



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole de biologie

**Validation of a metagenomic pipeline for discovery of rare bile acid-
transforming bacteria**

**Travail de Maîtrise universitaire ès Sciences en Sciences moléculaires du
vivant**

Master Thesis of Science in Molecular Life Sciences

par

Léonard Jequier

Directeur-Directrice : Dr. Mauro Delorenzi

Superviseur-e-s: Dr. Andrew Janowczyk

Expert-e-s: Prof. Rizlan Bernier-Latmani

Swiss Institute of Bioinformatics - Bioinformatics Core Facility

Date (Mai 2020)

Table of content

I.	Introduction.....	4
A.	Background.....	4
1.	Bile acids and the microbiota	4
2.	Studying the microbiota: technical evolution	6
B.	Motivation and goals.....	10
II.	Methods	11
A.	Experimental dataset	11
1.	Coverage calculation and assumptions.....	11
B.	Artificial dataset	11
C.	Pipeline implementation and usage.....	14
D.	Performance assessment	15
1.	Raw reads QC	15
2.	Read trimming QC	15
3.	Assembly QC.....	15
4.	Binning QC	16
E.	Optimisation:.....	17
F.	Deduplication	17
G.	Read trimming.....	18
H.	Assembly.....	19
1.	Overlap layout consensus	19
2.	K-mer strategy.....	19
3.	Implementation in the pipeline.....	20
I.	Binning.....	20
J.	Validation	21
III.	Results	23
A.	Optimisation	23
1.	Read deduplication.....	23
2.	Read quality trimming	24
3.	Assembly.....	25
4.	Binning.....	26
A.	Validation	29
1.	Assembly:	29
2.	Binning:.....	30
IV.	Discussion	31
A.	Artificial datasets.....	31

B.	Optimisation	31
1.	Optimisation process.....	31
C.	Validation	32
V.	Conclusion	33
VI.	Glossary	34
VII.	Supplementary tables.....	34
VIII.	Scripts	34
IX.	List of figures:	35
X.	List of tables.....	36
XI.	Bibliography.....	37

Abstract: Bile acids (BA) transforming bacteria, or more specifically, bacteria performing 7-dehydroxylation (7-DH), are suspected to play a crucial part in the endocrine role of BA and the interaction between the gut microbiota and the host. 7-DH bacteria's low abundance levels in the gut microbiota, however, render them difficult to study. During this project, a bioinformatic pipeline for the analysis of whole metagenomic shotgun sequencing datasets was developed with the aim of maximizing sensitivity towards low abundance species, thus enabling the potential discovery of new 7-DH bacteria. This pipeline was consolidated into a singular script, which enables parameters tested in this study to be easily modified by command line arguments, facilitating the usage for future users. Further, this pipeline has been containerized with docker so that specific versions of tools and dependencies are pre-installed, making the results reproducible and the pipeline compatible across computers. Finally, realistic artificial datasets simulating the mouse gut microbiota were created to optimise and validate the pipeline. With our validated parameters, the pipeline produced a Metagenome-Associated Genome (MAG) with 80% completeness and <5% contamination for *C. scindens*, a model for 7-DH bacteria, at an abundance of 0.05%. These results indicate that for most of the abundance range at which 7-DH bacteria are expected, the pipeline presented here will likely be able to produce MAGs with a sufficient quality to better grasp the diversity of 7-DH bacteria and discover new genes related to BA metabolism.

Résumé : Les bactéries qui dégradent les acides biliaires, plus spécifiquement, celles qui réalisent la réaction nommée 7-déhydroxylation (7-DH), sont suspectées de jouer une part cruciale dans le rôle hormonal dans acides biliaire et dans l'interaction entre le microbiote intestinal et l'hôte. Cependant, leur faible abondance dans le microbiote intestinal les rend difficile à étudier. Durant ce projet, un pipeline bio-informatique pour l'analyse de données métagénomique de séquençage type « shotgun » fut développée. Le but était de maximiser sa sensibilité envers les espèces bactérienne peu abondantes, afin de permettre la découverte potentielle de nouvelles espèces bactériennes capable de réaliser la 7-DH. Le pipeline a été développé sous la forme d'un script bash qui peut être lancé en une seule ligne de commande. De plus, tous les paramètres testés dans l'étude peuvent être modifiés simplement via un argument en ligne de commande. Il est accompagné d'un container « Docker » qui contient tous les outils et modules nécessaires au bon fonctionnement du pipeline, rendant les résultats reproductibles et le pipeline compatible sur d'autres ordinateur. Pour finir, des jeux de données artificiels réaliste, simulant le microbiote intestinal murin, ont été créés pour optimiser et valider le pipeline. Avec les paramètres validés, cette étude démontre que des génomes issus de données métagénomiques (MAG) complets à 80% et avec un taux de contamination inférieur à 5% peuvent être créés pour *C. scindens*, un modèle de bactérie capable de réaliser la 7-DH, à une abondance de 0.05%. Ces résultats semblent indiquer que le pipeline pourra, pour la plupart des niveaux d'abondance auxquels les bactérie 7-DH sont attendues, produire des MAG d'une qualité suffisante pour mieux cerner la diversité des bactéries 7-DH et découvrir de nouveaux gènes liés au métabolisme des acides biliaires.

I. Introduction

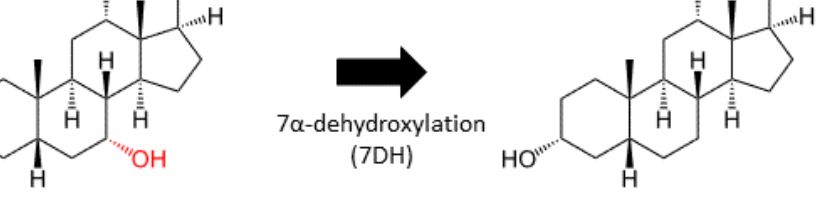
A. Background

1. Bile acids and the microbiota

A microbiota is the microbial community living in a specific niche, while the microbiome refers to the ensemble of genetic elements carried by the microbes in that community. In humans, the gut microbiota is increasingly recognized to be associated with host health. Notably, microbiota composition has been shown to effect obesity in human and mice. For example, obese individuals generally present with a lower microbiome diversity than lean individuals¹.

In addition to their well-known role in facilitation of lipid absorption, bile acids (BA, see [Glossary](#)) are signalling molecules involved in the regulation of glucose metabolism, lipid metabolism, and body weight⁴. These effects are mainly initiated by interactions with two receptors. The first is FXR, found in the liver, intestine, and pancreas. The second, TGR5, is present in the thyroid gland, intestine, and pancreas⁵, influencing energy expenditure in brown adipose tissue and skeletal muscle⁶.

A)

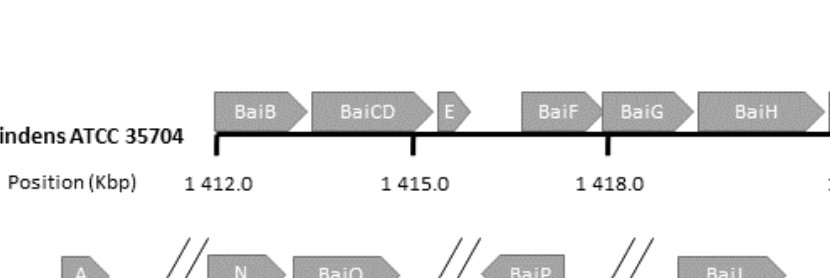


Cholic acid (CA)

7 α -dehydroxylation (7DH)

Deoxycholic acid (DCA)

B)



C. scindens ATCC 35704

Position (Kbp)

1 412.0 1 415.0 1 418.0 1 422.0

2 409.3 2 410.1 2 952.2 2 955.1 3 223.2 3 224.5 3 383.6 3 385.3

A N BaiO BaiP BaiJ

5

A recent *in silico* study indicates there is important inter-individual variation in the secondary BA production potential of the gut microbiome¹⁰. It also shows that bile-acid transformations depend on a community-wide metabolic network. Indeed, none of the ~700 analysed bacteria can produce all of the secondary BA alone, suggesting that an in-depth knowledge of 7-DH bacteria diversity will be required to fully understand these interactions.

So far, genes from the *bai* operon were found in a limited number of bacterial species. Most of them are part of the Clostridiales order¹¹, in the families *Clostridium*, *Blautia* and *Lachnospirillum*¹⁰. However, the diversity of 7-DH bacteria could be currently underestimated due to their very low abundance in the gut microbiota. Indeed, known 7-DH bacteria only compose about 0.0001% of the human microbiota¹² and between 0.4 and 0.01% of the mouse microbiota, with some variation depending on the diet¹³. Consequently, they are very challenging to study because the sequencing depth must be substantially increased to provide sufficient coverage of the bacteria of interest (see section II.A.1). In turn, this increases the cost of the experiment and creates large datasets requiring specialized analysis pipelines to analyse the data within reasonable time and computational resources. Developing and validating such a pipeline is precisely the aim of this master thesis. Therefore, the next section will discuss in depth the different methods available to study the microbiota as well as their limitations, especially regarding the detection of low-abundance species.

2. Studying the microbiota: technical evolution

a) 16S rRNA amplicon sequencing

16S rRNA amplicon sequencing leverages the property of the 16S ribosomal DNA subunit that it has variable regions interspaced by regions conserved in most of the bacteria¹⁴. This technique consists of extracting all the DNA in a biological sample (e.g., soil or gut microbiota), amplifying one or more 16S rRNA variable regions by PCR using primers targeting the conserved flanking regions and sequencing the PCR products (Figure 2). This allows an estimation of the taxonomic composition of the community and to measure the relative abundance of each taxa¹⁵.

While this technique is cultivation-independent, and therefore avoids the selection bias associated with the traditional approach to bacterial study¹⁶ (i.e., cultivation and isolation with subsequent individual sequencing), it has its own set of limitations. First, relative abundances tend to be biased as a result of differences between species in 16S gene copy number¹⁷ and in PCR amplification efficiency¹⁸. Secondly, the short length and the globally high conservation of the 16S rRNA does not always allow for the differentiation between closely related taxa¹⁹. This problem is amplified by a clustering step, where sequences diverging by only a few nucleotides are merged. This step is necessary to reduce the noise due to sequencing errors but it also hides potential variants and merges closely related strain or species²⁰. The choice of primers also influences the capacity to taxonomically assign the sequence as no singular region can distinguish all bacteria²¹. Finally, the taxonomy of the sequences can only be assigned by matching them with existing references. The interpretation of the results is therefore limited by database completeness.

