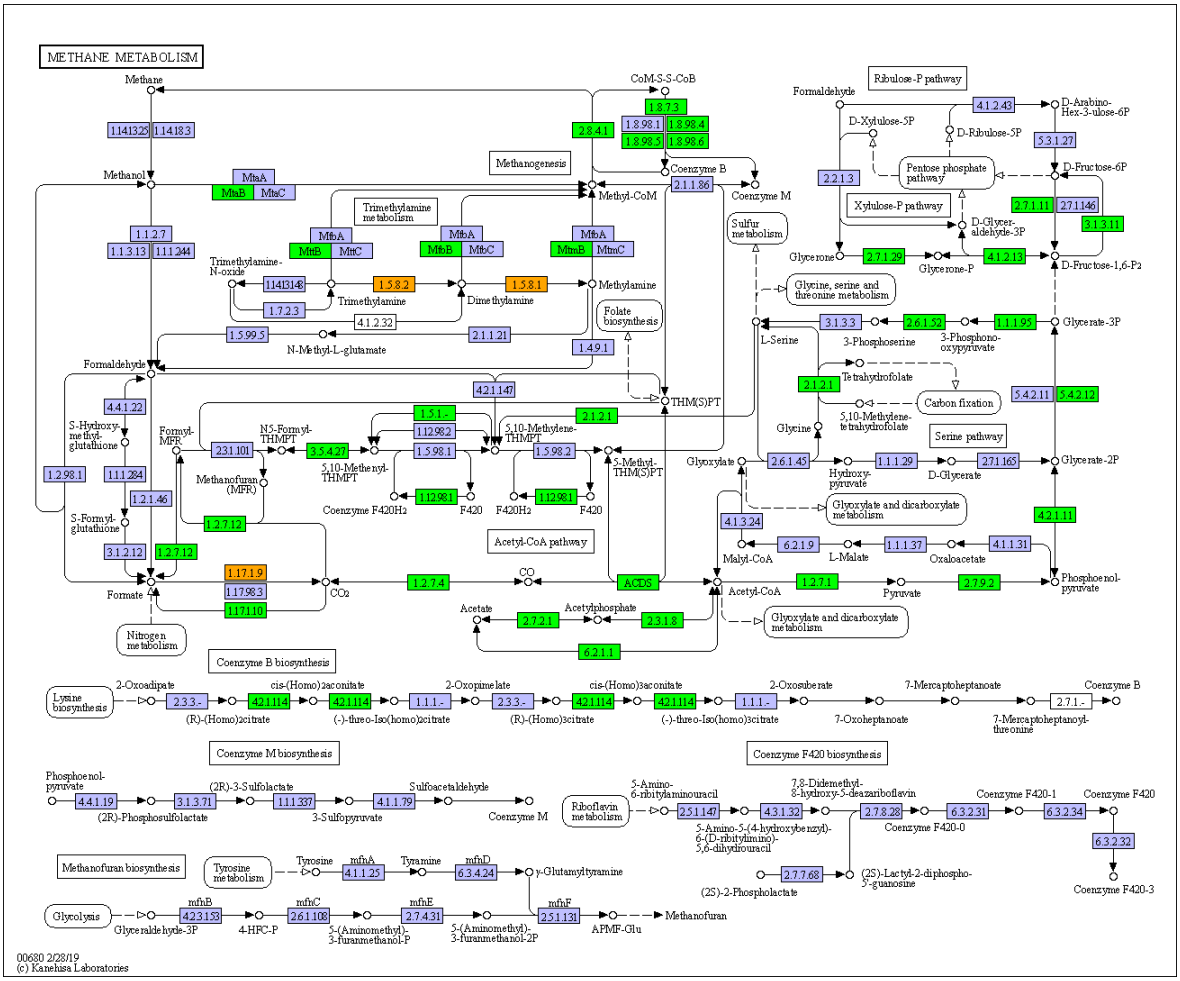


Figure 8- The methane metabolism pathway with the “expressed genes” highlighted in green and the “non-expressed genes” highlighted in orange.

5.6.3 The carbon metabolism

The carbon metabolism pathway is an example of a pathway highlighted by the parser. In the sample tested, the carbon metabolism contains a total of 383 “genes” in

the RNAseq data. We can see from figure n°9 that most of the pathway is expressed by the RNAseq data while being present in the bins MAGs. The number of “genes” present in both the bins and the RNAseq data (denominated as “expressed”) as well as the number of “genes” present only in the bins but not in the RNAseq data (denominated as “non-expressed”) are in the APPENDIX n°2. In the APPENDIX n° 2 is also present the figure of the “non-expressed genes” and the figure of the RNA “genes.”

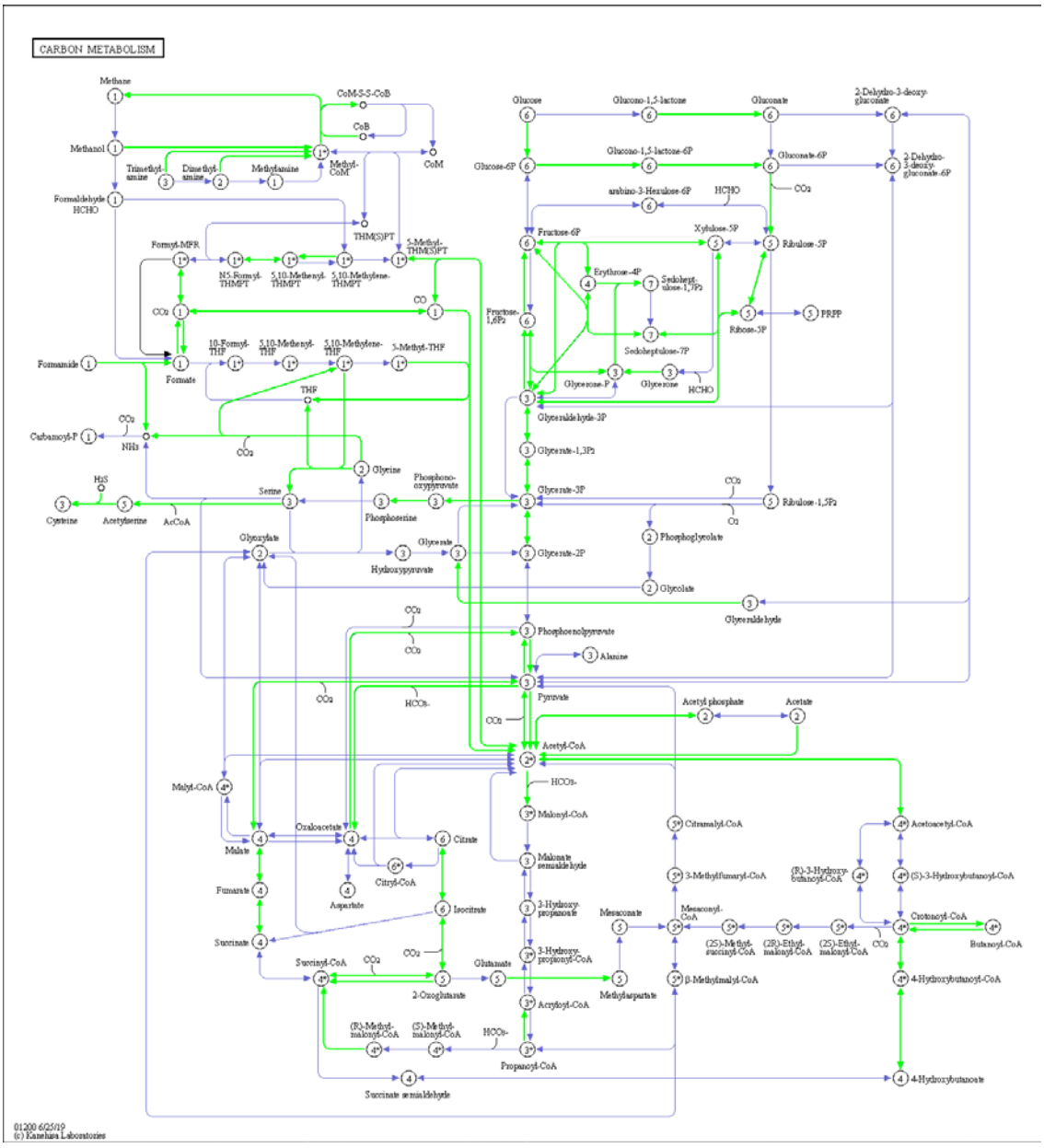


Figure 9- The carbon metabolism pathway with the “expressed genes” highlighted in green.

6 Discussion

6.1 Using Hybrid assembly

The long-read assembly has the advantages of avoiding the repeats and the gaps that can be produced in the assembly of the short reads. While short-read assembly is more accurate on the base level but has a higher risk of misassemblies through gaps and repeats.

6.2 Using three binning methods and a binning refiner

The different binning algorithms all have errors and weaknesses. That is why the use of binning refinement such as in MetaWRAP (Uritskiy, DiRuggiero, and Taylor 2018), DAS_Tool (Sieber et al. 2018), and Binning refiner (Song and Thomas 2017) is developed. The bin refinement uses the bins obtained from different methods to analyse and characterise their accuracy and then use the best elements of each method to output the best bins. This is showed by an improvement of the completeness and a diminution of the contamination assessed by CheckM.

6.3 Use GTDB with sourmash for classification

Two factors motivated the use of sourmash with the genome taxonomy database (GTDB). Sourmash is proven to be an efficient classification software using MinHash sketches of genomic data. The main advantages of this software are the high processing speed of sourmash as it is based on hashes produced from public databases and not the said databases. This also helps in the size of the database and the accessibility; a laptop can run sourmash on a database with no issue whatsoever and in an acceptable time (Breitwieser, Lu, and Salzberg 2019).

The GTDB was converted to a sourmash database with sourmash. The use of the GTDB is essential in the classification of MAGs from biogas reactor as GTDB include good quality draft genomes from such samples besides the RefSeq and Genbank databases.

6.4 Why use eggNOG to annotate

The eggNOG-Mapper is a tool that revolves around the adaptation of the annotation requirement (speed, accuracy). It can use both a HMMs database with HMMer3 (Eddy 2011) to map the query sequence then creates orthologous groups, or it can use a protein database with diamond to obtain the seed eggNOG orthologs that are then analysed the same way (Buchfink, Xie, and Huson 2015).

The use of the diamond method combined with the eggNOG 5.0 database leads to a fast and accurate annotation. Where Blast and InterProScan show a slightly worse result for a higher computational time (Huerta-Cepas et al. 2017).

6.5 Gene expression

The gene expression of the RNAseq data has a limited reach. Indeed, the output of the quantification only gives a TPM normalized quantification of the transcripts in the sample. Thus, the quantification can solely be useful for the understanding of expression level in the sample at a specific time. Due to the RNA sequencing of the microbial community in the sample and not independent organisms, only interpretation on the sample level can be made.

6.6 Graphical display

One of the significant issues with the actual display of the pathway with “genes” presence is the limitation of 119 “genes” entry at a time for the graphical display. It was chosen to redirect the pathway directly to the online KEGG DB as the number of figures to download and store to display with an offline mode would be too excessive.

The graphical display of the pathway is an additional feature that is helpful for the visualisation and comprehension of the “genes” involvement in the pathways activities. Nevertheless, when it is impossible to access, the use of manual search or reduction of the list of genes to display can offer more limited information.

6.7 MUFFIN limitations

Creating an automated pipeline also leads to some limitations. The use of the three steps of the pipeline can increase the computational charge when the aim is on a smaller number of bins or a particular organism. That is why each step can also be run individually, making a stop after the first and second steps to allow the manual narrowing of the data to analyse. For example, a sample gives 42 bins. You can

decide either to use all of them in the next steps, or reduce the number of bins, or decide after the CheckM quality check and the taxonomic classification to keep only the bins over XX% of completeness and from the YY lineage.

Another limitation is the lack of liberty for the user to tweak each software, and this was a choice made to have an ergonomic and straightforward pipeline. People who will want to configure everything manually will tend to run each software individually with the desired parameter.