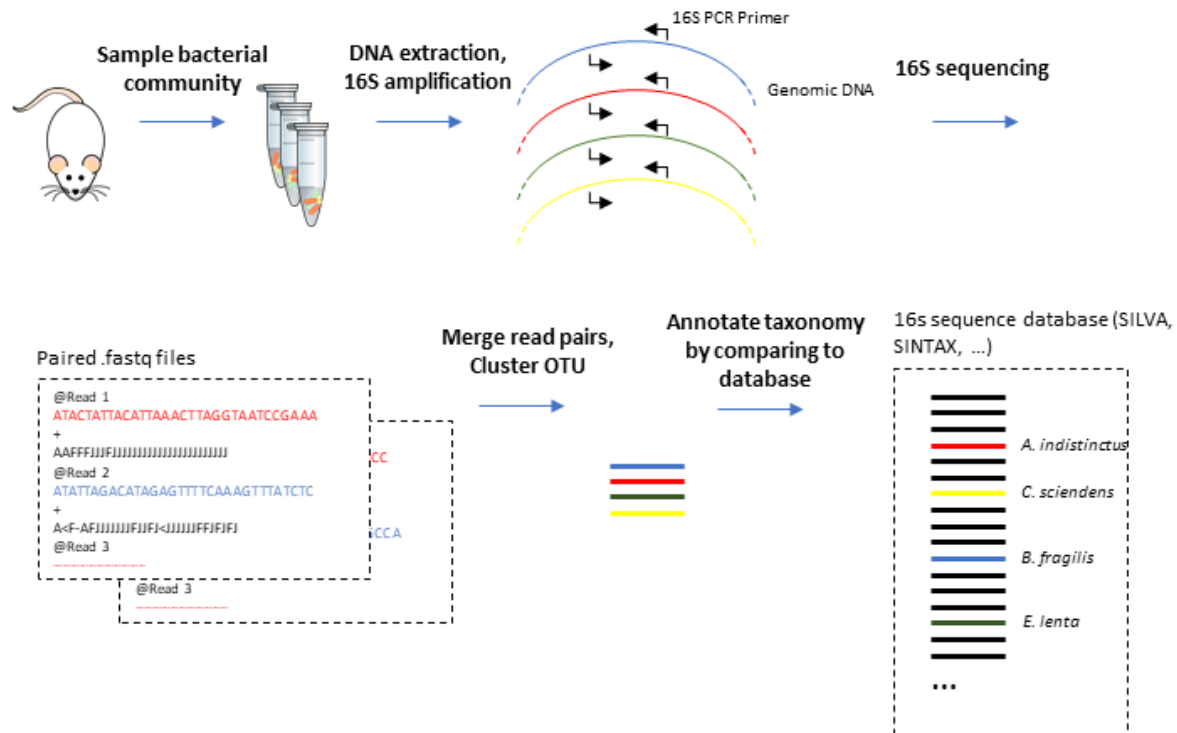


Figure 2: Schematic procedure of a 16S sequencing experiment to measure the composition of a bacterial community. The 16S rRNA-encoding gene is amplified using specific PCR primers and the resulting fragments are sequenced. Then the overlapping read pairs are merged. Very similar sequences are clustered in Operational Taxonomic Units (OTU) to prevent the added complexity of sequencing errors. The relative abundance of each OTU is estimated with the number of reads mapping to it. Finally, the resulting OTU are taxonomically annotated by finding their closest match in a database.

b) metagenome shotgun

During the last decade, the improvement in quality, throughput, and cost of DNA sequencing technologies has greatly improved. Methods such as Illumina sequencing can produce hundreds of millions of short reads (75 to 200 bp) in a single run. This facilitated the development of a revolutionary method to analyse microbial communities: Whole metagenome shotgun sequencing (WMGS). It consists of 1) extracting DNA from a biological sample (e.g. soil, water, gut microbiota) and 2) sequencing it without any cultivation or amplification steps. Then, 3) the short reads can be *de novo* assembled into larger DNA fragments called **contigs**. Finally, 4) contigs originating from the same species are clustered into bins called metagenome-assembled genomes (MAGs) in a step called binning²². This last step is important because genomes are almost never fully assembled into a single contig. Ideally, this entire workflow yields a set of MAGs that closely match the individual genomes of each species in the community.

De novo metagenomic assembly is more informative than 16S rRNA amplicon sequencing because it provides information on the entire genome of the bacteria in the community. Open Reading Frames (ORF), sequences that begin with a start codon end with a stop codon, can be searched in the draft genomes and be functionally annotated²³. Subsequently, nutrient requirements, antibiotic resistances, sporulation capacities or other traits of specific bacteria in the community can be predicted. That information can, for example, be used to design better-adapted cultivation mediums²⁴. Finally, the resulting genomes can help design gene-specific PCR primers to study different variants in the community²⁵.

However, these advantages come at the price of significantly more challenging downstream bioinformatical analysis. For each step, open-source tools designed for WMGS datasets were released

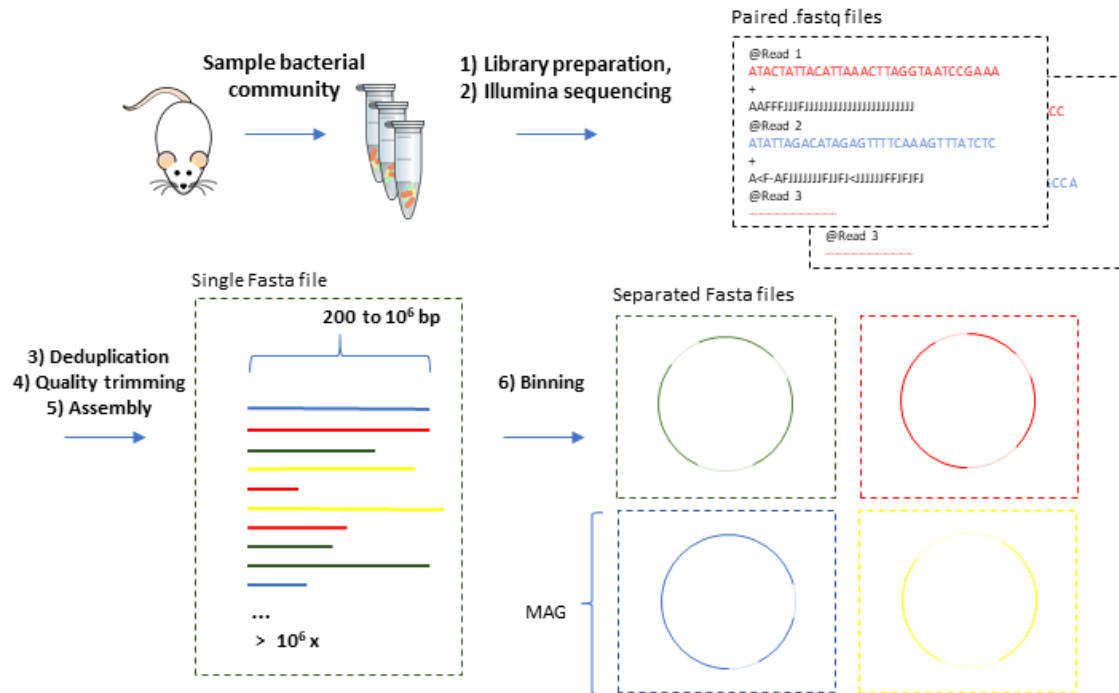


Figure 3: Schematic procedure of a WMGS experiment followed by an assembly-based analysis. The DNA is fragmented and sequenced without any amplification step. Then the short reads are assembled in contigs with a length ranging from hundreds to millions of base-pairs. Finally, the binning step clusters the contigs according to their predicated species of origin. This process produces MAGs, which can be further analysed in a similar way to typical draft genomes.

in the last decade^{26–28}, which facilitate this analysis process. Also, benchmarking studies were performed to compare their efficacy on synthetic²² and experimental datasets²⁹. However, choosing the appropriate tools, and fine-tuning their parameters to account for the constraints imposed by the microbial community of interest remains very challenging.

In the next sections, the steps and tools associated with our whole metagenomic sequencing pipeline are discussed, from the library preparation to the MAG binning. For the bioinformatic components, additional information such as the specific implementation of the algorithms used in the final design of the pipeline are available in Methods.

(1) Library preparation

Before sequencing, the target DNA must be prepared in a platform-specific fashion. For Illumina sequencing, library preparation includes 1) cutting DNA into small fragments of about 500 bp, and 2) attaching small oligonucleotides, called adapters, to the extremities of the fragment using PCR. These adapters allow the DNA sequence to bind to the flow cell, where the sequencing process happens on Illumina platforms. A flow cell is a surface covered with oligonucleotides (short sequence of DNA >20bp) complementary to the adapter's sequence and thus bind DNA to its surface.

Unfortunately, the PCR step induces bias in the output by overrepresenting sequences with medium GC content and artificially duplicating some DNA fragments. More recent library preparation techniques were developed to attach adapters without this PCR step³⁰. PCR-free techniques tend to be less biased toward medium GC content genomes and the exhibited duplication rate is often very low (0.04%)³¹.

(2) Illumina sequencing

The prepared DNA fragments are loaded on the flow cell and amplified, creating fixed clusters of identical sequences. Sequencing is performed by incorporating modified nucleotides one by one to these DNA fragments. The nucleotides release fluorescent particles when incorporated. This is monitored by a camera and the nucleotides incorporated in each cluster of fragments on the flow cell can be identified based on the colour of the fluorescence. The sequencing can be performed from both sides successively, creating 150 bp reads in pair, separated by a known distance of about 200 bp.

(3) Deduplication

It is common to find duplicated reads in sequencing datasets, mostly due to the library preparation process³⁰. These duplicates increase computational cost of the analysis due to data redundancy. Also, they skew the sequencing depth of the duplicated region, which is an important metric during assembly and contig binning.

As a result of a PCR-free library preparation process, low duplication levels ($< 0.04\%^{31}$) are expected in our experimental datasets. Given the exponential time requirement for pre-assembly duplicate removal leading to unreasonable computing time, this step was replaced by a quality control procedure: after assembly, duplication levels are measured to ensure they do not exceed the expected level. Should they exceed this level, leveraging the assembled contigs as a reference genome allows for the usage of more efficient duplication detection algorithms (see section II.F).

(4) Quality trimming

DNA sequencers use statistical models to estimate the probability that nucleotides were wrongly identified. These statistical models were trained by comparing the relationship between the signal measured by the sequencer to the accuracy of the nucleotide prediction in large datasets of reads with known accuracy. This data is output in the form of a quality-score associated with each base sequenced.

Each sequencing platform imparts specific artefacts on its output. For Illumina sequencers, common issues regarding quality include: 1) lower quality score at the extremities of the reads, indicating an increased probability of sequencing errors, and 2) reads containing components of the adapter sequence which binds the DNA fragments to the flow cell.

The trimming process consists of finding an appropriate trade-off between removing low quality bases that are problematic for the assembly and retaining as much of the reliable sequencing data as possible.

(5) Assembly

Assembly is the process of constructing long continuous sequences of DNA (i.e., contigs) using the short fragments produced by the sequencer. Intuitively, two reads can be assembled if the 3' end of the first overlaps significantly with the 5' end of the second. Two main algorithms exist to perform this process: overlap layout consensus and the K-mer strategy (see section II.H). Note that it was shown that the K-mer approach was especially useful for sequencing data with relatively short reads, such as Illumina. The overlap layout consensus on the other hand, is better suited for longer read technologies³².

Based on recent benchmarking studies^{22,29}, Megahit, an assembler using the K-mer strategy, appeared to be the best candidate for the task at hand. Especially since it uses a more efficient representation of the De Bruijn graph, called succinct De Bruijn Graph (sDBG), to reduce computation time and the memory required for the analysis²⁶.

(6) Binning

There are two main lines of evidence that can be used to cluster the contigs according to their predicted species. Some programs use one³³ and some both²⁸.

- Sequence composition: Bacterial genomes present species-specific patterns in their K-mer frequency. This information can be used to predict if two contigs likely originate from the same bacterial species.
- Read coverage: Due to experimental variation, changes in relative abundance of each bacterial species between samples are expected. Also, these changes should similarly affect the coverage of all contigs originating from the same species. Indeed, if the relative abundance of a species is lower in one sample, it will contain fewer reads from that species, resulting in a simultaneous drop in coverage for all contigs originating from that species. In other words, the coverage of contigs originating from the same species is expected to covariate between sample.

c) Existing analysis pipelines

Notably, complete pipelines for assembly-based analysis of WMGS datasets already exist. However, none of them is precisely appropriate for the *low-level abundances* expected in this experiment. EBI metagenomics³⁴, MetAMOS³⁵, and MOCAT³⁶ do not perform binning. InteMap³⁷ does not use one of the most recently published assemblers which is also one of the best performing in benchmarking studies: Megahit. The best existing pipeline could arguably be IMP³⁸, as it can perform all of the steps described above. However, it was specifically designed for co-analysis of metagenomic and meta-transcriptomic datasets, and thus may produce inferior results given that meta-transcriptomic data are not available in our study. Additionally, the binning program CONCOCT²⁸ used in our pipeline was shown to have better recall^{22,39} than the binning algorithm employed by IMP, MAXbin 2.0³⁹. Given the specific challenge of seeking to identify low abundance bacterial in our dataset, a higher recall is likely more appropriate. Nevertheless, IMP may make for an interesting comparison against the pipeline validated in this study as future work.

B. Motivation and goals

This thesis takes place in the context of two larger projects in collaboration with the Laboratory of Environmental Microbiology of Pr. Rizlan Bernier-Latmani at the École polytechnique fédérale de Lausanne (EPFL). Both projects aim to better elucidate the interaction between the host, the metagenome, and secondary BA production in mammals. The first, in humans, aims at better understanding the effect of bariatric surgery on BA's. The second, in mice, explores the contribution of host genetics and diet on the gut microbiota composition and its secondary bile acid production. During these projects, WMGS datasets of (a) human and (b) mice gut microbiota are going to be processed to create MAGs. These MAGs will then be screened for bile acid metabolism-related genes to find novel putative 7-DH bacteria species.

The aim of this work was thus to:

- 1) Create a realistic synthetic dataset which mimics the gut microbiota community of mice and test different metagenomic assembly pipeline parameters.
- 2) Develop the WMGS bioinformatic pipeline needed for these studies. More specifically, select tools for the pre-processing, the assembly of the reads, and the binning of the contigs. The main challenge, and differentiating factor from previously developed pipelines, is the assembly and binning of the 7-DH bacteria despite their very low abundances (0.0001% in human microbiota¹²; 0.4 - 0.01% in mouse microbiota¹³). Therefore, an important aspect of this project is tuning each tool's parameters

(optimisation) and measuring the final performance of the pipeline on artificial datasets to verify the anticipated sensitivity is sufficient to observe the targeted bacteria with 0.01% abundance. Abundance levels corresponding to the mouse gut microbiota were selected for this study as the associated real-world data will be available first for the mouse model. Further validation will be required before applying the method to find 7-DH bacteria in the human gut microbiota, for which the abundance levels are estimated to be two orders of magnitude lower.

II. Methods

A. Experimental dataset

The pipeline is aimed at analysing WMGS data derived from the gut microbiota of mice with different genetic backgrounds (BXD34, BX40, BXD11, BXD63) and diets (**Chow** vs **High Fat**). These conditions result in 8 strain/diet combinations (n=72, 4 strains x 3 biological replicates x 3 technical replicates x 2 diets). For simplicity, going forward, strain/diet combinations will be referred to as “conditions”. In the final experiment, the library preparation will employ the Illumina TruSeq DNA PCR-Free kit. This greatly reduces the number of duplicate reads (< 0.04%)³¹ and facilitates the downstream analysis (see section II.F). The DNA will then be sequenced on an Illumina Xten sequencer, creating paired-ends reads of 150 bp.

1. Coverage calculation and assumptions.

For a single genome, overlapping reads across the entirety of its length are required for assembly. This property is measured by the sequencing coverage. Assuming a uniform coverage, if the number of bases sequenced equals the size of the genome of interest, the coverage would be of 1x on average. To calculate coverage in a metagenome, one must also consider the abundance of each species in the community:

$$\text{Equation 1)} \quad \text{Coverage} = \frac{\# \text{ nucleotides sequenced} \times \text{Abundance}}{\text{Genome size}}$$

Where “Abundance” is the relative abundance of the species of interest, i.e., the number of cells of the species divided by total number of bacterial cells.

It is generally recommended to have a 20x coverage for reconstruction of individual genomes in a metagenomic dataset⁴⁰. However, due to the cost of sequencing it is not currently feasible to generate the amount of data that would yield such coverage for species with an abundance as low as 0.01% (coverage 2.84x). Indeed, the experimental dataset will be created using 12 lanes of an Illumina sequencer therefore sequencing 100Gbp per condition, which is notably a 20 fold increase compared to previous datasets⁴¹. Assuming an average genome size of 3.5 Mbp and that all reads come from the microbiome and not the host, the minimal abundance at which bacteria will have 20x coverage is 0.07% (Relative abundance=0.0007). This is still in the low-end of the range at which we expect to see the 7-DH bacteria in the mouse gut microbiota, and we are therefore confident that with a properly designed pipeline we will be able to reconstruct MAG of good quality for bacteria at these levels of abundance.

B. Artificial dataset

Realistic synthetic data mimicking the mouse gut microbiota were created to optimise and validate the pipeline. The datasets were created by sampling reads in various proportions from reference genomes. The taxonomic composition of the synthetic dataset is based on abundance measures from a study by Xiao et al, 2015⁴¹. This article describes the analysis of 180 WMGS datasets for mice of various strains.

Each dataset was assembled into contigs and ORF positions were predicted and taxonomically assigned. This resulted in 2.6 million ORFs with an average length of 762 bp and an associated taxonomic annotation. They were then clustered into co-abundance gene groups (CAG) based on cross-sample coverage covariance using the canopy algorithm⁴². Among the CAG found, 541 had a high quality: with >700 genes, a high correlation in their abundance profile, and consistent taxonomical annotation. These CAG were therefore considered as “Metagenomic species” (MGS)^a. Note that MGS are different from MAGs as they only contain predicted ORF and not the full contigs.

The 541 MGS were assigned to 96 different taxa at the species level. Relative abundance measures of the MGS with identical species taxonomy were merged. The abundance of the MGS was averaged across five selected mice (C57/BL6, high-fat diet), resulting in 66 non-zero MGS abundance values.

To generate our dataset, for each of the 66 taxa, a reference genome was retrieved from the NCBI assembly database^{43,44}. In this database, assemblies are assigned 4 quality-levels: “Reference”, “Representative”, “Complete” and “Incomplete”. For each unique MGS, the best available assembly was downloaded. If no “Reference” or “Representative” genome was available, the complete genome with the highest N50 (see section II.D.3 for definition of N50) was chosen. This process was automated with the python script “genome_download2.py” (see **Scripts**).

To measure the sensitivity of the pipeline toward low abundance species, 2 artificial datasets were created, with decreasing abundance level of *C. scindens* ATCC 35704, a known bile-acid transforming bacterium. To do so, the abundance of *C. scindens* was manually modified to respectively 0.0005 and 0.0001. New abundances for the other bacteria in the dataset were calculated by:

$$\text{Equation 2) } \text{New abundance} = (1 - C. \textit{scindens} \text{ new abundance level}) * \frac{\text{old abundance level}}{1 - C. \textit{scindens} \text{ old abundance level}}$$

The abundance calculations were performed with Microsoft Excel and the results can be found in the supplemental file “Abundance_profiles.xls”^b.

To be in concordance with the anticipated real-world experimental dataset, the artificial dataset will be 100Gbp large and have 9 replicates per condition. To simulate the experimental variation in abundance between replicates, a slightly modified abundance vector was used for each replicate. This abundance vector was created by adding gaussian noise constrained between +10% and -10% of the original abundance level to each element of the base abundance vector. If the abundance modified abundance vector does not sum to one, a normalisation step is automatically performed by the sampling function used.

The sample() function of the “R” software was subsequently used to calculate the number of reads to generate per reference genome and per sample. Briefly, for each replicate, the target number of read (100Gbp = 37*10⁶ reads x 2 x 150 bp x 9 replicates) was sampled with replacement from a vector containing the reference genomes names. The abundance vector with added noise for that replicate was used as the probability for each read to be sampled from a given genome. This sampling method ensures a fixed number of reads per sample and realistic abundances profile in the whole dataset. This process was programmed in the script “Xiao_samplig.R” and the creation of multiple datasets was automated with the bash script “Xiao_based_script2.bash” (see **Scripts**).

^a Abundance profile found at <http://gigadb.org/dataset/view/id/100114/token/mZlMYJf04LshpgP>, file mouse_CAGs_75Q_profiles.csv.gz

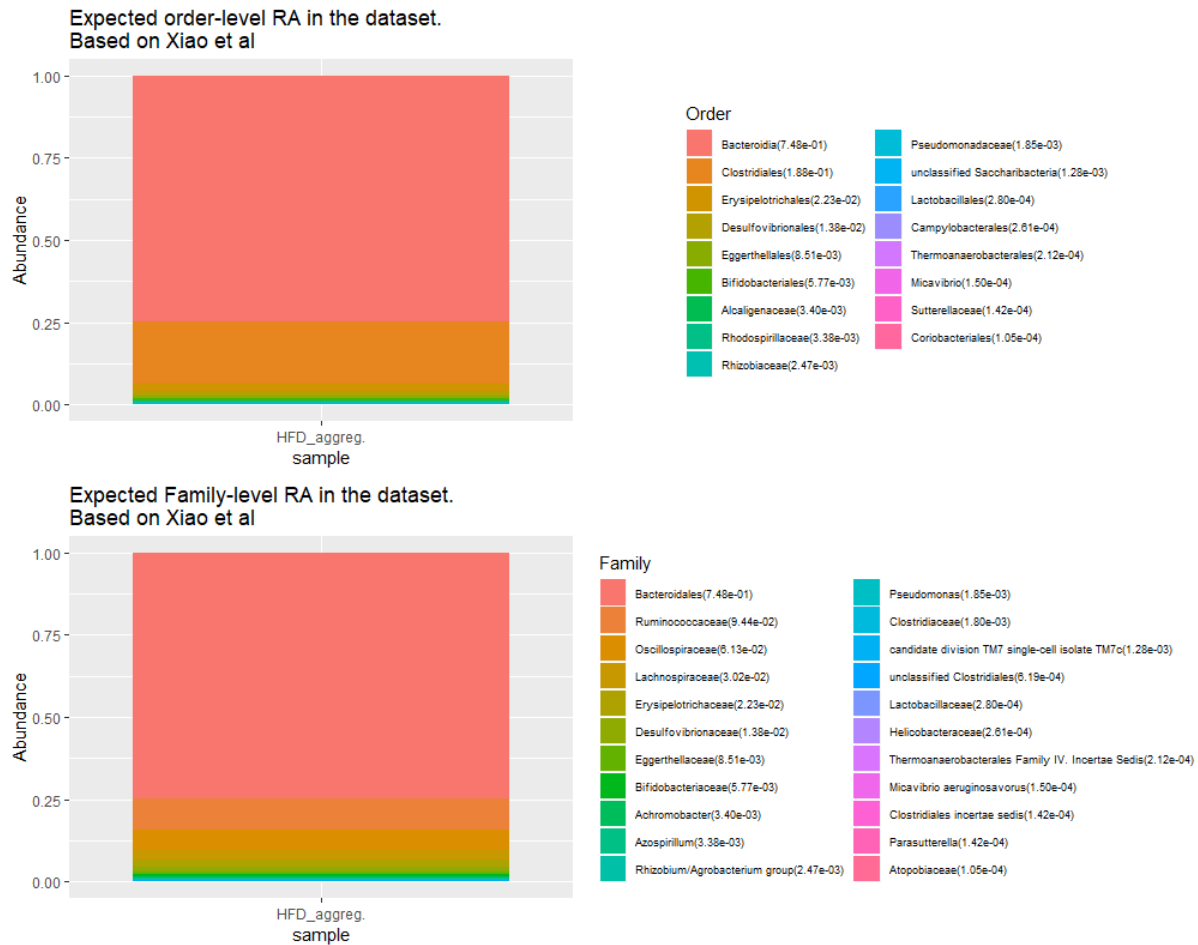


Figure 4: Order- and family-level taxonomic composition of the artificial datasets based on taxonomically assigned MGS ($n=541$) in Xiao et al, 2015. Relative abundance measures of the MGS with identical species taxonomy were merged, resulting in 96 unique taxa. For each of them, the average relative abundance in five mice of the same strain and diet (C57_{BL}, HFD) was calculated, resulting in 66>0 MGS.

Once the reference genomes and the number of reads to create for each replicate were chosen, ART Illumina⁴⁵ was employed to generate reads from each reference genome. This program creates artificial reads by sampling short sequences from a reference genome and outputs them in fastq/sam format. It also simulates the error and quality profiles of a variety of real Illumina sequencers. The error models are an especially interesting feature of this software and were created by comparing real-world experimental data to corresponding reference genomes and measuring the frequency of incorrect base calling and insertion-deletions. The error model employed here was HiSeqX PCR free (150bp) by specifying the parameter “-ss HSXn”.

The number of reads to create is specified with the “-c” parameter. However, art_illumina generates this number of read **for each contig** in the reference genome, which was not convenient for our goal of creating a given number of read **per reference genome**. Therefore, each reference genome was merged in a single contig to facilitate read creation. They were also stripped of any unknown bases, denoted by “N”. This modification would likely detrimentally impact processes such as aligning the artificial read back to the original reference genome or performing gene prediction and annotation on the contigs generated with the artificial datasets, by potentially causing reading frames shift. As no such process is planned on the artificial dataset, and as it greatly simplifies the analysis of the pipeline performance, merging contigs and removing the unknown bases was considered a reasonable trade-off.