7 Conclusion and further perspectives

This project results in the creation of a metagenomics analysis pipeline supported by de novo transcriptomics (MUFFIN). That is reproducible, automated, and simple of utilisation. This pipeline can use hybrid assembly methods to increase the completeness but also the base level quality of the MAGs produced. It also produces taxonomic classification and bin qualities, bin and transcript annotation, transcript expression on the number of reads for each transcript (TPM normalised), and finally simple HTML summary files to show the pathway present in the bins and the involvement of the said bin in the pathways.

The data used to test this pipeline showed that the assembly and the binning steps produced a fair number of bins (35 bins) with overall good quality, over 70% completeness and less than 2% contamination for all bins. The taxonomic classification showed similar hits as in other studies while also opening to new potential discoveries. The annotation is also a source of various information that can be utilized for further and more in-depth researches on the microbial population and interaction.

Further research on bins 6, 21 and 23 could be of great interest as the result of the pipeline show a good level of completeness with low contamination (APPENDIX n°1). The classification indicates them as “Thermacetogeniaceae,” “unclassified clostridium,” and “Caldatribacteriaceae,” respectively (Appendix n°1), and the annotation shows their substantial involvement in the methane and carbon metabolism pathways (APPENDIX n°2).

MUFFIN could benefit in the future of different improvements such as a file that summarises all the taxonomic classification effectuate by MetaWRAP, CheckM, eggNOG, and Sourmash; creating a list of all the found ko IDs (from KEGG DB) that are not involved in a pathway according to the eggNOG annotation.

It could also benefit of the addition some more statistics and information in the HTML such as the total of ko IDs found versus the ko IDs found with a KEGG pathway; the percentage of ko present in the pathway versus what is found in the RNA, bins, individual bin; a distribution of the pathways (e.g., the most abundant pathways in the bins).

To place the classified bin in a graphical phylogenetic tree, create the option to output the graphical pathway with the complete set of genes.

Another improvement would be to create an advance user and wizard user configuration file that would allow the user to tweak all the different parameters of all the software as desired.

MUFFIN could also benefit from the addition of new analysis software such as differential expression analysis, short reads assembly methods. Another improvement of the MUFFIN pipeline would be to diversify the sources for the reads input. The use of Pacific Biosciences sequenced reads for example.

Acknowledgments

I would first like to thank my supervisors Erik Bongcam-Rudlof, Bettina Müller and Christian Brandt for their support through this work, the different suggestion to improve the pipeline, the exciting discussion, and the opportunity to integrate into their groups at the Swedish University of Agricultural Sciences.

I would like to thanks Hadrien Gourlé and Mortiz Buck for their advice and suggestions.

Also, I would like to thanks David Coornaert, Françoise Besanger and Aline Leonet for being the sparks of my motivation toward scientific research and bioinformatics. I also thank them for the chance they gave me to discover Uppsala and the SLU during my bachelor internship.

Lastly, I would like to thanks my family for their support through this journey and especially my dad for making me curious about everything and giving me the stubbornness of continuing my projects no matter what.

References

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, et al. 2012. 'Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome'. *PLOS Computational Biology* 8 (6): e1002358. https://doi.org/10.1371/journal.pcbi.1002358.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. 'Binning Metagenomic Contigs by Coverage and Composition'. *Nature Methods* 11 (11): 1144–46. https://doi.org/10.1038/nmeth.3103.

Andrews, Simon, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. 2012. *FastQC*. Babraham, UK.

Angelidaki, Irini, Dimitar Karakashev, Damien J. Batstone, Caroline M. Plugge, and Alfons J. M. Stams. 2011. 'Chapter Sixteen - Biomethanation and Its Potential'. In *Methods in Enzymology*, edited by Amy C. Rosenzweig and Stephen W. Ragsdale, 494:327–51. Methods in Methane Metabolism, Part A. Academic Press. https://doi.org/10.1016/B978-0-12-385112-3.00016-0.

Arafat, Hassan A., Kenan Jijakli, and Amimul Ahsan. 2015. 'Environmental Performance and Energy Recovery Potential of Five Processes for Municipal Solid Waste Treatment'. *Journal of Cleaner Production*, Decision-support models and tools for helping to make real progress to more sustainable societies, 105 (October): 233–40. https://doi.org/10.1016/j.jclepro.2013.11.071.

Atelge, M. R., David Krisa, Gopalakrishnan Kumar, Cigdem Eskicioglu, Dinh Duc Nguyen, Soon Woong Chang, A. E. Atabani, Alaa H. Al-Muhtaseb, and S. Unalan. 2018. 'Biogas Production from Organic Waste: Recent Progress and Perspectives'. *Waste and Biomass Valorization*, December. https://doi.org/10.1007/s12649-018-00546-0.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing'. *Journal of Computational Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Bastide, Amandine, and Alexandre David. 2018. 'Interaction of RRNA with MRNA and TRNA in Translating Mammalian Ribosome: Functional Implications in Health and Disease'. *Biomolecules* 8 (4). https://doi.org/10.3390/biom8040100.

Botello Suárez, Wilmar Alirio, Juliana da Silva Vantini, Rose Maria Duda, Poliana Fernanda Giachetto, Leandro Carrijo Cintra, Maria Inês Tiraboschi Ferro, and Roberto Alves de Oliveira. 2018. 'Predominance of Syntrophic Bacteria, Methanosaeta and Methanoculleus in a Two-Stage up-Flow Anaerobic Sludge Blanket Reactor Treating Coffee Processing Wastewater at High Organic Loading Rate'. *Bioresource Technology* 268 (November): 158–68. https://doi.org/10.1016/j.biortech.2018.06.091.

Breitwieser, Florian P, Jennifer Lu, and Steven L Salzberg. 2019. 'A Review of Methods and Databases for Metagenomic Classification and Assembly'. *Briefings in Bioinformatics* 20 (4): 1125–36. https://doi.org/10.1093/bib/bbx120.

Brown, C., and Luiz Irber. 2016. 'Sourmash: A Library for MinHash Sketching of DNA'. Journal of Open Source Software. 14 September 2016. https://doi.org/10.21105/joss.00027.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. ‘Fast and Sensitive Protein Alignment Using DIAMOND’. *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. ‘Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor’. *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

Clark, Karen, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2016. ‘GenBank’. *Nucleic Acids Research* 44 (D1): D67–72. https://doi.org/10.1093/nar/gkv1276.

De Coster, Wouter, Svenn D’Hert, Darrin T. Schultz, Marc Cruts, and Christine Van Broeckhoven. 2018. ‘NanoPack: Visualizing and Processing Long-Read Sequencing Data’. *Bioinformatics* 34 (15): 2666–69. https://doi.org/10.1093/bioinformatics/bty149.

Eddy, Sean R. 2011. ‘Accelerated Profile HMM Searches’. *PLoS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. ‘Anvi’o: An Advanced Analysis and Visualization Platform for ‘omics Data’. *PeerJ* 3 (October): e1319. https://doi.org/10.7717/peerj.1319.

Eriksson, Mattias, Ingrid Strid, and Per-Anders Hansson. 2015. ‘Carbon Footprint of Food Waste Management Options in the Waste Hierarchy – a Swedish Case Study’. *Journal of Cleaner Production* 93 (April): 115–25. https://doi.org/10.1016/j.jclepro.2015.01.026.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Johannes Alneberg, Harshil Patel, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2019. ‘Nf-Core: Community Curated Bioinformatics Pipelines’. *BioRxiv*, April, 610741. https://doi.org/10.1101/610741.

Ghosh, Arpita, Aditya Mehta, and Asif M. Khan. 2019. ‘Metagenomic Analysis and Its Applications’. In *Encyclopedia of Bioinformatics and Computational Biology*, edited by Shoba Ranganathan, Michael Gribskov, Kenta Nakai, and Christian Schönbach, 184–93. Oxford: Academic Press. https://doi.org/10.1016/B978-0-12-809633-8.20178-7.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. ‘De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis’. *Nature Protocols* 8 (8): 1494–1512. https://doi.org/10.1038/nprot.2013.084.

Huerta-Cepas, Jaime, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. 2017. ‘Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper’. *Molecular Biology and Evolution* 34 (8): 2115–22. https://doi.org/10.1093/molbev/msx148.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. ‘EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses’. *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Jensen, Lars Juhl, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. 2008. ‘EggNOG: Automated Construction and Annotation of Orthologous Groups of Genes’. *Nucleic Acids Research* 36 (Database issue): D250–54. https://doi.org/10.1093/nar/gkm796.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. ‘MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities’. *PeerJ* 3: e1165. https://doi.org/10.7717/peerj.1165.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. ‘Assembly of Long, Error-Prone Reads Using Repeat Graphs’. *Nature Biotechnology* 37 (5): 540–46. https://doi.org/10.1038/s41587-019-0072-8.

Lee, Joonyeob, Eunji Kim, Gyuseong Han, Jovale Vincent Tongco, Seung Gu Shin, and Seokhwan Hwang. 2018. ‘Microbial Communities Underpinning Mesophilic Anaerobic Digesters Treating

Food Wastewater or Sewage Sludge: A Full-Scale Study'. *Bioresource Technology* 259 (March). https://doi.org/10.1016/j.biortech.2018.03.052.

Leipzig, Jeremy. 2017. 'A Review of Bioinformatic Pipeline Frameworks'. *Briefings in Bioinformatics* 18 (3): 530–36. https://doi.org/10.1093/bib/bbw020.

Li, Heng. 2018. 'Minimap2: Pairwise Alignment for Nucleotide Sequences'. *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, Heng, and Richard Durbin. 2009. 'Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform'. *Bioinformatics* 25 (14): 1754–60. https://doi.org/10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Liu, Yuchen, and William B. Whitman. 2008. 'Metabolic, Phylogenetic, and Ecological Diversity of the Methanogenic Archaea'. *Annals of the New York Academy of Sciences* 1125 (1): 171–89. https://doi.org/10.1196/annals.1419.019.

Ma, Zhanshan (Sam), Lianwei Li, Chengxi Ye, Minsheng Peng, and Ya-Ping Zhang. 2019. 'Hybrid Assembly of Ultra-Long Nanopore Reads Augmented with 10x-Genomics Contigs: Demonstrated with a Human Genome'. *Genomics* 111 (6): 1896–1901. https://doi.org/10.1016/j.ygeno.2018.12.013.

Manzoor, Shahid, Erik Bongcam-Rudloff, Anna Schnürer, and Bettina Müller. 2016. 'Genome-Guided Analysis and Whole Transcriptome Profiling of the Mesophilic Syntrophic Acetate Oxidising Bacterium Syntrophaceticus Schinkii'. *PLOS ONE* 11 (11): e0166520. https://doi.org/10.1371/journal.pone.0166520.

Mesle, Margaux, Gilles Dromart, and Phil Oger. 2013. 'Microbial Methanogenesis in Subsurface Oil and Coal'. *Research in Microbiology* 164 (July). https://doi.org/10.1016/j.resmic.2013.07.004.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, EM Glass, M. Kubal, T. Paczian, et al. 2008. 'The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes'. *BMC Bioinformatics* 9 (1): 386. https://doi.org/10.1186/1471-2105-9-386.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. 'MetaSPAdes: A New Versatile Metagenomic Assembler'. *Genome Research* 27 (5): 824–34. https://doi.org/10.1101/gr.213959.116.

Nurul, Ashyikin Noor Ahmad, Danish-Daniel Muhammad, Victor Tosin Okomoda, and Ariffin Asma Bt. Nur. 2019. '16S RRNA-Based Metagenomic Analysis of Microbial Communities Associated with Wild Labroides Dimidiatus from Karah Island, Terengganu, Malaysia'. *Biotechnology Reports* 21 (January). https://doi.org/10.1016/j.btre.2019.e00303.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. 'Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation'. *Nucleic Acids Research* 44 (D1): D733-745. https://doi.org/10.1093/nar/gkv1189.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. 'A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life'. *Nature Biotechnology* 36 (10): 996–1004. https://doi.org/10.1038/nbt.4229.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. 'CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes'. *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. 'Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life'. *Nature Microbiology* 2 (11): 1533–42. https://doi.org/10.1038/s41564-017-0012-7.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. 'Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression'. *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Pelletier, E., A. Kreimeyer, S. Bocs, Z. Rouy, G. Gyapay, R. Chouari, D. Riviere, et al. 2008. '"Candidatus Cloacamonas Acidaminovorans": Genome Sequence Reconstruction Provides a First Glimpse of a New Bacterial Division'. *Journal of Bacteriology* 190 (7): 2572–79. https://doi.org/10.1128/JB.01248-07.