The creation of the reads from each reference genome for each sample by art_illumina is automated in the script “Xiao_based_script4.sh” (see **Scripts**). Briefly, it goes through the .csv file outputted by “Xiao_sampling.R” and calls art_illumina to create the correct number of reads for the corresponding reference genome. Once the reads for each reference are created, the art_illumina fastq output are concatenated according to the sample number. The reads within the file are shuffled using the software “fastq-shuffle”^c to avoid any potential ordering bias.

C. Pipeline implementation and usage

The pipeline was developed as a bash script accompanied by a docker container where all the required tools and their dependencies were installed. The docker container is based on an Ubuntu 16.04 image.

The script is run from the command line and requires as input the path to a folder containing the

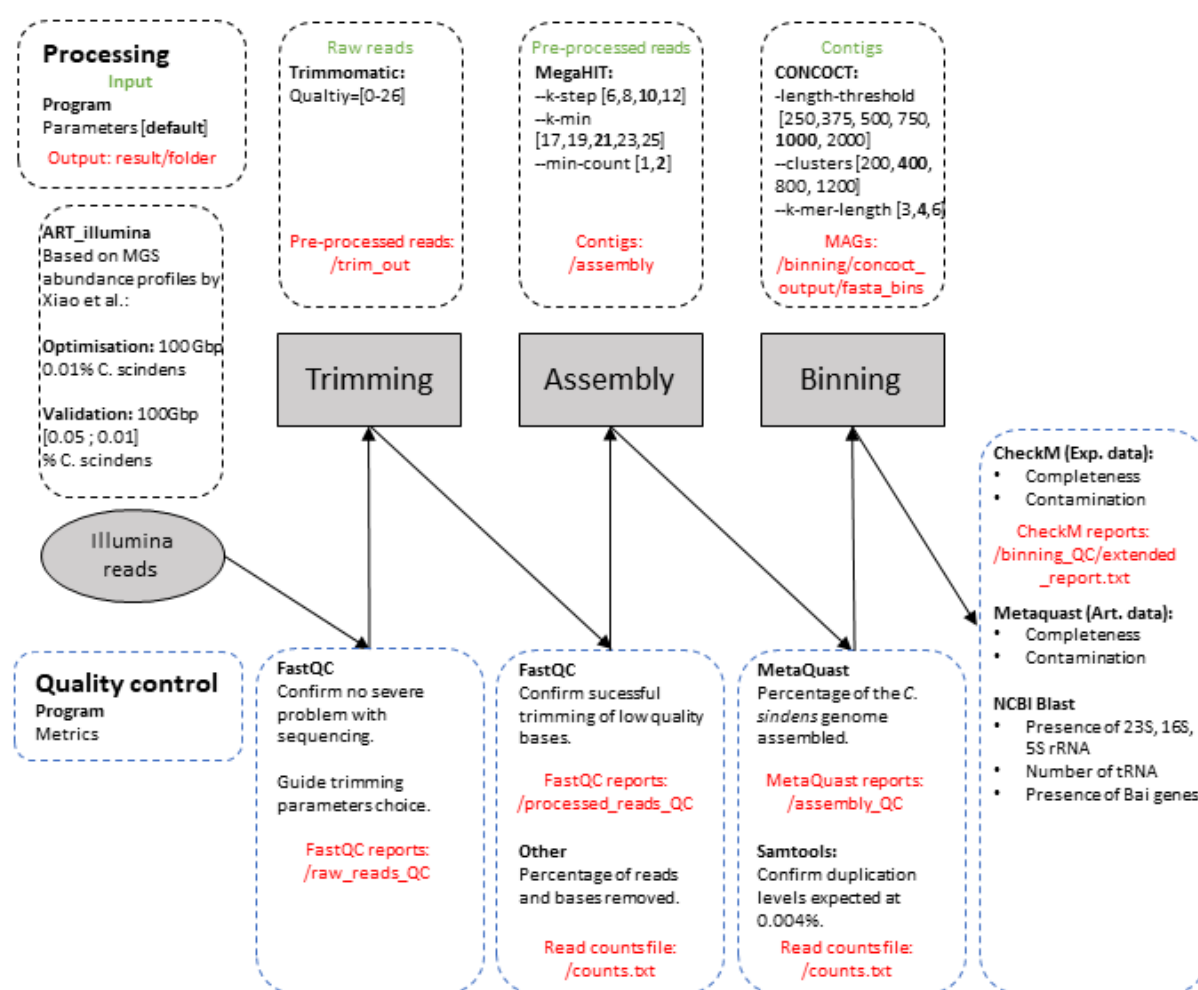


Figure 5: Flowchart presenting an overview of the programs and parameters used in the pipeline. The black dotted rectangles above each step highlights for each processing step, the input (green), the output (red) and the parameters tested during the optimisation process. The blue dotted rectangles below present the quality control step and the metrics computed.

sequencing data in fastq format. Additionally, the user can provide the path to a folder containing full

^c <https://github.com/chloroExtractorTeam/fastq-shuffle>

reference genomes, which improves the assembly quality control process. The default value of several tool parameters described below can be modified by passing a named argument to the script. Similarly, the user can specify the location of the output directory, the memory to allocate for the analysis, as well as the number of threads to use.

The pipeline outputs a result folder containing, among others, the trimmed reads, the assembled contigs, and the bins predicted by CONCOCT in fasta format. It also contains multiple quality control files, such as the read quality control plots produced by FastQC⁴⁶ before and after trimming, the assembly quality control report from MetaQUAST²⁷, and the output of the binning quality control tool CheckM⁴⁷. It also records the number of reads before/after pre-processing, the number of reads mapping to the assembly, the number of read duplicates, and the time consumed by each step of the analysis. Importantly, the script makes use of parallelization⁴⁸ to make the analysis more efficient.

D. Performance assessment

At multiple points in the pipeline, the quality of the results is monitored (**Figure 5**). This section will describe different quality-control (QC) metrics relevant to this project, how they are computed, and what aspect of the performance they are measuring.

1. Raw reads QC

The first quality control step measures the quality of the input data using FastQC⁴⁶ (**Figure 5**). It produces an html report containing several plots that facilitate the identification of issues imparted during the sequencing process. For example, one plot illustrates the quality score of each base against the position in the read (**Figure 6**). Another plot demonstrates the distribution of the average quality score among reads. Both can help in choosing appropriate trimming parameters. The other plots are mainly aimed at detecting issues with library preparation or the sequencing process, for example by plotting the average quality score of the reads against the position on the flow cell.

2. Read trimming QC

A new FastQC report is produced after the quality trimming (**Figure 5**) to visualize the quality profile of the remaining bases. It helps ensure that the quality trimming process removed the expected low-quality bases and that no adapter sequence remain at read extremities.

3. Assembly QC

MetaQUAST²⁷ was chosen to assess the quality of the assemblies (**Figure 5**). It computes various reference-independent metrics, such as:

- **N50 and L50:** the first is the length of the largest contig that, summed with all the larger contigs, contains 50% of the nucleotides of the total assembly. The second is the minimal number of contigs required to cumulate 50% of the total assembly. They are surrogate measures for continuity in the assembly. The absolute numbers tend not to be informative, yet they enable a comparison between two assemblies wherein the assembly with the larger N50 and lower L50 is more continuous and therefore tend to be more informative. These values are at their optimum when each genome is assembled in a single contig.
- **Total length of the assembly:** the sum of all the contigs' lengths. It indicates the amount of assembled data. Ideally, it should be the sum of the length of the reference genomes.

These metrics will be particularly useful when the pipeline is applied to real-world experimental dataset, for which no reference genomes are available *a priori*. These measures could allow users to compare their experimental results to the ones on the artificial dataset described above. If results are

alike, it suggests that the assembly performed similarly: a bacterium with a given coverage will have an equivalent proportion of its genome assembled.

During the optimisation and the validation process on synthetic data, the best attainable performance is known: fully reconstructing each reference genome. If reference genomes are input through the “-r” argument, MetaQUAST can align the contigs to the reference genome and compute other useful metrics:

- **Genome fraction assembled (GFA):** What proportion of each reference genome was assembled by the pipeline. This metric ranges from 100% for each genome in the best case to 0% in the worst case. For each genome, the assembled proportion depends notably on the sequencing coverage of the genome (see **Figure 13**), which depends on the abundance of the bacteria in the community and the sequencing depth used (see section **I.A.1**).
- **NA50:** Corresponds to the N50 of contigs mapping to a given reference genome. It measures the continuity of the assembly for each reference genome in the dataset.
- **Number of mismatches per 100kbp:** Number of bases with a different nucleotide in the assembled contigs compared to the reference genomes, normalized per 100kbp.

Finally, if no reference genomes are available, MetaQUAST will try to find appropriate surrogates by screening the contigs for 16S rRNA genes by comparing them to the SILVA database⁴⁹ using BLASTn⁵⁰. The program then downloads a reference genome from the NCBI database for the 50 bacterial/archaeal species with the best hit and keep the ones for which more than 10% of the genome is represented in the contigs. These results are to be interpreted carefully as they are dependent on the presence of a closely related strain in the NCBI database. Unfortunately, these metrics cannot be meaningfully used for assessing the quality of previously unknown organisms, limiting its potential in discovery type contexts.

4. Binning QC

After binning (**Figure 5**), the resulting MAGs will be evaluated using three metrics:

- **Completeness:** Fraction of the reference genome present in the MAG, calculated by mapping the resulting contigs to the reference genome and calculating the ratio between assembled and non-assembled genomes.
- **Contamination:** Fraction of the MAG not mapping to the reference, contigs that should have been clustered in another bin.
- **Presence of 23S, 16S, 5S rRNA and number of tRNA present.** Ideally, each bin has the correct number of each of these genes. Unfortunately, assembly is especially difficult in these regions of the genome²⁹ as they tend to be highly conserved between species and strain. As a result, the assembler requires reads longer than the conserved regions to perform well in there. Notably, the sequencing technology employed in this study produces reads of 150 bp long, which may be too short to successfully assemble some of these complex regions.

Depending on the presence/absence of reference genomes, these metrics can be calculated by two methods. In de novo scenarios, CheckM⁴⁷ identifies lineage-specific marker genes in each contig and uses them to calculate if a MAG contains contig markers from multiple lineages (contamination) as well as determining what proportion of markers for the lineage are present (completeness).

If reference genomes are available, MetaQUAST can be employed to produce more reliable results. It treats each bin as an independent assembly, and maps the contigs it contains to the set of reference genomes. Then, it creates a report detailing for each bin, the number of contigs mapping to each reference, their lengths, the genome fraction it represents and reports any missassemblies. This

enables the quantification of completeness (fraction of the main reference genome contained in the bin) and contamination (total length of contigs mapping to other references) with more precision than solely employing marker genes.

Based on these metrics, the MIMAG standard⁵¹ will be applied to label the bin as high, medium or low quality, following the criteria displayed in Table 1.

Bin quality	Completeness	Contamination	Presence of rRNA/tRNA
High	>90%	<5%	23S, 16S, 5S rRNA and at least 18 tRNA
Medium	>50%	<10%	-
Low	<50%	>10%	-

Table 1: MIMAG standard quality label and associated criteria.

E. Optimisation:

As discussed in the introduction, tool selection and parameter optimisation are especially challenging steps in building a WMGS analysis pipeline.

Optimization implies testing multiple sets of parameters for each tool in the pipeline, and comparing the results to determine parameters that optimally performs for the specific challenge. Again, our challenge was to discover low-abundance 7-DH bacteria, as such the set of parameters were compared based on their capacity to successfully recover the genome of *C. scindens* ATCC 35704 at a relative abundance of 0.01%.

The naïve approach for identifying the optimal parameter set involves executing the pipeline once on each possible parameter combination. Given that each run takes between 30 to 50 hours on a 100 Gbp dataset, this would not be feasible given the 10⁵84 combination of parameter values requiring validation. Therefore, a supervised step-wise optimisation strategy was employed, wherein each component of the pipeline was optimized individually with the subsequent step employing the output from the previous step. This was feasible since each step has its own unique output for which the best possible performance is known *a priori*. Indeed, the read pre-processing and assembly can be optimized using the GFA of *C. scindens* as the metric to maximize. Also, the binning process on the best assembly was optimized in regard to the fraction of the *C. scindens* genome binned in a single MAG, while keeping the contamination <10%.

Method sections F. to I. detail the tools used at each step of the pipeline, their specific algorithm, which candidate parameters were tested during the optimisation phase, and why they were selected for optimisation.

F. Deduplication

Read deduplication algorithms can either employ a reference genome or not, thus creating two categories for consideration.

Without a reference genome, the most naïve solution involves comparing each read to the other reads to determine if they are sufficiently dissimilar for retention. Different approaches propose heuristics to make this process more efficient, but ultimately this category of approaches remains very memory- and time-consuming.

When reference genomes are available, a more efficient method can be applied. First, reads are aligned to the reference genomes. Then, the tool iterates through each reference and searches for multiple reads starting and ending at identical positions on the reference, implying they have the same

sequence. Finally, the tool either marks duplicates within the alignment file or entirely removes the duplicates, retaining a single copy. The improved computational efficiency of this approach on large datasets is due to the fact that the best alignment algorithms scale in linear time complexity as a function of query length⁵². Unfortunately, previously known reference genomes are required and are thus not applicable in the case of metagenome *de novo* assembly.

Earlier versions of our pipeline included a reference-free deduplication step, using HTS Super Deduper⁵³. Its heuristic approach involves only comparing short segments of each read, and retaining only a single read if the sequence matches multiple reads. This approach proved to be more efficient at removing duplicates than other pre-alignment deduplication methods, such as FastUniq or Fulcrum⁵³. In addition, for this type of whole metagenomic shotgun sequencing dataset, the probability that two reads start exactly at the same position in a full genome is extremely low. This is even more improbable for low-abundance species which coincide with a low coverage. However, even with this simplification, computation time scales exponentially with the number of reads. At the size of the experimental dataset, the estimated computation time of 50 hours was not reasonable, and an alternative approach had to be found.

Noting that due to the PCR-free library preparation process the expected duplication levels in the experimental dataset is anticipated to be very low ($< 0.04\%$ ³¹), a quality control step after the assembly using the contigs as a reference was included: the reads are mapped to the contigs using BWA-MEM⁵² (a necessary step for the binning algorithm later) and the duplication level measured with samtools⁵⁴, to ensure it is indeed at the expected level. In the unlikely scenario where the duplication level is higher than expected, the user can choose to deduplicate the input fastq files using the contigs as a reference using a tool such as samtools rmdup and restart the pipeline employing the deduplicated reads.

G. Read trimming

As described in the introduction, DNA sequencers provide a quality-score associated with each base sequenced. They do so using statistical models trained on large datasets of reads with known accuracy. Quality-scores are then calculated with the following formula:

$$\text{Equation 3)} \quad Q = -10 \log_{10} P$$

where Q is the PHRED quality score and P is the probability of a wrong base call estimated by the sequencer.

Read quality trimming consists of removing nucleotides with low quality scores as they have a higher risk to contain sequencing errors and thus represent a major challenge for the assembly (see section II.H.2). In the proposed pipeline, Trimmomatic was employed as it is able to remove both adapter sequences and perform quality trimming.

The read quality trimming was performed using the “sliding window” algorithm: the program slides a window of length L along the read in the 5' to 3' direction. The mean quality in the window is calculated for each window, and if it drops below a given score Q , the bases from the 5' end of the window to the 3' end of the read are discarded. The value for Q is user-specified. The parameter Q was extensively tested during optimisation process, with values between 26 and 0 and a fixed window size of $L=5$.

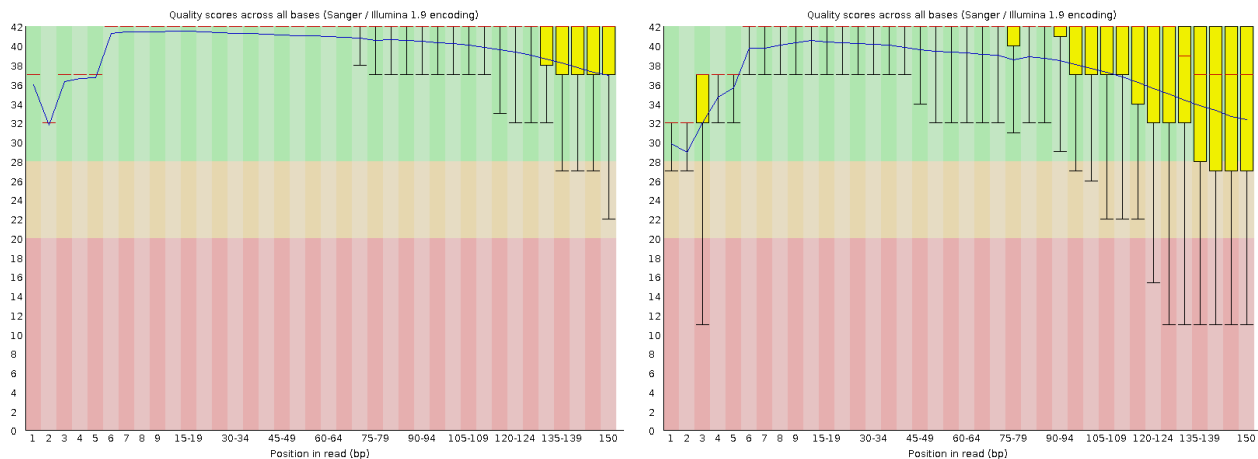


Figure 6: Example of base quality-score profiles of reads in the artificial dataset. Left panel is R1 and right panel is R2. Base 1, on the left of the x-axis, is the 5' end of the reads and base 150, on the right, is the 3' end. Plots created using FastQC. As expected with Illumina sequencer, a drop in base quality scores can be observed at the 3' end of the read, especially in R2.

H. Assembly

As mentioned in the introduction, there are two approaches to assemble the short reads into longer continuous fragments called contigs.

1. Overlap layout consensus

This assembly strategy is perhaps the most intuitive, it consists of comparing each read pair and detecting overlaps in nucleotide sequences. This information is used to generate a graph, called a layout, where each read is a node and overlaps between read pairs are edges. Finally, a consensus sequence can be derived by finding a Hamiltonian path in the graph. This corresponds to a path that goes through each node of the graph exactly once. However, finding such a path is known to be a NP hard problem³² and therefore this technique tends to be excessively memory intensive and is best suited to datasets of long reads with low coverage³².

An improved version of this strategy based on a structure called string-graphs can be used to assemble short-read WNGS datasets. Assemblers using this strategy are faster and more memory-efficient than overlap-consensus based ones⁵⁵. However, they still have prohibitively high memory requirement for very large datasets as compared to assembler using the K-mer strategy.

2. K-mer strategy

This algorithm consists of dividing each read into overlapping nucleotides strings of length K (called K-mers). Then, a directed graph is constructed where the nodes are K-mers and an edge is created between them if they are adjacent to each other in a read (overlap of K-1 bases). With the graph constructed, contigs can be formed by finding a path that goes through each edge exactly once. This is known as an Eulerian path problem and efficient algorithms exist to solve it⁵⁶.

Even with this efficient approach, K-mer strategies still have limitations²⁶. First, they are especially sensitive to sequencing errors, as each sequencing error will wrongly create K additional edges in the graph. To mitigate this problem, some assemblers start by counting each K-mer in the dataset and only those present more than a given threshold are used for the graph construction.

Secondly, if a DNA region is repeated, K-mers longer than the repeated region are needed to find the correct path. However, using a large value for K is less efficient in low-coverage regions, which in

metagenomes, is particularly problematic for low-abundance species. This challenge can be addressed by constructing the graph in multiple iterations with an increasing value of K and merging the new contigs with the results of the previous iteration. However, the length K is ultimately limited by the read length. If a repeated region is larger than the read size, it cannot be resolved and will result in separated contigs.

3. Implementation in the pipeline

The real-world experimental datasets will be composed of 9 replicates (3 biological and 3 technical). To optimize the assembly in the most similar conditions as possible, the same design was applied to the synthetic datasets. To maximize the coverage of low-abundance genomes, all replicates will be co-assembled. This means that they will all be treated as a single sample instead of being assembled separately.

The chosen assembler, Megahit²⁶, employs the K-mer strategy. To compensate for the sequencing error limitation mentioned above, Megahit begins by counting each K-mer in the dataset and only those present at least twice are used for graph construction (threshold modifiable, see `--min-count` below). Megahit mitigates the challenges associated with the choice of the K-mer size by constructing the graph in multiple iterations with an increasing value of K. At each iteration, in addition to the reads, Megahit uses the contig information created at the previous iteration to construct a new De Bruijn graph.

The following parameters were selected for optimization. The values to tested are between brackets and the **default value is in bold**:

- **--k-min**: The K-mer size used during the first iteration. With the K-mer strategy, for two reads to be assembled, they must overlap at least K bases. Therefore, setting a high K-min value will prevent reads with only minimal overlap from being assembled. This will be especially frequent in low-coverage regions and therefore problematic in low abundance genomes. However, the lower the K-min value is, the more complex the graph becomes, as shorter K-mer will tend to be repeated more often in the metagenome. The Megahit manual recommends setting higher values for large and/or complex datasets. For this experiment, the dataset will be large but a high sensitivity in the low-coverage region is also a strict requirement. Therefore, both values above and below the default will be evaluated. [17,19,**21**,23,25]
- **--k-step**: The step value by which the K-mer is increased at each iteration. The Megahit manual states that a smaller k-step is more appropriate for lower-coverage datasets. [6,8,**10**,12]
- **--min-count**: Before each assembly iteration, the K-mers with a count below this threshold will be considered sequencing errors and removed from subsequent consideration. As stated before, this filtering is aimed at avoiding sequencing errors, that are a major challenge for Megahit. However, this filtering is particularly harsh for low-coverage regions and risks hiding rare species variants. Therefore, lowering this threshold to one will also be considered during the optimization process. [1,**2**]

I. Binning

As stated previously, this step consists of clustering the contigs that are predicted to come from the same species into so-called MAGs. Based on benchmarking studies^{22,29}, CONCOCT²⁸ was selected. This program uses two metrics, calculated for each contig, to perform the clustering: (a) K-mer frequency metrics and (b) read coverage metrics.

K-mer frequency metric: It was shown that different species tend to present short patterns (K-mers) at different frequency in the genome³³. For each contig, CONCOCT measures the frequency of each

non-palindromic tetramer (K-mer of length 4, 136 possibilities) and store the results in a matrix of dimension $N_{\text{contigs}} \times N_{\text{K-mers}}$. The matrix is normalized to account for differences in contig length.

Read Coverage metric: Since contigs from the same genome are expected to covariate across different samples of the same metagenome, CONCOCT measures the coverage of each contig in each sample and stores it in a second matrix of dimension $N_{\text{contigs}} \times N_{\text{samples}}$. This value is normalized within and across samples and then log-transformed. Additionally, the total coverage across samples for each contig is also stored, as it may also help in distinguishing between different species bringing the final dimension to $N_{\text{contigs}} \times (N_{\text{samples}} + 1)$.

These two matrices containing one row per contig are merged, forming a matrix of dimension $N_{\text{contigs}} \times (N_{\text{samples}} + 1 + N_{\text{K-mers}})$. A principal component analysis (PCA) is performed on the matrix to reduce the dimensionality and only the dimensions needed to explain 90% of the variance are retained. Then the contigs in the reduced dataset are clustered using a Gaussian Mixture Model (GMM) fitted using a variational Bayesian approximation. CONCOCT starts with a large number of clusters, initialized by a K-means clustering algorithm and aims to identify a model which maximizes the likelihood of the data given the set of clusters by alternatively updating (a) the weight of dataset columns and (b) the parameters of the model.