Raboni, Massimo, and Giordano Urbini. 2014. 'Production and Use of Biogas in Europe: A Survey of Current Status and Perspectives'. *Revista Ambiente e Agua* 9 (June): 191–202. https://doi.org/10.4136/ambi-agua.1324.

Ragsdale, Stephen W., and Elizabeth Pierce. 2008. 'Acetogenesis and the Wood-Ljungdahl Pathway of CO2 Fixation'. *Biochimica et Biophysica Acta* 1784 (12): 1873–98. https://doi.org/10.1016/j.bbapap.2008.08.012.

Sessitsch, Angela, Stephen Gyamfi, Nancy Stralis-Pavese, Alexandra Weilharter, and Ulrike Pfeifer. 2002. 'RNA Isolation from Soil for Bacterial Community and Functional Analysis: Evaluation of Different Extraction and Soil Conservation Protocols'. *Journal of Microbiological Methods* 51 (2): 171–79. https://doi.org/10.1016/S0167-7012(02)00065-9.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. 'Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy'. *Nature Microbiology* 3 (7): 836–43. https://doi.org/10.1038/s41564-018-0171-1.

Solli, Linn, Othilde Elise Håvelsrud, Svein Jarle Horn, and Anne Gunn Rike. 2014. 'A Metagenomic Study of the Microbial Communities in Four Parallel Biogas Reactors'. *Biotechnology for Biofuels* 7 (1): 146. https://doi.org/10.1186/s13068-014-0146-2.

Song, Wei-Zhi, and Torsten Thomas. 2017. 'Binning_refiner: Improving Genome Bins through the Combination of Different Binning Programs'. *Bioinformatics* 33 (12): 1873–75. https://doi.org/10.1093/bioinformatics/btx086.

Steen, Andrew D., Alexander Crits-Christoph, Paul Carini, Kristen M. DeAngelis, Noah Fierer, Karen G. Lloyd, and J. Cameron Thrash. 2019. 'High Proportions of Bacteria and Archaea across Most Biomes Remain Uncultured'. *The ISME Journal* 13 (12): 3126–30. https://doi.org/10.1038/s41396-019-0484-y.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. 'Structure and Function of the Global Ocean Microbiome'. *Science* 348 (6237). https://doi.org/10.1126/science.1261359.

Tommaso, Paolo Di, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. 'Nextflow Enables Reproducible Computational Workflows'. *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Torrijos, Michel. 2016. 'State of Development of Biogas Production in Europe'. *Procedia Environmental Sciences* 35 (December): 881–89. https://doi.org/10.1016/j.proenv.2016.07.043.

Uritskiy, Gherman V., Jocelyne DiRuggiero, and James Taylor. 2018. 'MetaWRAP—a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis'. *Microbiome* 6 (1): 158. https://doi.org/10.1186/s40168-018-0541-1.

Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. ‘Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads’. *Genome Research* 27 (5): 737–46. https://doi.org/10.1101/gr.214270.116.

Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. ‘Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement’. *PloS One* 9 (11): e112963. https://doi.org/10.1371/journal.pone.0112963.

Wellinger, Arthur, Jerry D. Murphy, and David Baxter. 2013. *The Biogas Handbook: Science, Production and Applications*. Elsevier.

Westreich, Samuel T., Michelle L. Treiber, David A. Mills, Ian Korf, and Danielle G. Lemay. 2018. ‘SAMSA2: A Standalone Metatranscriptome Analysis Pipeline’. *BMC Bioinformatics* 19 (1): 175. https://doi.org/10.1186/s12859-018-2189-z.

Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. ‘Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads’. *PLoS Computational Biology* 13 (6): e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. ‘MaxBin: An Automated Binning Method to Recover Individual Genomes from Metagenomes Using an Expectation-Maximization Algorithm’. *Microbiome* 2 (1): 26. https://doi.org/10.1186/2049-2618-2-26.

Appendix 1 - CheckM and sourmash (GTDB) results

Appendix 1; Table 1 *- CheckM quality check*

Bin Id	genomes	markers	marker sets	0	1	2	3	4	5+	Complete-ness	Contamina-tion	Strain heterogeneity
bin.01	35	420	196	16	394	10	0	0	0	94.19	2.16	0.00
bin.02	3167	126	75	1	124	1	0	0	0	98.67	1.33	0.00
bin.03	930	213	118	4	207	2	0	0	0	97.25	1.13	0.00
bin.04	5443	103	58	9	94	0	0	0	0	90.86	0.00	0.00
bin.05	304	250	143	60	189	1	0	0	0	81.77	0.70	0.00
bin.06	100	295	158	11	283	1	0	0	0	96.13	0.16	100.00
bin.07	387	223	124	28	195	0	0	0	0	81.25	0.00	0.00
bin.08	304	250	143	3	247	0	0	0	0	97.90	0.00	0.00
bin.09	1324	176	102	10	165	1	0	0	0	91.18	0.49	0.00
bin.10	90	234	153	32	202	0	0	0	0	86.60	0.00	0.00
bin.11	930	213	118	6	203	4	0	0	0	95.73	1.98	0.00
bin.12	5443	103	58	1	97	5	0	0	0	98.28	3.97	40.00
bin.13	433	273	183	8	264	1	0	0	0	96.72	0.27	100.00
bin.14	172	263	149	47	212	4	0	0	0	77.14	2.35	0.00
bin.15	930	213	118	6	206	1	0	0	0	97.46	0.42	0.00
bin.16	148	187	124	1	186	0	0	0	0	99.60	0.00	0.00
bin.17	1324	176	102	13	156	7	0	0	0	89.51	5.43	0.00
bin.18	5443	105	59	1	98	6	0	0	0	98.31	6.78	66.67
bin.19	350	316	210	15	299	2	0	0	0	93.57	0.71	0.00
bin.20	174	149	89	35	113	1	0	0	0	71.16	0.56	0.00
bin.21	1324	176	102	20	155	1	0	0	0	83.82	0.98	0.00
bin.22	930	213	118	12	201	0	0	0	0	90.47	0.00	0.00
bin.23	5443	105	59	18	86	1	0	0	0	74.58	1.69	0.00
bin.24	3167	126	75	37	87	2	0	0	0	72.32	0.78	50.00
bin.25	1324	176	102	6	166	4	0	0	0	94.61	3.43	0.00
bin.26	1324	176	102	5	168	3	0	0	0	95.10	1.96	0.00
bin.27	1324	175	101	12	163	0	0	0	0	89.60	0.00	0.00
bin.28	930	213	118	14	198	1	0	0	0	88.77	0.85	0.00
bin.29	1318	179	104	5	166	8	0	0	0	95.67	3.93	0.00
bin.30	1318	179	104	8	168	3	0	0	0	92.31	1.92	0.00
bin.31	1324	176	102	3	171	2	0	0	0	98.01	1.47	0.00
bin.32	930	207	114	2	202	3	0	0	0	98.25	0.95	33.33
bin.33	930	213	118	11	201	1	0	0	0	91.74	0.85	0.00
bin.34	1324	176	102	10	163	3	0	0	0	91.67	1.05	0.00
bin.35	5443	105	59	4	101	0	0	0	0	94.76	0.00	0.00

Appendix 1; Table 2 *- CheckM Lineage*

Bin Id	Marker lineage	UID
bin.01	c__Clostridia	(UID1085)
bin.02	k__Bacteria	(UID2328)
bin.03	p__Firmicutes	(UID241)
bin.04	k__Bacteria	(UID209)
bin.05	o__Clostridiales	(UID1120)
bin.06	p__Firmicutes	(UID1022)
bin.07	c__Clostridia	(UID1118)
bin.08	o__Clostridiales	(UID1120)
bin.09	p__Firmicutes	(UID239)
bin.10	p__Euryarchaeota	(UID54)
bin.11	p__Firmicutes	(UID241)
bin.12	k__Bacteria	(UID209)
bin.13	k__Bacteria	(UID2570)
bin.14	o__Clostridiales	(UID1212)
bin.15	p__Firmicutes	(UID241)
bin.16	p__Euryarchaeota	(UID3)
bin.17	p__Firmicutes	(UID239)
bin.18	k__Bacteria	(UID209)
bin.19	p__Bacteroidetes	(UID2605)
bin.20	k__Bacteria	(UID2329)
bin.21	p__Firmicutes	(UID239)
bin.22	p__Firmicutes	(UID241)
bin.23	k__Bacteria	(UID209)
bin.24	k__Bacteria	(UID2328)
bin.25	p__Firmicutes	(UID239)
bin.26	p__Firmicutes	(UID239)
bin.27	p__Firmicutes	(UID239)
bin.28	p__Firmicutes	(UID241)
bin.29	p__Firmicutes	(UID240)
bin.30	p__Firmicutes	(UID240)
bin.31	p__Firmicutes	(UID239)
bin.32	p__Firmicutes	(UID241)
bin.33	p__Firmicutes	(UID241)
bin.34	p__Firmicutes	(UID239)
bin.35	k__Bacteria	(UID209)

Appendix 1; Table 3 *- Sourmash taxonomic lineage (superkingdom to order)*

Bin Id	status	superkingdom	phylum	class	order
bin.01	found	d__Bacteria	p__Firmicutes_B	c__Syntrophomonadia	o__Syntrophomonadales
bin.02	found	d__Bacteria	p__Firmicutes	c__Bacilli	o__ML615J-28
bin.03	found	d__Bacteria	p__Firmicutes_G	c__UBA4882	o__UBA10575
bin.04	found	d__Bacteria	p__Thermotogota	c__Thermotogae	o__Petrotogales
bin.05	found	d__Bacteria	p__Firmicutes_A	c__Clostridia	o__Acetivibrionales
bin.06	found	d__Bacteria	p__Firmicutes_B	c__Syntrophomonadia	o__Thermacetogeniales
bin.07	found	d__Bacteria	p__Firmicutes_A	c__Clostridia	o__4C28d-15
bin.08	nomatch				
bin.09	found	d__Bacteria	p__Firmicutes_G	c__SHA-98	o__UBA4971
bin.10	found	d__Archaea	p__Halobacterota	c__Methanomicrobia	o__Methanomicrobiales
bin.11	found	d__Bacteria	p__Firmicutes_G	c__Limnochordia	o__DTU010
bin.12	found	d__Bacteria	p__Thermotogota	c__Thermotogae	o__Petrotogales
bin.13	disagree	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Bacteroidales
bin.14	found	d__Bacteria	p__Firmicutes_A	c__Clostridia	o__Acetivibrionales
bin.15	found	d__Bacteria	p__DTU030	c__DTU030	o__DTU030
bin.16	found	d__Archaea	p__Thermoplasmatota	c__Thermoplasmata	o__Methanomassiliicoccales
bin.17	nomatch				
bin.18	found	d__Bacteria	p__Caldatribacteriota	c__Caldatribacteriia	o__Caldatribacteriales
bin.19	found	d__Bacteria	p__Bacteroidota	c__Bacteroidia	o__Bacteroidales
bin.20	nomatch				
bin.21	found	d__Bacteria	p__Firmicutes_G	c__Limnochordia	o__DTU080
bin.22	nomatch				
bin.23	found	d__Bacteria	p__Caldatribacteriota	c__Caldatribacteriia	o__Caldatribacteriales
bin.24	found	d__Bacteria	p__Firmicutes	c__Bacilli	o__ML615J-28
bin.25	found	d__Bacteria	p__Firmicutes_G	c__SHA-98	o__DTUO25
bin.26	disagree	d__Bacteria	p__Firmicutes_G		
bin.27	found	d__Bacteria	p__Firmicutes_E	c__DTU015	o__D8A-2
bin.28	nomatch				
bin.29	found	d__Bacteria	p__Firmicutes_A	c__Thermovenabulia	o__Thermovenabulales
bin.30	found	d__Bacteria	p__Firmicutes_G	c__Limnochordia	o__DTU010
bin.31	nomatch				
bin.32	found	d__Bacteria	p__Firmicutes_F	c__Halanaerobiia	o__Halanaerobiales
bin.33	found	d__Bacteria	p__Firmicutes_D	c__Dethiobacteria	o__DTU022
bin.34	found	d__Bacteria	p__Firmicutes_G	c__DTU065	o__DTU065
bin.35	nomatch				

Appendix 1; Table 4 *- Sourmash taxonomic classification (family to species)*

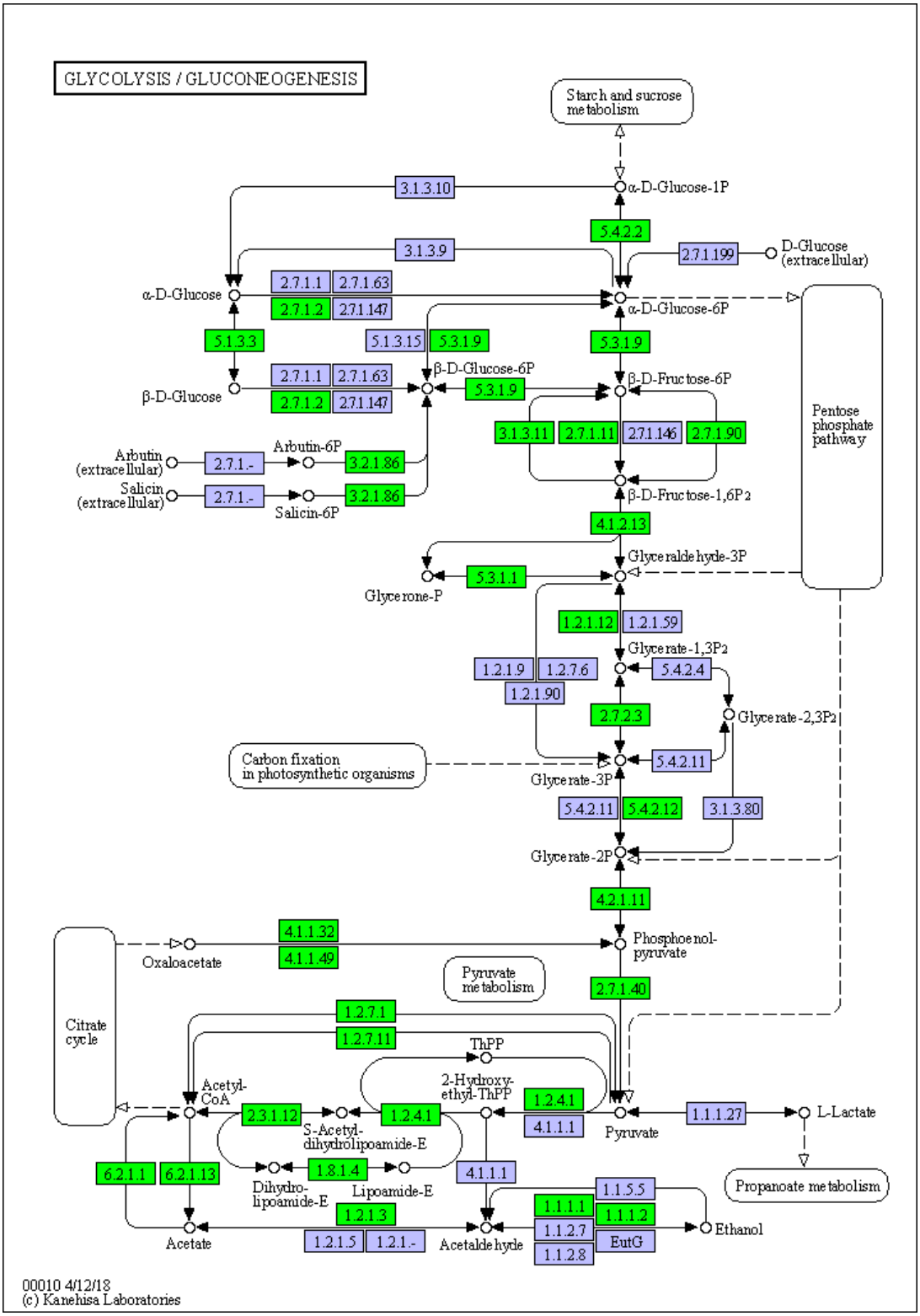
Bin Id	status	family	genus	species
bin.01	found	f__Syntrophomonadaceae	g__DTU018	s__DTU018 sp003444615
bin.02	found	f__CAG-698	g__UBA3946	s__UBA3946 sp002385755
bin.03	found	f__UBA3943	g__UBA3943	s__UBA3943 sp002385625
bin.04	found	f__Petrotogaceae	g__Defluviitoga	s__Defluviitoga tunisiensis
bin.05	found	f__Acetivibrionaceae	g__DTU013	s__DTU013 sp002385815
bin.06	found	f__Thermacetogeniaceae	g__DTU068	s__DTU068 sp001513545
bin.07	found	f__DTU072	g__DTU072	s__DTU072 sp001512685
bin.08	nomatch			
bin.09	found	f__UBA4971	g__UBA2557	s__UBA2557 sp900019985
bin.10	found	f__Methanocullaceae	g__Methanoculleus	s__Methanoculleus thermohydrogenotrophicum
bin.11	found	f__DTU010	g__DTU010	s__DTU010 sp002391385
bin.12	found	f__Petrotogaceae	g__Defluviitoga	s__Defluviitoga tunisiensis
bin.13	disagree	f__Dysgonomonadaceae	g__UBA4179	
bin.14	found	f__Acetivibrionaceae	g__Herbivorax	s__Herbivorax saccincola
bin.15	found	f__DTU030	g__DTU030	s__DTU030 sp001513125
bin.16	found	f__Methanomassiliicoccaceae	g__DTU008	s__DTU008 sp001512965
bin.17	nomatch			
bin.18	found	f__Caldatribacteriaceae	g__UBA3950	s__UBA3950 sp002385475
bin.19	found	f__DTU049	g__DTU049	s__DTU049 sp001512885
bin.20	nomatch			
bin.21	found	f__DTU080	g__DTU080	s__DTU080 sp001513395
bin.22	nomatch			
bin.23	found	f__Caldatribacteriaceae	g__UBA3950	s__UBA3950 sp002385475
bin.24	found	f__CAG-698	g__DTU056	s__DTU056 sp001512985
bin.25	found	f__DTU025	g__DTU025	s__DTU025 sp001513145
bin.26	disagree			
bin.27	found	f__D2	g__DTU015	s__DTU015 sp001513185
bin.28	nomatch			
bin.29	found	f__Tepidanaerobacteraceae	g__DTU063	s__DTU063 sp001512695
bin.30	found	f__DTU012	g__DTU012	s__DTU012 sp900019385
bin.31	nomatch			
bin.32	found	f__DTU029	g__DTU029	s__DTU029 sp001512435
bin.33	found	f__DTU022	g__DTU022	s__DTU022 sp001512835
bin.34	found	f__DTU065	g__DTU065	s__DTU065 sp001512545
bin.35	nomatch			

Appendix 2 – Parser results

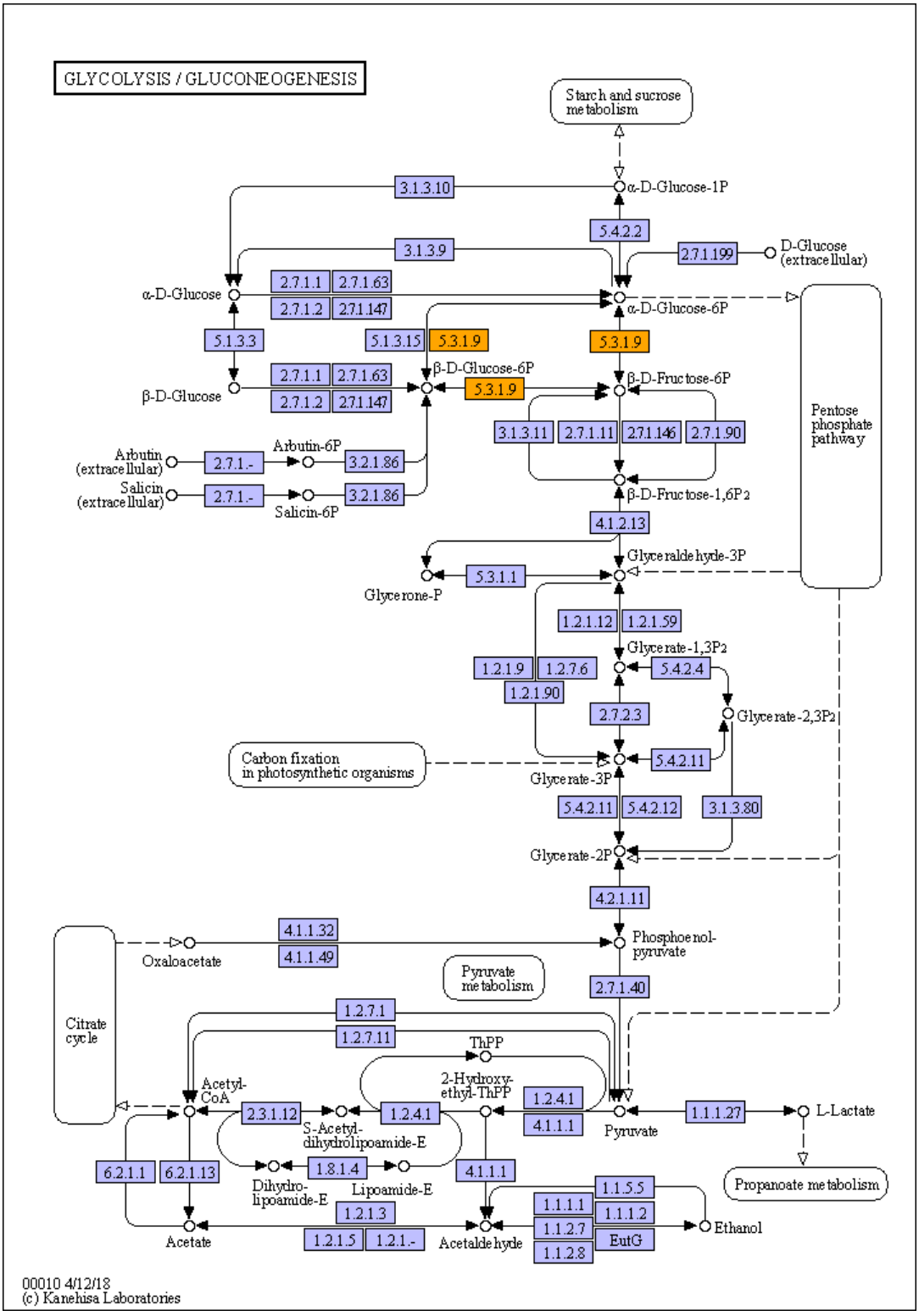
A. The glycolysis

Appendix 2; Table 1 *- the gene "expression." of the glycolysis pathway in the bins*

Bin ID	"Expressed genes."	"Non-expressed genes."	Sourmash family
bin.1	7	0	f__Syntrophomonadaceae
bin.2	6	0	f__CAG-698
bin.3	5	0	f__UBA3943
bin.4	9	0	f__Petrotogaceae
bin.5	21	0	f__Acetivibrionaceae
bin.6	11	0	f__Thermacetogeniaceae
bin.7	2	0	f__DTU072
bin.8	1	0	
bin.9	7	0	f__UBA4971
bin.10	10	0	f__Methanocullaceae
bin.11	9	0	f__DTU010
bin.12	7	0	f__Petrotogaceae
bin.13	5	0	f__Dysgonomonadaceae
bin.14	11	0	f__Acetivibrionaceae
bin.15	2	0	f__DTU030
bin.16	0	0	f__Methanomassiliicoccaceae
bin.17	1	0	
bin.18	16	1	f__Caldatribacteriaceae
bin.19	5	0	f__DTU049
bin.20	1	0	
bin.21	4	0	f__DTU080
bin.22	3	0	
bin.23	16	4	f__Caldatribacteriaceae
bin.24	3	0	f__CAG-698
bin.25	1	0	f__DTU025
bin.26	2	0	
bin.27	0	0	f__D2
bin.28	11	0	
bin.29	1	0	f__Tepidanaerobacteraceae
bin.30	7	0	f__DTU012
bin.31	8	2	
bin.32	11	0	f__DTU029
bin.33	1	0	f__DTU022
bin.34	4	0	f__DTU065
bin.35	2	0	



Appendix 2; Figure 1- The glycolysis pathway with the "expressed genes" highlighted in green.