The following parameters were selected for optimisation. The values to test are between brackets and the default value is in bold:

- **--length-threshold:** contigs shorter than this value will not be included. A high value greatly reduces the complexity of the dataset and therefore the computing time, which is crucial in this experiment given the sequencing depth selected. However, some information is also discarded as a result. There is thus an important trade-off to be specifically tuned for this experiment. [250, 375, 500, **1000**, 2000]
- **--clusters:** max number of clusters for the GMM fitting. It should be set at least two to three times the number of expected clusters/species, because the algorithm should start with more centroids than necessary and then iteratively discards the unsupported ones. [**400**, 800, 1200]
- **--kmer length:** The size of the K-mer used to calculate the composition vector. Using 6-mers is interesting biologically because it corresponds to two amino acid codons. [3,4,6]

J. Validation

During this step, the performance of the final pipeline was tested on a new synthetic dataset to check if the optimisation did not over-fit the parameters to the first one. The new dataset was generated with the same procedure as the first one except for the seeds used to initialize the pseudorandom processes (adding noise to species abundance vector, sampling the number of read per species, sampling the reads with art illumina). The GFA of *C. scindens* at an abundance of 0.01% recovered in a single MAG was measured. The GFA metric has a range from 0% at worse to 100% at best. Given the currently available tools and their reported metrics, a value of between 50% and 70% was deemed to be acceptable for a genome of this low abundance. This decision was motivated by Megahit being reported to assemble genomes with 16X and 32X coverage at 85% and 95%, in less complex datasets²² as well as accounting for the potential loss of small contigs during the binning process.

The pipeline was also evaluated on an additional dataset with an abundance level of 0.05% for *C. scindens* to better assess the minimal abundance threshold required for the bacteria of interest to be detected. The pipeline will be considered successful at a given abundance level if it can produce at least a medium quality MAG for *C. scindens*. This implies a MAG with >50% completeness and <10% contamination.

Aiming at a high-quality label would be unrealistic for such low abundance species, especially given the requirement of rRNA and tRNA presence (**Table 1**). Indeed, rRNA and tRNA are the most challenging regions for the assembly as they are conserved across bacteria and as the most frequent variants tend to hide the rarest ones. In addition, the presence of rRNA and tRNA in the MAG would not be informative regarding the bile-acid degradation potential of the MAG, which will be a key point of the downstream analysis. Rather, the ability to capture bile acid-transforming gene depends on the completeness of the MAGs, as more incomplete MAG will be more likely to lack the region encoding bile-acid degrading genes. Therefore, we will not aim for bins with high quality labels as defined by the MIMAG standard, but instead will aim for medium quality bins and focus on improving the completeness as much as possible ensuring the contamination level remains below 10%.

III. Results

A. Optimisation

This section describes, for each step of the pipeline, how the different parameter values tested impact the assembly and binning of low-abundance bacteria using as a model the genome of *C. scindens* at an abundance of 0.01%. Relevant general trends on bacterial community as a whole are also presented.

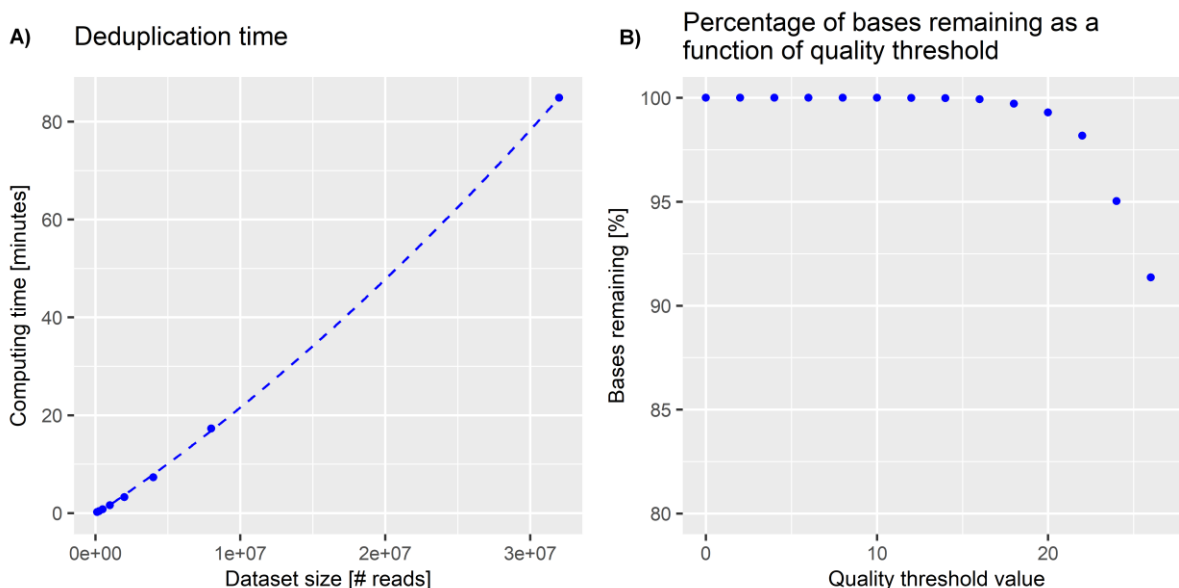


Figure 7: **A)** Estimation of the computing time needed for read deduplication. The computing time was measured with datasets of 125K to 32M reads. In R, a polynomial model of degree two was fit to the data and extrapolated to the final size (100Gbp or 333M reads). This resulted in an estimated computing time of 50.8 hours. Model: $y = 2.152e-14 * x^2 + 1.975e-06 * x - 35$. **B)** Percentage of the dataset remaining depending on the threshold selected for quality trimming.

1. Read deduplication

The read deduplication tool Super Deduper was selected for the initial implementation of the pipeline due to its demonstrated high throughput⁵³. Unfortunately, it was observed that its computation time scaled exponentially with the number of reads. As a result, the 100Gbp or 3.33M reads of the validation dataset would have taken approximately 50.8 hours (**Figure 7.A**), becoming a major bottleneck of the pipeline. Since duplication levels in the real-world experimental dataset are expected to be very low ($< 0.04\%$ ³⁰) due to the PCR-free library preparation process, this step was removed in favour of a quality control step after the assembly, using the alignment-based approach discussed in **Method F**. The reads are aligned to the assembly and the number of duplicates is measured using Samtools. Given that aligning the reads to the assembly is also a pre-requisite for CONCOCT, this quality control step only increases the computation time by about 30 minutes. In the unlikely scenario where duplication levels are higher than expected, the user can use Samtools to remove duplicated reads and restart the pipeline on the deduplicated fastq files.

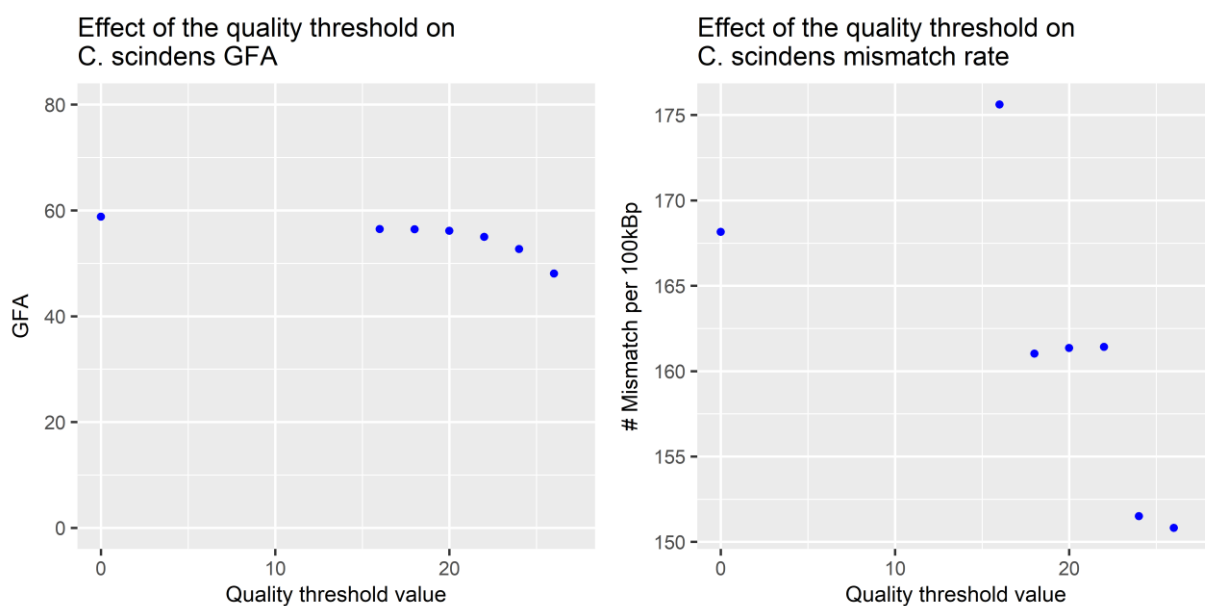


Figure 8: **A)** Genome fraction assembled for *C. scindens* at abundance 0.01% as a function of quality threshold. A lower quality threshold appears to be associated with an increased GFA. The GFA was measured with Metaquast by comparing the contigs to the reference genomes. **B)** Mismatches per 100kbp for *C. scindens* in function of the quality threshold. Mismatches were measured by Metaquast.

2. Read quality trimming

The first metric monitored during the optimisation of the quality threshold for the read trimming was the number of nucleotides discarded. Reducing the quality threshold allows for the keeping of more nucleotides for the downstream analysis (**Figure 7.B**) but increases the risk of sequencing errors in the datasets. The number of nucleotides appears to decrease especially rapidly at quality threshold higher than 20.

The second metric monitored was the GFA of *C. scindens* in after the assembly. A lower quality threshold appears to increase the GFA of *C. scindens* (**Figure 8.A**). This might appear counterintuitive as sequencing errors are known to be one of the weaknesses of assemblers using the K-mer strategy, but this increase is likely due to fewer bases being removed by the read trimming, leading to a higher coverage. Indeed, as shown in **Figure 13**, coverage is a very important factor explaining the GFA of a given genome.

However, decreasing the quality threshold also leads to a slight increase in the rate of mismatches in the contigs (**Figure 8.B**), as sequenced bases of poorer quality kept in the trimmed read dataset are more likely to contain sequencing errors.

Considering that the metric to optimize is the GFA for low abundance bacteria, a quality threshold of 0, meaning no quality trimming at all, was chosen for the validation. The associated increase in mismatch rate was estimated to be a reasonable trade-off. Notably, final users can also choose a different quality threshold using the command line argument, should their analysis prioritize fewer mismatches in the contigs at the cost of a lower GFA.

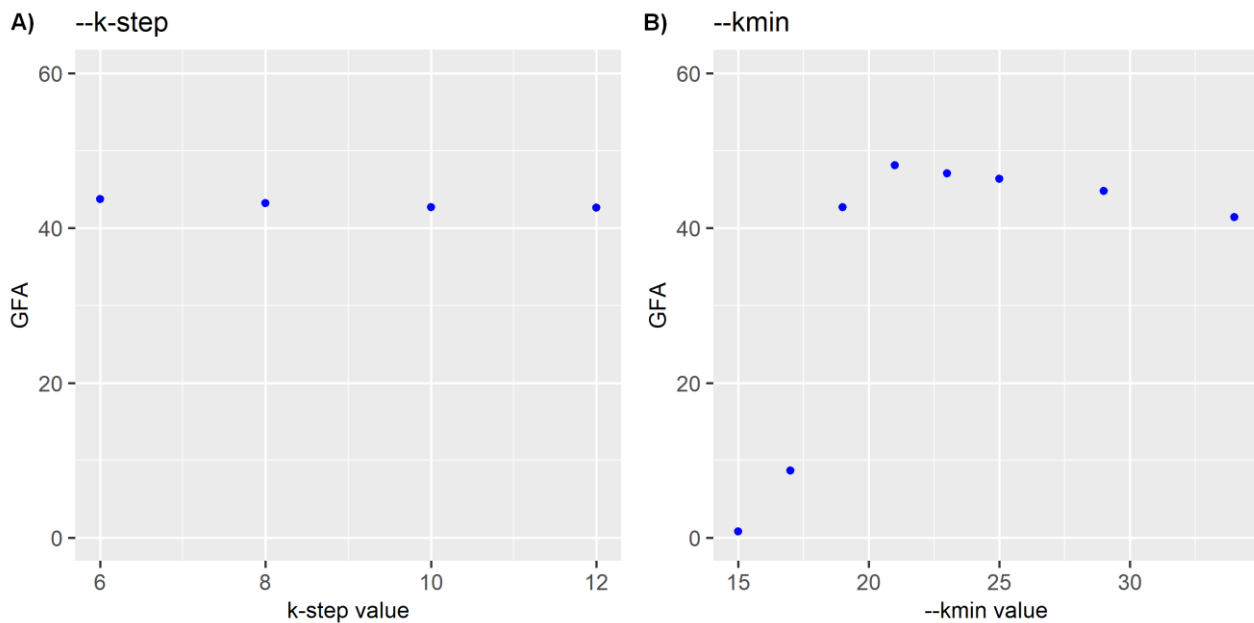


Figure 9: **A)** Effect of “k-step” on the *C. scindens* genome fraction assembled. 3 values were compared with otherwise identical parameters. The genome fraction assembled was measured by Metaquast. **B)** Same as A but showcasing the effect of “k-min” on *C. scindens* genome fraction assembled. The GFA peaks at 48% with a k-min size of 21.