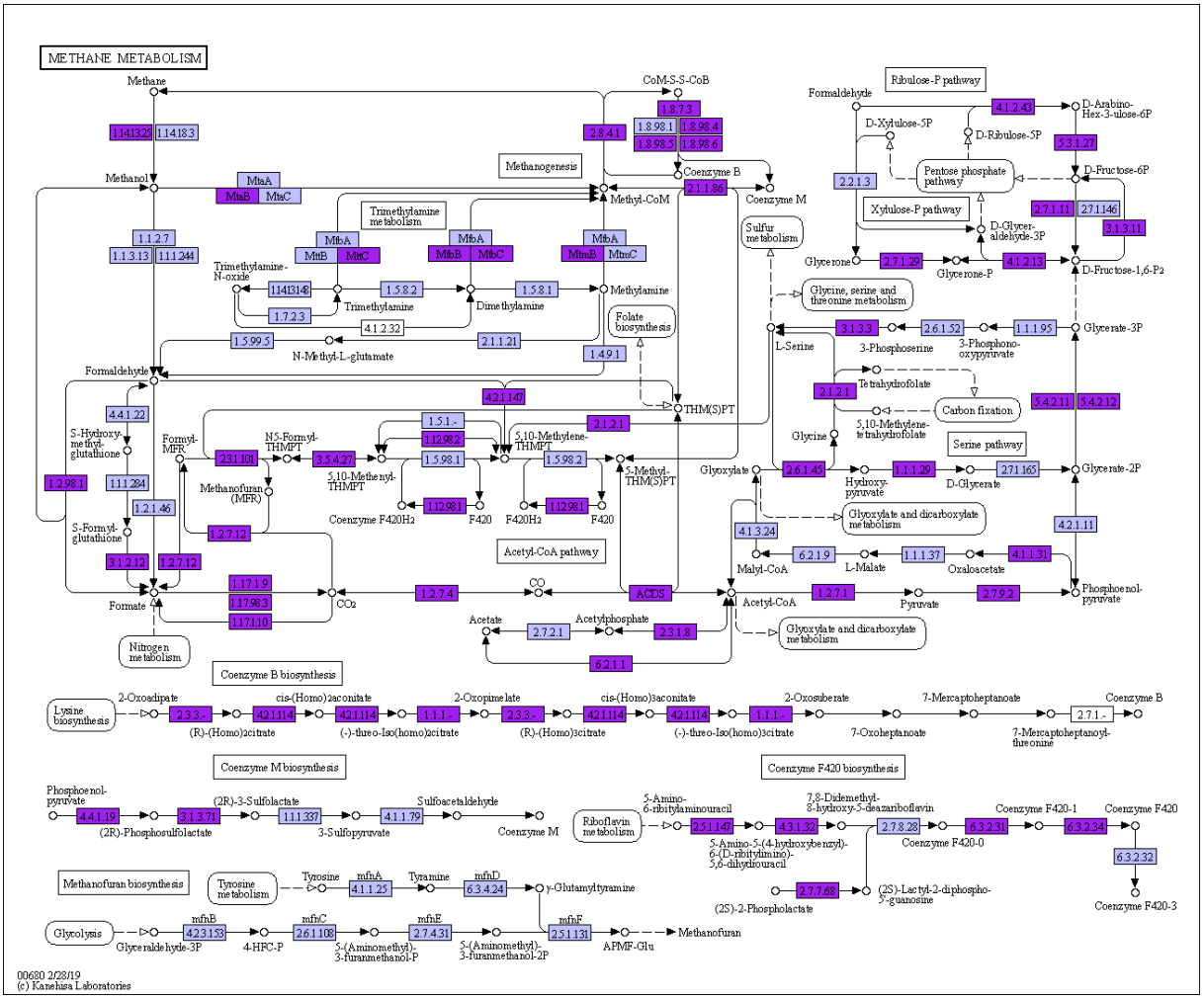


Appendix 2; Figure 2 - The glycolysis pathway with the "non-expressed genes" highlighted in orange.

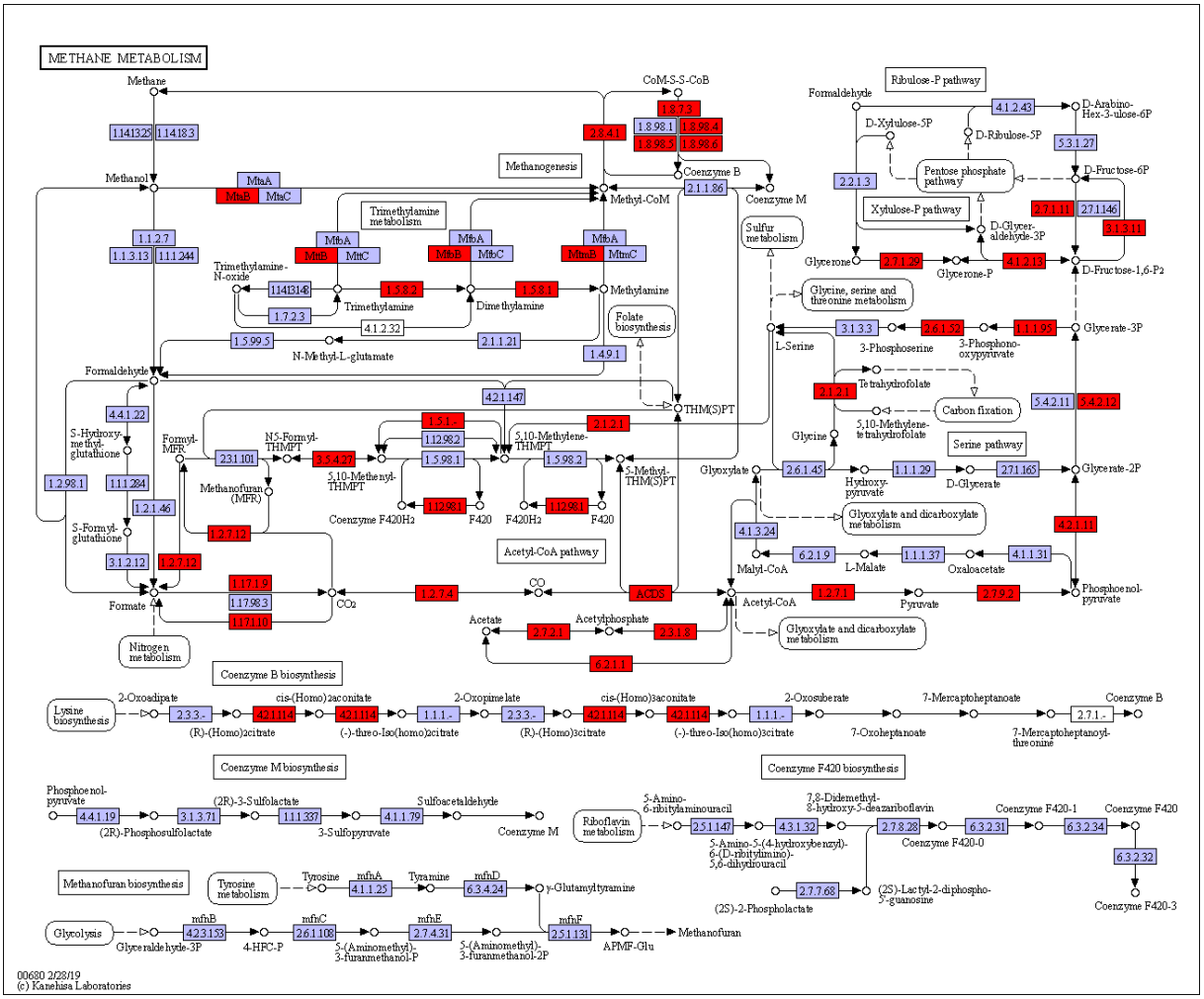
B. Methane metabolism

Appendix 2; Table 2 *- the gene "expression." of the methane metabolism pathway in the bins*

Bin ID	"Expressed genes."	"Non-expressed genes."	Sourmash family
bin.1	2	1	f__Syntrophomonadaceae
bin.2	4	0	f__CAG-698
bin.3	2	1	f__UBA3943
bin.4	4	0	f__Petrotogaceae
bin.5	6	0	f__Acetivibrionaceae
bin.6	20	0	f__Thermacetogeniaceae
bin.7	4	2	f__DTU072
bin.8	8	4	
bin.9	3	0	f__UBA4971
bin.10	9	0	f__Methanocullaceae
bin.11	5	0	f__DTU010
bin.12	4	0	f__Petrotogaceae
bin.13	4	1	f__Dysgonomonadaceae
bin.14	4	0	f__Acetivibrionaceae
bin.15	3	1	f__DTU030
bin.16	4	0	f__Methanomassiliicoccaceae
bin.17	0	0	
bin.18	12	0	f__Caldatribacteriaceae
bin.19	7	0	f__DTU049
bin.20	2	0	
bin.21	15	3	f__DTU080
bin.22	6	1	
bin.23	14	0	f__Caldatribacteriaceae
bin.24	2	0	f__CAG-698
bin.25	0	0	f__DTU025
bin.26	0	0	
bin.27	0	0	f__D2
bin.28	9	0	
bin.29	7	1	f__Tepidanaerobacteraceae
bin.30	0	0	f__DTU012
bin.31	9	3	
bin.32	2	0	f__DTU029
bin.33	9	0	f__DTU022
bin.34	1	0	f__DTU065
bin.35	5	0	



Appendix 2; Figure 3 - The methane metabolism pathway with the RNAseq "genes highlighted in purple.

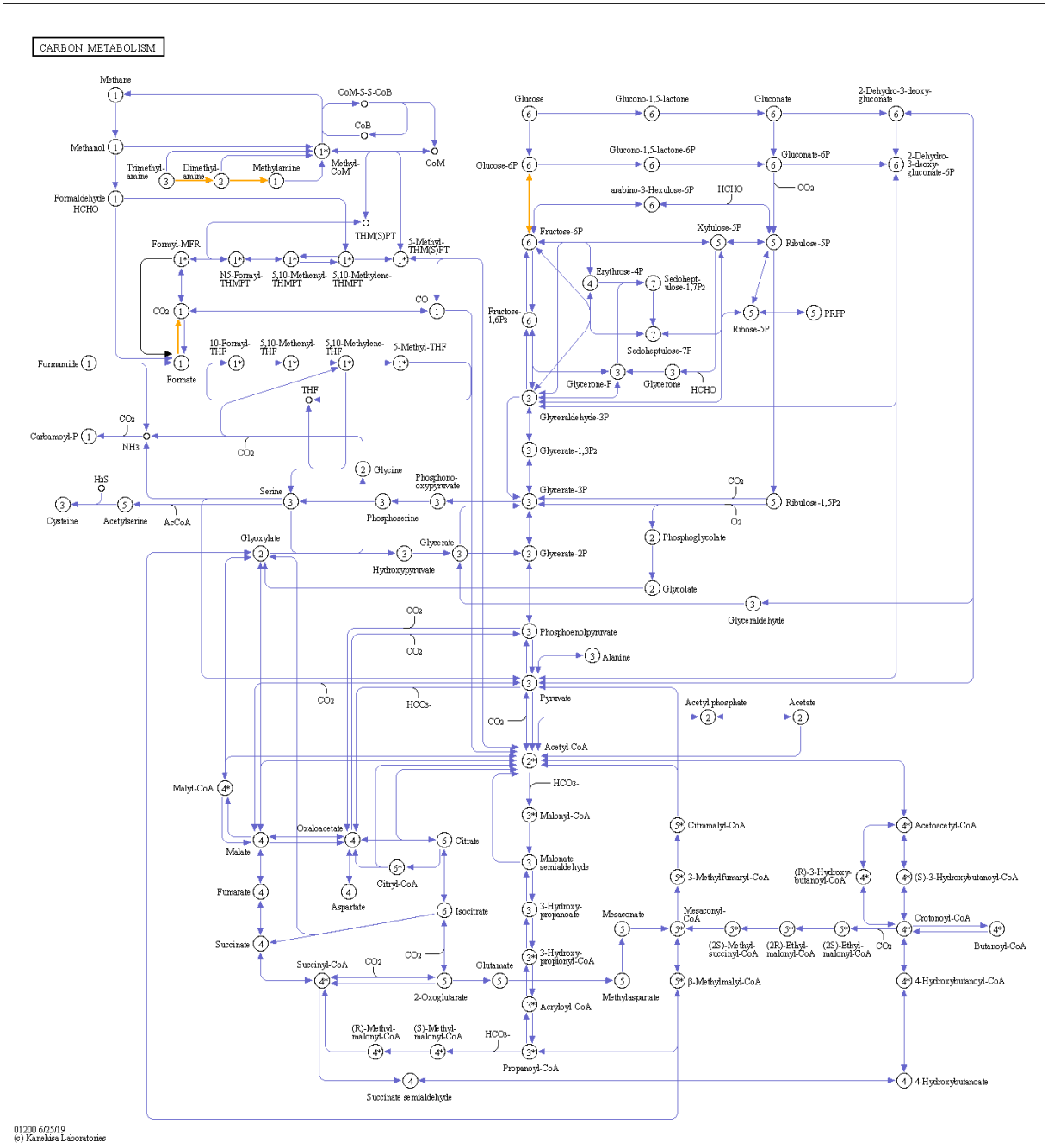


Appendix 2; Figure 4 - The methane metabolism pathway with all the genes present in the bins highlighted in red. No distinction between "expressed" and "non-expressed."

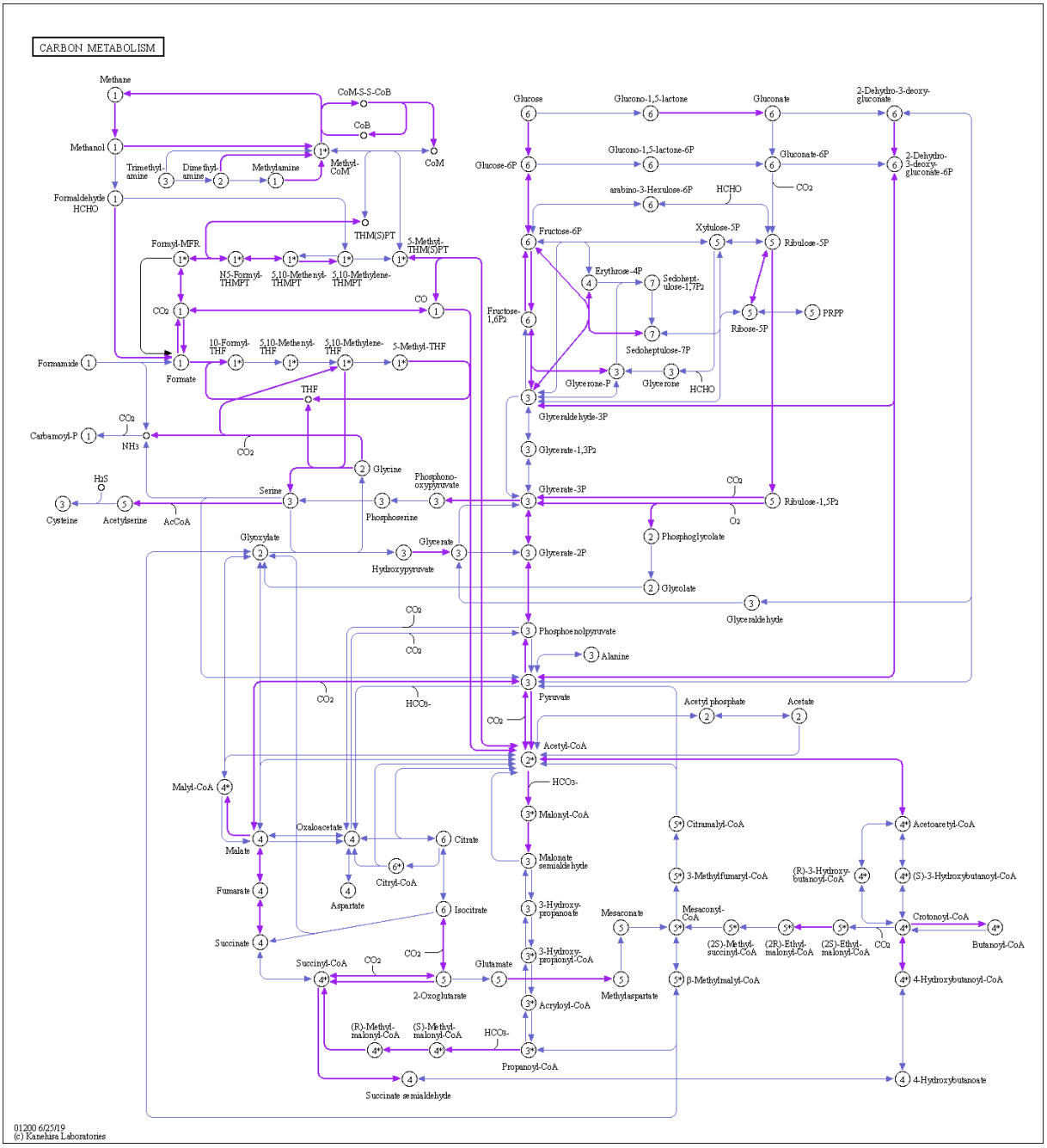
C. Carbon Metabolism

Appendix 2; Table 3 *- the gene "expression." of the carbon metabolism pathway in the bins*

Bin ID	"Expressed genes."	"Non-expressed genes."	Sourmash family
bin.1	23	2	f__Syntrophomonadaceae
bin.2	10	0	f__CAG-698
bin.3	5	1	f__UBA3943
bin.4	16	0	f__Petrotogaceae
bin.5	29	0	f__Acetivibrionaceae
bin.6	33	0	f__Thermacetogeniaceae
bin.7	9	3	f__DTU072
bin.8	18	4	
bin.9	11	0	f__UBA4971
bin.10	10	0	f__Methanocullaceae
bin.11	15	0	f__DTU010
bin.12	18	0	f__Petrotogaceae
bin.13	14	1	f__Dysgonomonadaceae
bin.14	5	0	f__Acetivibrionaceae
bin.15	17	0	f__DTU030
bin.16	3	0	f__Methanomassiliicoccaceae
bin.17	4	0	
bin.18	30	1	f__Caldatribacteriaceae
bin.19	20	0	f__DTU049
bin.20	1	0	
bin.21	24	3	f__DTU080
bin.22	8	1	
bin.23	26	2	f__Caldatribacteriaceae
bin.24	6	0	f__CAG-698
bin.25	6	0	f__DTU025
bin.26	4	0	
bin.27	2	0	f__D2
bin.28	32	0	
bin.29	12	1	f__Tepidanaerobacteraceae
bin.30	6	0	f__DTU012
bin.31	20	2	
bin.32	11	0	f__DTU029
bin.33	17	0	f__DTU022
bin.34	7	0	f__DTU065
bin.35	17	0	



Appendix 2; Figure 5 - The carbon metabolism pathway with the "non-expressed genes" highlighted in orange.



Appendix 2; Figure 6 - The carbon metabolism pathway with the RNAseq "genes" highlighted in purple.

Appendix 3 – MUFFIN manuscript

The manuscript for the publication of MUFFIN is available here https://www.biorxiv.org/content/10.1101/2020.02.08.939843v1, and in the following pages.

Metagenomics workflow for hybrid assembly, differential coverage binning, transcriptomics and pathway analysis (MUFFIN)

Renaud Van Damme[1,2], Martin Hölzer[4], Adrian Viehweger[3,4], Bettina Müller[1], Erik Bongcam-Rudloff[2], Christian Brandt[2,5]

1. Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden
2. Dept. Animal Breeding and Genetics, Bioinformatics section. Swedish University of Agricultural, Sciences, Uppsala, Sweden
3. Department of Medical Microbiology, University Hospital Leipzig, Germany
4. RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Jena, Germany
5. Institute for Infectious Diseases and Infection Control, Jena University Hospital, Jena, Germany

Abstract

Metagenomics has redefined many areas of microbiology. However, metagenome-assembled genomes (MAGs) are often fragmented, primarily when sequencing was performed with short reads. Recent long-read sequencing technologies promise to improve genome reconstruction. However, the integration of two different sequencing modalities makes downstream analyses complex. We, therefore, developed MUFFIN, a complete metagenomic workflow that uses short and long reads to produce high-quality bins and their annotations. The workflow is written by using Nextflow, a workflow orchestration software, to achieve high reproducibility and fast and straightforward use. This workflow also produces the taxonomic classification and KEGG pathways of the bins and can be further used by providing RNA-Seq data (optionally) for quantification and annotation. We tested the workflow using twenty biogas reactor samples and assessed the capacity of MUFFIN to process and output relevant files needed to analyze the microbial community and their function. MUFFIN produces functional pathway predictions and if provided *de novo* transcript annotations across the metagenomic sample and for each bin.

Author Summary

RVD did the development and design of MUFFIN and wrote the first draft; BM and EBR did the critical reading and correction of the manuscript; MH did the critical reading of the manuscript and the general adjustments for the metagenomic workflow; AV did the critical reading of the manuscript and adjustments for the taxonomic classifications. CB supervised the project, did the workflow design, helped with the implementation, and revised the manuscript.

Introduction

Metagenomics is widely used to analyze the composition, structure, and dynamics of microbial communities, as it provides deep insights into uncultivatable organisms and their relationship to each other [1–5]. In this context, whole metagenome sequencing is mainly performed using short-read sequencing technologies, predominantly provided by Illumina. Not surprisingly, the vast majority of tools and workflows for the analysis of metagenomic samples are designed around short reads. However, long-read sequencing technologies such as provided by PacBio or Oxford Nanopore Technologies (ONT) retrieve genomes from metagenomic datasets with higher completeness and less contamination [6]. The long-read information bridges gaps in a short-read-only assembly that often occur due to intra- and interspecies repeats [6]. Complete viral genomes can be already identified from environmental samples without any assembly step via nanopore-based sequencing [7]. Combined with a reduction in cost per gigabase [8] and an increase in data output, the technologies for sequencing long reads quickly became suitable for metagenomic analysis [9–12]. In particular, with the MinION, ONT offers mobile and cost-effective sequencing device for long reads that paves the way for the real-time analysis of metagenomic samples. Currently, the combination of both worlds (long reads and high-precision short reads) allows the reconstruction of more complete and more accurate metagenome-assembled genomes (MAGs) [6].

One of the main challenges and bottlenecks of current metagenome sequencing studies is the orchestration of various computational tools into stable and reproducible workflows to

analyze the data. A recent study from 2019 involving 24,490 bioinformatics software resources showed that 26 % of all these resources are not currently online accessible [13]. Among 99 randomly selected tools, 49 % were deemed 'difficult to install,' and 28 % ultimately failed the installation procedure. For a large-scale metagenomics study, various tools are needed to analyze the data comprehensively. Thus, already during the installation procedure, various issues arise related to missing system libraries, conflicting dependencies and environments or operating system incompatibilities. Even more complicating, metagenomic workflows are computing intense and need to be compatible with high-performance compute clusters (HPCs), and thus different workload managers such as SLURM or LSF. We combined the workflow manager Nextflow[14] with virtualization software (so-called 'containers') to generate reproducible results in various working environments and allow full parallelization of the workload on a higher degree.

Several workflows for metagenomic analyses have been published, including MetaWRAP(v1.2.1)[15], Anvi'o[16], SAMSA2[17], Humann[18], or MG-Rast[19]. Unlike those, MUFFIN allows for a hybrid metagenomic approach combining the strengths of short and long reads. It ensures reproducibility through the use of a workflow manager and reliance on either install recipes (Conda [20]) or containers (Docker[21]).

Design and implementation

MUFFIN integrates state-of-the-art bioinformatic tools via Conda recipes or Docker containers for the processing of metagenomic sequences in a Nextflow workflow environment (Figure 1). MUFFIN executes three steps subsequently or separately if intermediate results, such as MAGs, are available. As a result, a more flexible workflow execution is possible. The three steps represent common metagenomic analysis tasks and are summarized in Figure 1:

1. Assemble: Hybrid assembly and binning

2. Classify: Bin quality control and taxonomic assessment

3. Annotate: Bin annotation and KEGG pathway summary

The workflow takes paired-end Illumina reads (short reads) and nanopore-based reads (long reads) as input for the assembly and binning and allows for additional user-provided read sets for differential coverage binning. Differential coverage binning facilitates genome bins with higher completeness than other currently used methods [22]. Step 2 will be executed automatically after the assembly and binning procedure or can be executed independently by providing MUFFIN a directory containing MAGs in FASTA format. In step 3, paired-end RNA-Seq data can be optionally supplemented to improve the annotation of bins.