3. Assembly

For the assembly step, the results from optimising the following three parameters of Megahit were as follows:

--k-step: controlling K-mer size increase between assembly iterations. The values 6, 8, 10 (default) and 12 were tested. As no effect on the assembly of *C. scindens* could be observed by changing this parameter (**Figure 9.A**), the default step of 10 was selected for validation.

--k-min: size of the K-mer for the first assembly iteration. Values ranging from 15 to 39 were tested. The GFA for *C. scindens* increased rapidly at first, reaching an optimum of 48.1% GFA at a K-mer size of 21 and then slowly decreased (**Figure 9.B**). The worst results at low minimal K-mer size are likely due to shorter K-mer having a higher probability to be present in multiple genomes, therefore creating graphs too complex for Megahit to assemble correctly. On the contrary, as the K-mer strategy will require the reads to overlap by at least K nucleotides to assemble them into contigs, increasing the K-mer size requires reads with increased overlap for Megahit to assemble them (see section II.H.2). As genomes with low coverage tend to have reads with less overlap, this could explain the less complete assemblies obtained for *C. scindens* with K-mer sizes higher than 21.

--min-count: by default, this parameter filters K-mers appearing less than two times. It aims at reducing impact of sequencing errors in the dataset, which tend to increase memory consumption and computing time. However, it will at the same time prevent the assembly of low-coverage region (below 2x) and rare variants in the community (see section II.H.2). Therefore, an assembly with a --min-count of 1 was attempted, but it resulted in an error because the memory allocated (100Gb) was not enough for the process to complete. Even if lowering the threshold might be valuable to assemble low-abundance bacteria, this cannot be done in our case due to the important size of the dataset.

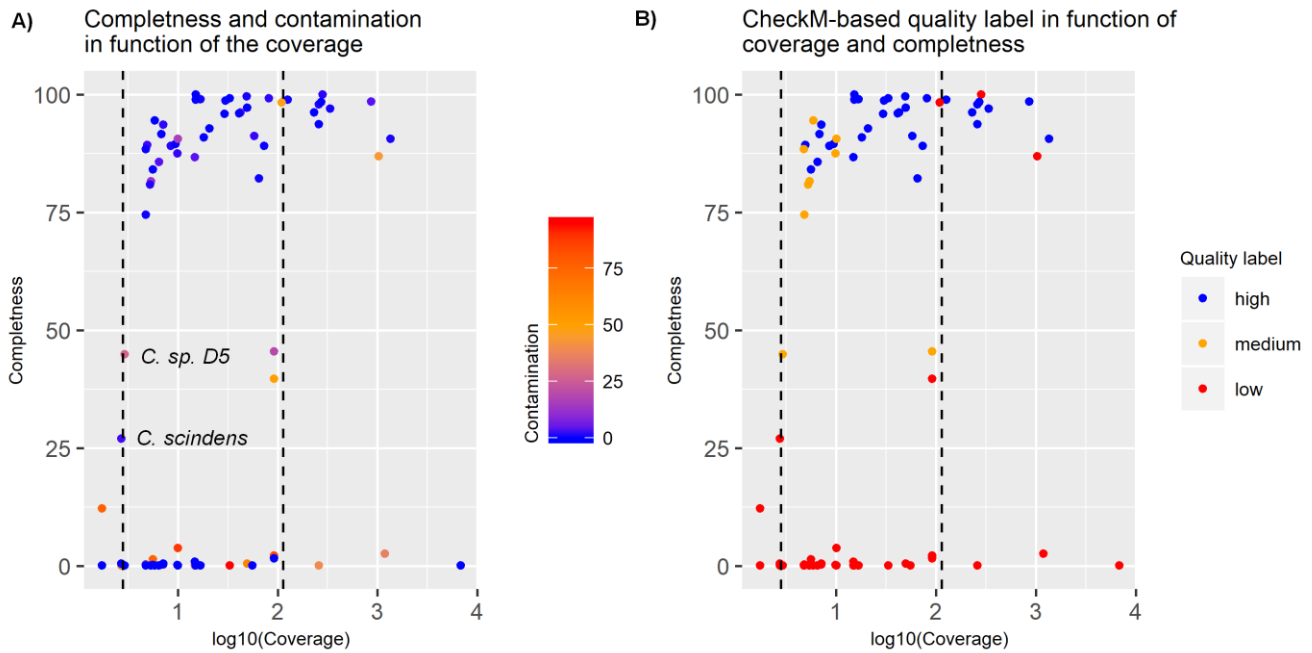


Figure 10: **A)** Completeness of each bin in function of the coverage of the main genome in the bin. The colour scale represents the contamination. *C. sp. D5* and *C. scindens* are labelled for comparison with figure 14. The dotted lines indicate the low and high ends of the coverage range expected for 7-DH bacteria. Completeness and contamination values were calculated by MetaQuast by comparison with the reference genomes. The coverage was measured with samtools. Binning parameters: length threshold 750; clusters 400. **B)** Presents the same data than A, but the colour represents the MIMIAG standard quality label for completeness and contamination. These labelled were assigned based on the result of Checkm, which does not require reference genomes.

4. Binning

The binning process globally reveals a strong relationship between coverage and completeness. Still, a non-negligible number of bins present with either low completeness or high contamination values, even at relatively high coverage (**Figure 10.A**, yellow points at the top). Fortunately, most of them can be identified as low-quality based on the quality control metrics measured by CheckM (**Figure 10.B**, same points marked in red). As this program does not require reference genomes to measure the completeness and contamination, it could help detect bins of lower quality even in the experimental dataset.

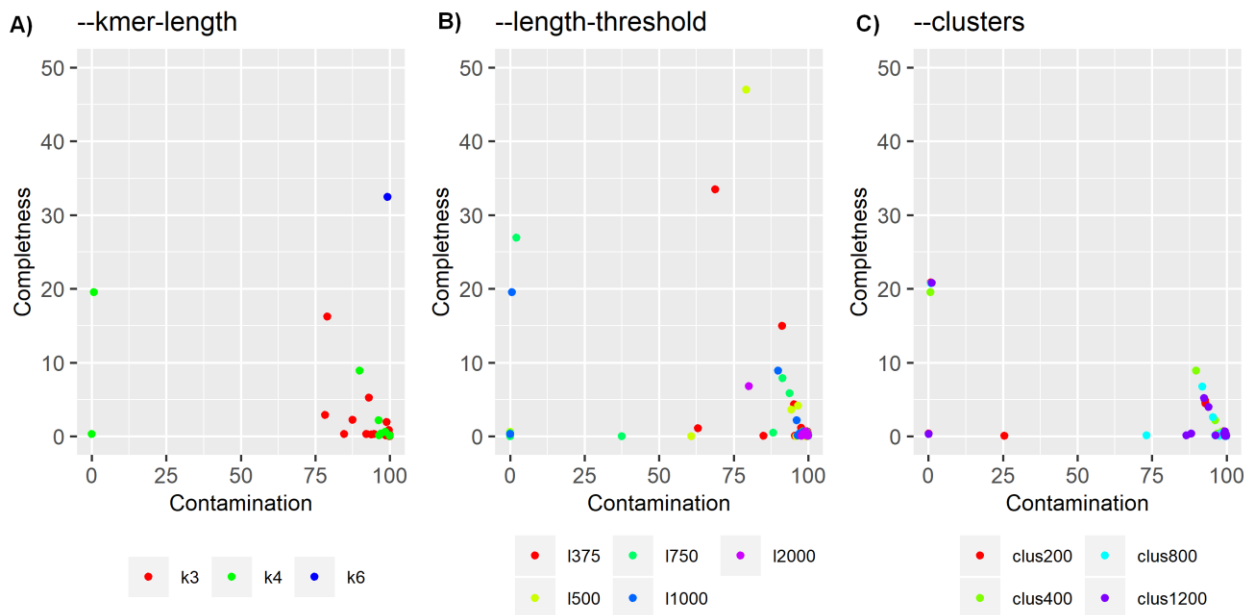


Figure 11: Only bins containing contigs from *C. scindens* are represented here. Bins with the higher quality will be found at the top left of each of these plots. **A)** Effect of different “--kmer_length” values on the completeness and contamination of all bins containing contigs mapping to *C. scindens*. The completeness represents the proportion of the *C. scindens* genome contained in the bin. The contamination represents the proportion of the bin that does not map to *C. scindens*. **B)** Same as A, but for the length threshold **C)** Same as A, but for the number of initial clusters.

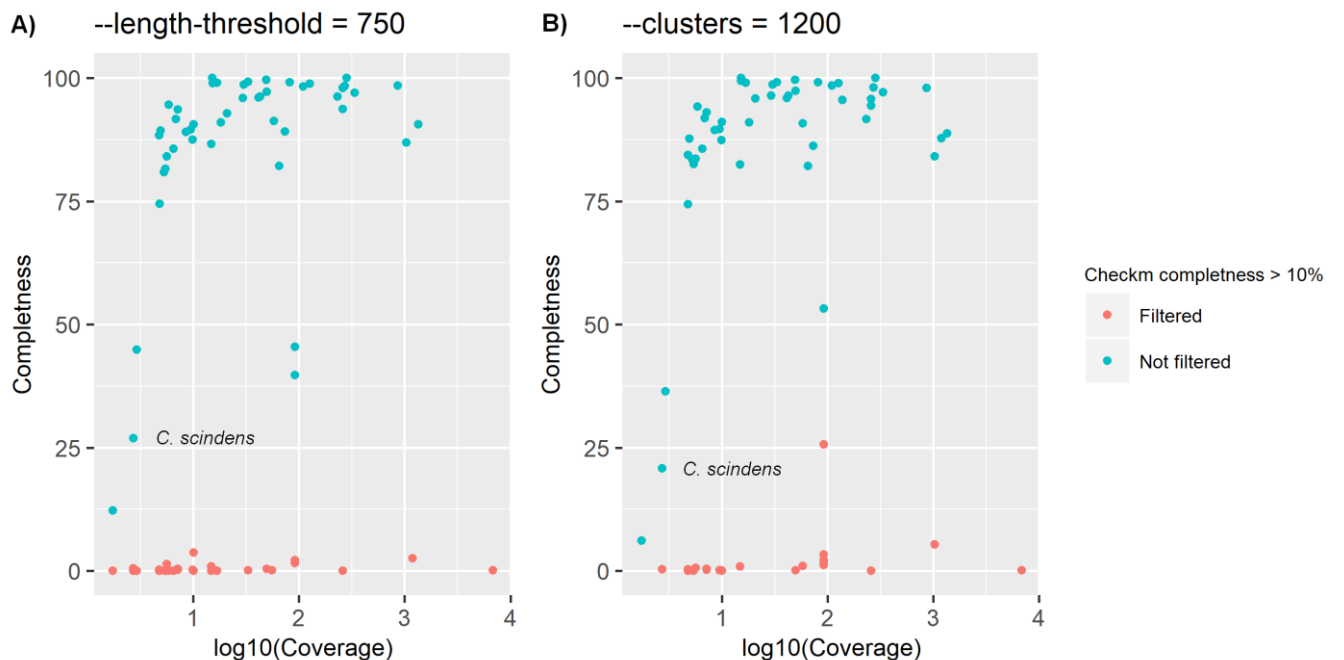


Figure 12: **A)** Highlights the high number of very low completeness bins created by using a length threshold of 750 and their efficient filtering by removing bins with CheckM completeness values <10 (in red). **B)** Same figure for the parameter “--cluster” = 1200.