On completion, MUFFIN provides various outputs such as the MAGs, KEGG pathways, and bin quality/annotations. Additionally, all mandatory databases are automatically downloaded and stored in the working directory or can be alternatively provided via an input flag.

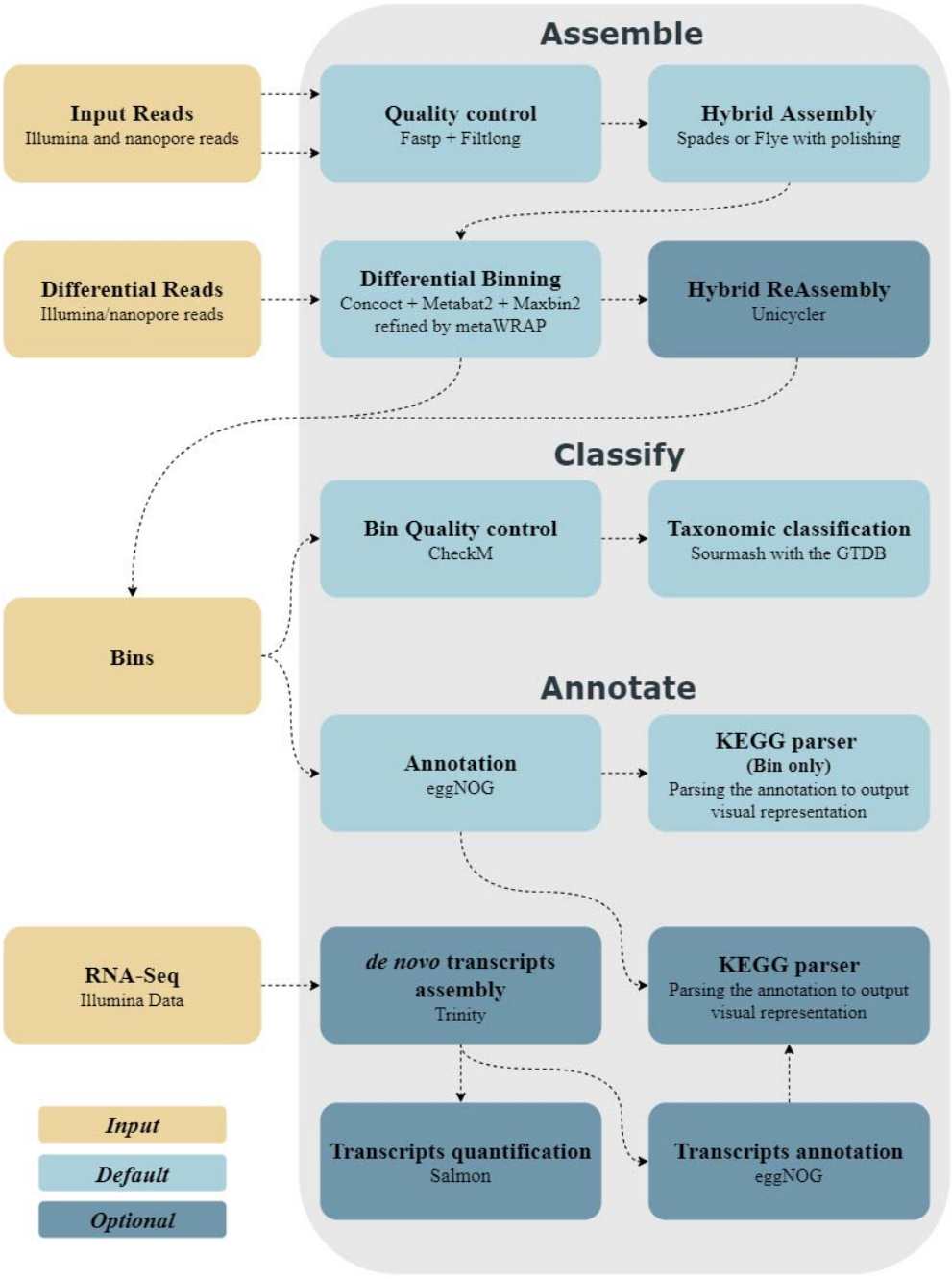


Figure 1: Simplified overview of the MUFFIN workflow. All three steps (Assemble, Classify, Annotate) from top to bottom are shown. The RNA-Seq data for Step 3 (Annotate) is optional.

Step 1 - Assemble: Hybrid assembly and binning

The first step (**Assembly and binning**), uses metagenomic nanopore-based long reads and Illumina paired-end short reads to obtain high-quality and highly complete bins. The short-read quality control is operated using fastp (v0.20.0) [23]. Optionally, Filtlong (v0.2.0) [24] can be

used to discard long reads below a length of 1000 bp [24]. The hybrid assembly can be performed according to two principles, which differ substantially in the read set to begin with. The default approach starts from a short-read assembly where contigs are bridged via the long reads using metaSPAdes (v3.13.1) [25–27]. Alternatively, MUFFIN can be executed starting from a long-read-only assembly using metaFlye (v2.6) [28,29] followed by polishing the assembly with the long reads using Racon (v1.4.7) [30] and medaka (v0.11.0) [31] and finalizing the error correction by incorporating the short reads using multiple rounds of Pilon (v1.23) [32].

Binning is the most crucial step during metagenomic analysis. Therefore, MUFFIN combines three different binning software tools, respectively CONCOCT (v1.0.0) [33], MaxBin2 (v2.2.4) [34], and MetaBAT2 (v2.14) [35] and refine these bins via MetaWRAP (v1.2.1)[15]. The user can provide additional read data sets (short or long reads) to perform automatically differential coverage binning to assign contigs to their bins better.

Moreover, an additional reassembly of bins has shown the capacity to increase the completeness and N50 while decreasing the contamination of the bins[15]. Therefore, MUFFIN allows for an optional reassembly to improve the continuity of the MAGs further. This re-assembly is performed by retrieving the reads belonging to one bin and doing an assembly with Unicycler (v0.4.8) [36].

To support a transparent and reproducible metagenomics workflow, all reads that cannot be mapped back to the existing high-quality bins (after the refinement) are available as an output for further analysis. These reads could be further analyzed by other tools or, e.g., used as a new input to run MUFFIN while providing other read sets for the differential coverage binning to extract additional high-quality bins.

Step 2 - Classify: Bin quality control and taxonomic assessment

In the second step (**Bin quality control and taxonomic assessment**), the quality of the bins is evaluated with CheckM (v1.0.18) [37] followed by assigning a taxonomic classification to the bins using sourmash (v2.0.0a10) [38] and the Genome Taxonomy Database (GTDB

release r89) [39]. The GTDB was chosen as it contains many unculturable bacteria and archaea – this allows for monophyletic species assignments, which other databases do not assure [40,41]. GTDB substantially improved overall downstream results [40]. The user can also analyze other bin sets in this step regardless of their origin by providing a directory with multiple FASTA files (bins).

Step 3 - Annotate: Bin annotation and KEGG pathway summary

The last step of MUFFIN (**Bin annotation and output summary**) comprises the annotation of the bins using eggNOG-mapper (v2.0.1) [42] and the eggNOG database (v5) [43]. If RNA-Seq data of the metagenome sample is provided (Illumina, paired-end), quality control using fastp (v0.20.0) [23] and a *de novo* transcript assembly using Trinity (v2.8.5) [44] followed by a quasi-mapping transcript quantification using Salmon (v0.15.0) [45] are performed. Lastly, the transcripts are annotated using eggNOG-mapper (v2.0.1) [42] again, followed by a parser to output the activity of the pathway graphically in relation to the sample level. The expression of low and high abundant genes present in the bins is shown. If only bin sets are provided without any RNA-Seq data, the pathways of all the bins are created based on gene presence alone. The KEGG pathway results are summarized in detail as interactive HTML files (example snippet: **Error! Not a valid bookmark self-reference.**).

Like step 2, this step can be directly performed with a bin set created via another workflow.

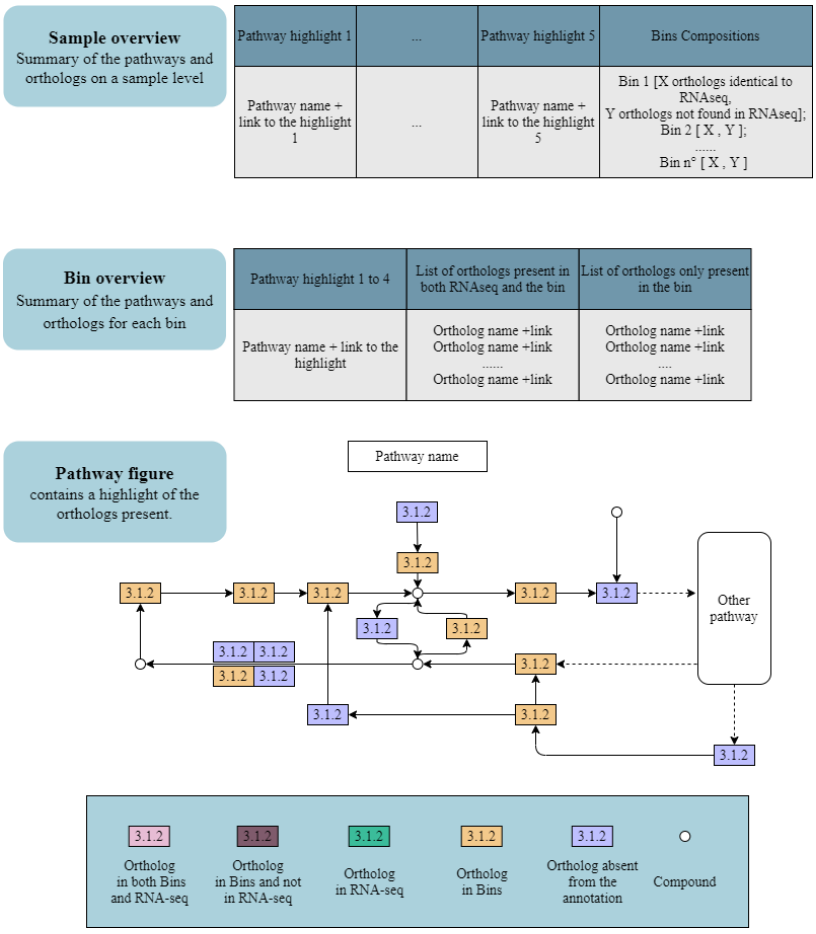


Figure 2: Example snippets of the sub-workflow results of step 3 (Annotate).

Running MUFFIN and version control

MUFFIN requires only two dependencies, which allows an easy and user-friendly workflow execution. One of them is the workflow management system Nextflow [14] and the other can be either Conda [20] as a package manager or Docker [21] to use containerized tools. A detailed Installation process is available on https://github.com/RVanDamme/MUFFIN. Each MUFFIN release specifies the Nextflow version it was tested on, to avoid any version conflicts between MUFFIN and Nextflow at any time. A Nextflow-specific version can always be directly downloaded as an executable file from https://github.com/nextflow-io/nextflow/releases, which can then be paired with a compatible MUFFIN version via the -r flag.

Results

We chose Nextflow for the development of our metagenomic workflow because of its direct cloud computing support (Amazon AWS, Google Life Science, Kubernetes), various ready-to-use batch schedulers (SGE, SLURM, LSF), state-of-the-art container support (Docker, Singularity) and accessibility of a widely used software package manager (Conda). Moreover, Nextflow [14] provides a practical and straightforward intermediary file handling with process-specific work directories and the possibility to resume failed executions where the work ceased. Additionally, the workflow code itself is separated from the 'profile' code (which contains Docker, Conda, or cluster related code), which allows for a convenient and fast workflow adaptation to different computing clusters without touching or changing the actual workflow code.

The entire MUFFIN workflow was executed on 20 samples from the Bioproject PRJEB34573 (available at ENA or NCBI) using the Cloud Life Sciences API (google cloud) with docker containers. This metagenomic bioreactor study provides paired-end Illumina and nanopore-based data for each sample [41]. We used five different Illumina read sets of the same project for differential coverage binning, and the workflow runtime was less than two days for all samples. MUFFIN was able to retrieve 1122 MAGs with genome completeness of at least 70 % and contamination of less than 10 % (Figure 3). In total, MUFFIN retrieved 654 MAGs with genome completeness of over 90 %, of which 456 have less than 2% contamination out of the 20 datasets. For comparison, a recent study was using 134 publicly available datasets from different biogas reactors and retrieved 1,635 metagenome-assembled genomes with genome completeness of over 50% [46].

Exemplarily, we investigated the impact of additional re-assembly of each bin for five samples (Figure 3). The N50 was increased by an average of 6-7 fold across all samples. Twenty-six bins of the five samples had an N50 ranging between 1 to 3 Mbases. Some bins benefit more of this step as the re-assembly performance depends on the number of reads available for each bin.

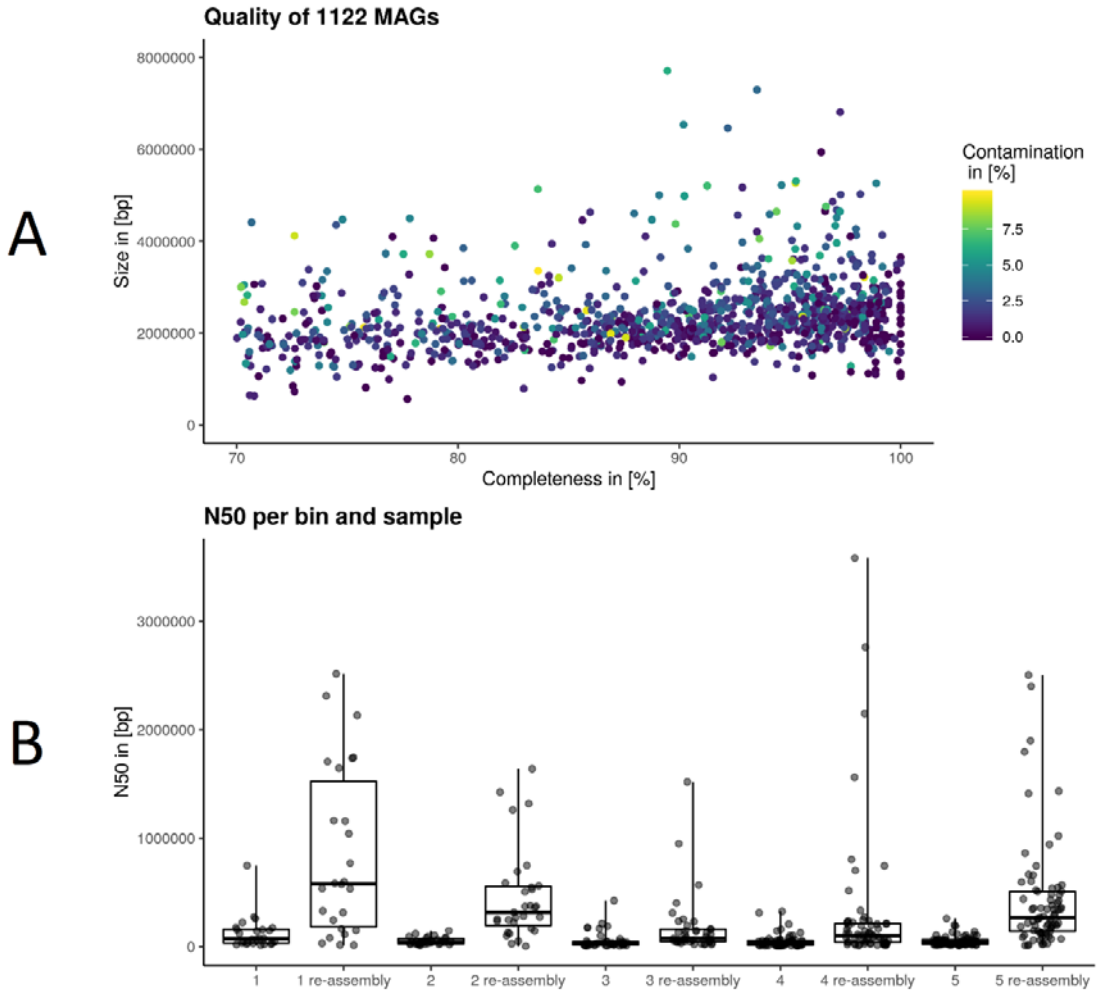


Figure 3: A: Quality overview of 1122 meta-assembled genomes (MAGs) by plotting size to completeness and coloring based on contamination level. B: N50 comparison between each bin of five selected samples from the Bioproject PRJEB34573 before and after individual bin reassembly.

Discussion

The analysis of metagenomic sequencing data evolved as an emerging and promising research field to retrieve, characterize, and analyze organisms that are difficult to cultivate. There are numerous tools available for individual metagenomics analysis tasks, but they are mainly developed independently and are often difficult to install and run. The MUFFIN workflow gathers the different steps of a metagenomics analysis in an easy-to-install, highly reproducible, and scalable workflow using Nextflow which makes them easily accessible to researchers.

MUFFIN utilizes the advantages of two sequencing technologies, whereas short reads can provide a better representation of low abundant species due to their higher coverage. This aspect is further utilized via the final re-Assembly step after binning, which is an optional step due to the additional computational burden which solely aims to improve genome continuity.

Another critical aspect is the full support of differential binning, for both long and short reads, via a single input option. The additional coverage information from other read sets of similar habitats allows for the generation of more concise bins with higher completeness and less contamination because more coverage information is available for each binning tool to decide which bin each contig belongs.

With supplied RNA-Seq data, MUFFIN is capable of enhancing the pathway results present in the metagenomic sample by incorporating this data as well as the general expression level of the genes. Such information is essential to further analyze a metagenomic data sets in-depth, for example, to define the origin of a sample or to improve environmental parameters for production reactors such as biogas reactors. Knowing whether an organism expresses a gene is a crucial element in deciding whether a more detailed analysis of that organism in the biotope where the sample was taken is necessary or not.

Availability and future directions

MUFFIN is an ongoing workflow project that gets further improved and adjusted. The modular workflow setup of MUFFIN using Nextflow allows for fast adjustments as soon as future developments in hybrid metagenomics arise, including the pre-configuration for other workload managers. MUFFIN can directly benefit from the addition of new bioinformatics software such as for differential expression analysis and short-read assembly that can be easily plugged into the modular system of the workflow. Another improvement is the creation of an advanced user and wizard user configuration file, allowing experienced users to tweak all the different parameters of all the different software as desired.

MUFFIN will further benefit from different improvements, in particular by graphically comparing the generated MAGs via a phylogenetic tree. Furthermore, a convenient approach to include negative controls is under development to allow the reliable analysis of super-low abundant organisms in metagenomic samples.

MUFFIN is publicly available at https://github.com/RVanDamme/MUFFIN under the GNU general public license v3.0. Detailed information about the program versions used and additional information can be found in the GitHub repository. All tools used by MUFFIN are listed in the supplementary table S1. The Docker images used in MUFFIN are prebuilt and publicly available at https://hub.docker.com/u/nanozoo, and the GTDB formatted for sourmash(v2.0.0a10)[38] usage is publicly available at https://osf.io/wxf9z / and was created by C. Titus Brown (associate professor at UC DAVIS, http://ivory.idyll.org/blog/2019-sourmash-lca-db-gtdb.html).

Acknowledgment

We want to thank Hadrien Gourlé and Moritz Buck for the valuable insights into metagenomic analysis and Annotation.

References

1. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* **5**, R245–R249 (1998).
2. De, R. Metagenomics: aid to combat antimicrobial resistance in diarrhea. *Gut Pathogens* **11**, 47 (2019).
3. Mukherjee, A. & Reddy, M. S. Metatranscriptomics: an approach for retrieving novel eukaryotic genes from polluted and related environments. *3 Biotech* **10**, 71 (2020).
4. Grossart, H.-P., Massana, R., McMahon, K. D. & Walsh, D. A. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography* **65**, S2–S20 (2020).

5. Carabeo-Pérez, A., Guerra-Rivera, G., Ramos-Leal, M. & Jiménez-Hernández, J. Metagenomic approaches: effective tools for monitoring the structure and functionality of microbiomes in anaerobic digestion systems. *Appl Microbiol Biotechnol* **103**, 9379–9390 (2019).
6. Overholt, W. A. *et al.* Inclusion of Oxford Nanopore long reads improves all microbial and phage metagenome-assembled genomes from a complex aquifer system. *bioRxiv* 2019.12.18.880807 (2019) doi:10.1101/2019.12.18.880807.
7. Beaulaurier, J. *et al.* Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities. *bioRxiv* 619684 (2019) doi:10.1101/619684.
8. Wetterstrand, K. A. DNA Sequencing Costs: Data. *www.genome.gov/sequencingcostsdata* www.genome.gov/sequencingcostsdata.
9. Somerville, V. *et al.* Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol* **19**, 143 (2019).
10. Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, (2019).
11. Driscoll, C. B., Otten, T. G., Brown, N. M. & Dreher, T. W. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* **12**, (2017).
12. Suzuki, Y. *et al.* Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome* **7**, 119 (2019).
13. Mangul, S., Martin, L. S., Eskin, E. & Blekhman, R. Improving the usability and archival stability of bioinformatics software. *Genome Biol.* **20**, 47 (2019).
14. Tommaso, P. D. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).

15. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

16. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

17. Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I. & Lemay, D. G. SAMSA2: a standalone metatranscriptome analysis pipeline. *BMC Bioinformatics* **19**, 175 (2018).

18. Abubucker, S. *et al.* Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLOS Computational Biology* **8**, e1002358 (2012).

19. Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).

20. Anaconda Software distribution. Anaconda | The World's Most Popular Data Science Platform. *https://anaconda.com* https://www.anaconda.com/.

21. Boettiger, C. An introduction to Docker for reproducible research. *SIGOPS Oper. Syst. Rev.* **49**, 71–79 (2015).

22. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533–538 (2013).

23. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

24. Wick, R. *rrwick/Filtlong*. (2020).

25. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).

26. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

27. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

28. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).

29. Kolmogorov, M., Rayko, M., Yuan, J., Polevikov, E. & Pevzner, P. metaFlye: scalable long-read metagenome assembly using repeat graphs. *bioRxiv* 637637 (2019) doi:10.1101/637637.

30. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

31. *nanoporetech/medaka*. (Oxford Nanopore Technologies, 2020).

32. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112963.

33. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144–1146 (2014).

34. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).

35. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities [PeerJ]. https://peerj.com/articles/1165/.

36. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005595.

37. 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).

38. Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software* https://joss.theoj.org (2016) doi:10.21105/joss.00027.

39. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996–1004 (2018).

40. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. Correcting index databases improves metagenomic studies. *bioRxiv* 712166 (2019) doi:10.1101/712166.

41. Brandt, C., Bongcam-Rudloff, E. & Müller, B. *Abundance tracking by long-read nanopore sequencing of complex microbial communities in samples from 20 different biogas/wastewater plants*. https://www.researchsquare.com/article/d9fdc53c-a0a7-44a1-bce4-4d1fafa60f9e/v1 (2019) doi:10.21203/rs.2.17734/v1.

42. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115–2122 (2017).

43. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).

44. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).

45. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).

46. Campanaro, S. *et al.* The anaerobic digestion microbiome: a collection of 1600 metagenome-assembled genomes shows high species diversity related to methane production. *bioRxiv* 680553 (2019) doi:10.1101/680553.

Funding Disclosure

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – BR 5692/1-1 and BR 5692/1-2. This material is based upon work supported by Google Cloud.

BM was funded by FORMAS, grant number 942-2015-1008. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

357 MH is supported by the Collaborative Research Centre AquaDiva (CRC 1076 AquaDiva) of
358 the Friedrich Schiller University Jena, funded by the DFG. MH appreciates the support of the
359 Joachim Herz Foundation by the add-on fellowship for interdisciplinary life science.

360