For CONCOCT, the three following parameters were optimized:

--kmer_length: the length of the K-mers used for the K-mer frequency vector, which serves to cluster the contigs based on the sequence content. The best performance, more specifically the highest completeness possible for *C. scindens* while maintaining a contamination level below 10%, was obtained with the default value: 4. It resulted in a bin containing 19.5% of *C. scindens* genome, with only 0.6% contamination (**Figure 11.A**, green point on the left) and for the other species, 64 bins were created, 39 of which are of medium quality or higher (**Table 3**).

K-mer length = 3 results in fewer bin, both overall and of medium to high quality. Some bin contains contigs from *C. scindens*, but they have a lower completeness as compared to the default parameter and are all highly contaminated (**Figure 11.A**, red points on the right have contamination levels of 75% or higher).

Increasing the K-mer length to 6 results in only 12 bins in total, of which only 6 have a medium quality or higher (**Table 3**). One bin contains 30% of the genome of *C. scindens* but 80% of the bin is made of contigs from various other species, making it not useful in our particular use case (**Figure 11.A**, blue point at the top right).

--length_threshold: the minimal length threshold for CONCOCT to include the contigs into its analysis. By default, it is set at 1000 (in blue on **Figure 11.B**). Setting it to 750 creates a bin with higher completeness (26.9%) for *C. scindens*, while keeping the contamination well below 5% (light green point at the top left, on **Figure 11.B**). Indeed, lowering this threshold can be especially useful for low abundance bacteria, as they also tend to have shorter contigs, which might otherwise not even be considered by CONCOCT. On the other hand, it also creates redundant, almost empty bins for some species. Fortunately, this issue can be efficiently solved by filtering bins with a CheckM completeness score below 10%, as illustrated by the red points on **Figure 12.A**. Therefore, the value 750 was selected for the validation despite this minor drawback.

Setting the length threshold to even lower values leaves less contigs with no bin assigned but also tends to increase both the number of wrongly assigned contigs and the computation time of the analysis. When set to 250, the analysis was interrupted after 10 days without completion. Setting this parameter to 375 results in a very high number of bins with only 8 of medium to high quality (**Table 3**).

Increasing it to 2000 increases the number of bins obtained, overall and for medium to high quality, but results in less complete bins in low-coverage regions, as demonstrated by the poor performance on *C. scindens* (**Figure 11.B**, the purple points are all located near the bottom right of the plot).

--clusters: the initial number of clusters. The results for *C. scindens* were similar for all the values tested from 400, the default, to 1200. Nevertheless, the value 1200 performed better overall, producing more medium to high-quality bins (**Table 3**). Similarly to the performance seen when setting the length threshold to 750, redundant and very low-completeness bins were created for some species when increasing the number of clusters. Again, as this issue can be mitigated by filtering the bins with a CheckM quality threshold below 10% (**Figure 12.B**), the value of 1200 was considered preferable and was selected for the optimisation.

A. Validation

This section will present the results obtained on the held-out validation dataset. First, the general trends on the bacterial community as a whole will be presented. Results obtained for *C. scindens* at the abundance level 0.01% and 0.05%, the model used for low-abundance bacteria, will be also be highlighted.

	Quality trimming	Assembly			Binning		
Parameter	Threshold	--k-step	--k-min	--min-count	--kmer-length	--length-threshold	--clusters
Validation 1	0	10	21	2	4	750	1200
Validation 2	0	10	21	2	4	1000	400

Table 2: Parameters used for the different tools during the validation runs. The first run was performed with the parameters selected based on the optimisation process (validation 1). A second run was completed after the proposed parameters appeared to be overfit to the training set. This second time (validation 2) the parameters for the binning tool, CONCOCT, were reverted to the defaults.

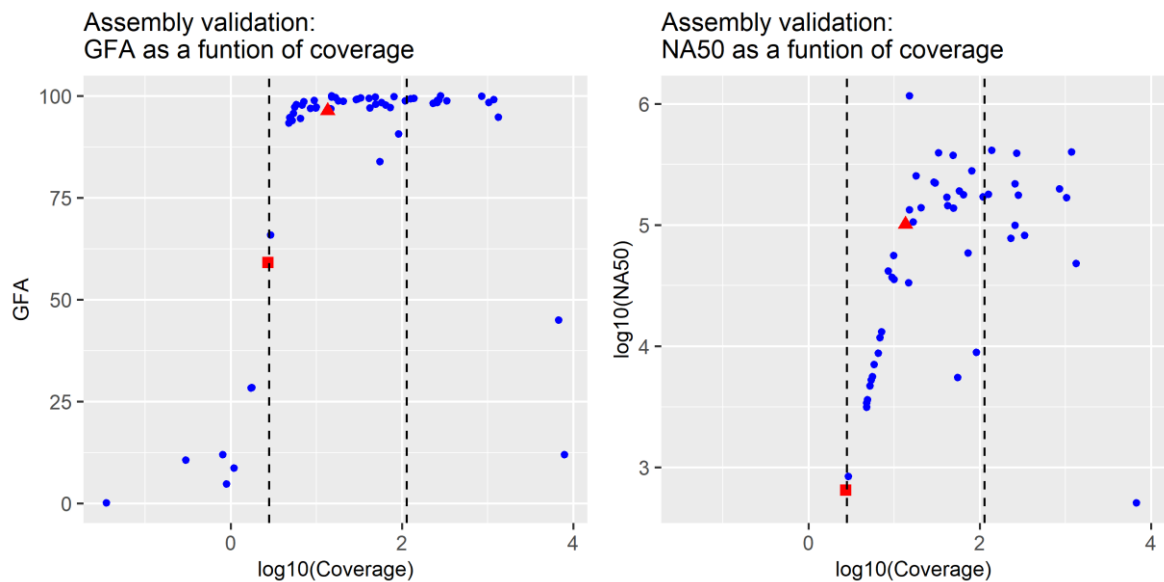


Figure 13: **A)** GFA in function of the average coverage for each genome in the dataset with *C. scindens* at an abundance 0.01%. The red points represent *C. scindens*, with a square and a triangle for abundances 0.01% and 0.05%, respectively. To allow a better readability, other genomes in the 0.05% datasets are not represented as the results are almost identical. **B)** Same as A) but for NA50 values.

1. Assembly:

The GFA for a given bacteria appears to be highly linked to its coverage in the validation dataset, especially at lower coverages. GFA reaches >80% for most of the species as soon as the coverage reaches >5x (**Figure 13.A**). For *C. scindens*, 59.1% and 96.4% of the genome is assembled at an abundance of 0.01% and .05%, respectively.

A similar relationship can also be observed with the NA50 (**Figure 13.B**). It appears more clearly in the low-coverage genomes because the maximum attainable NA50 depends on the length of the reference

genome, contrarily to the GFA, which is bound between 0 and 100%. Altogether, these results correspond well to the ones obtained on the optimisation dataset.

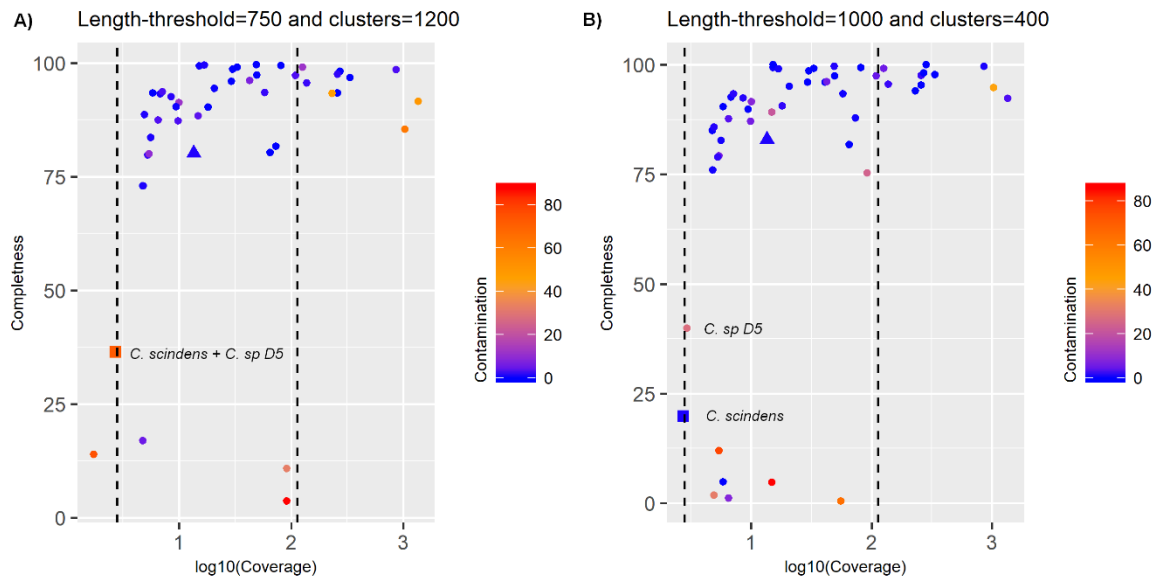


Figure 14: A) Results of the validation with the optimized parameters. The square and the triangle represent *C. scindens* at an abundance of 0.01 and 0.05%, respectively. The colour scale represents the contamination. Bin with higher quality will be at the top left and coloured blue. The dotted lines indicate the low and high ends of the coverage range expected for 7-DH bacteria. Completeness and contamination values were calculated by MetaQuast by comparison with the reference genomes. The coverage was measured with samtools B) Same as A) but for the second validation, with the default binning parameters

2. Binning:

With “--length-threshold” = 750 and “--clusters” = 1200 (Table 2), the results for genomes with a coverage higher than 5x are similar to the results obtained on the optimisation dataset (Figure 10): most bin have a completeness of 80% or above and contamination below 10%. For example, a bin with 80% completeness was reconstructed for *C. scindens* at an abundance of 0.05% (Figure 14.A, represented by the blue triangle). However, the completeness drops at 36%, and more importantly, the contamination reaches 72% when *C. scindens* is only present at an abundance of 0.01% (Figure 14.A, orange square). This resulting bin is composed of 27.4% of *C. scindens* (coverage 2.7x), 54.5% of *Clostridium* sp. D5 (coverage, 2.9x) 9,1% of *Blautia hydrogenotrophica* (coverage 1.7x) and small amounts of 19 other species. On the optimisation dataset, it can be noticed that these two species were separated in two different bins, one containing *C. scindens*, the other containing *C. sp D5* as well as some contaminants (Figure 10.A)

Results on the 0.01% dataset are surprisingly poor compared to the contamination levels obtained on the optimisation dataset. As it was theorized that the high contamination level was due to a detrimental interaction between the two binning parameters (see section IV.C), a second clustering step was attempted with only “--length-threshold” = 750, but it produced identical results. Therefore, the hypothesis of detrimental interaction was rejected.

Finally, a last validation run was completed with the default binning parameters: “--length-threshold” = 1000 and “--clusters” = 400 (Table 2). This resulted in bins with 20% completeness and 0% contamination for *C. scindens* at 0.01% (Figure 14.B, blue square), and 80% completeness with 0.5% contamination at 0.05% (Figure 14, blue triangle). It therefore appears that the default parameters

tend to be more robust than the optimised parameters, as they avoided merging the bin containing *C. sp. D5* with *C. scindens* even on the validation dataset.

IV. Discussion

A. Artificial datasets

For this project, state-of-the-art artificial datasets were created for the optimisation and validation of the pipeline. They simulate as realistically as possible the error profile of the sequencer, the experimental variation in abundance between samples, and the taxonomic profile of a mouse gut microbiota community based on previously established empirical measures. For each species in the profile, reads were sampled from the reference genomes with the highest N50 in the NCBI Refseq database.

Still, it likely underestimates intra-species variation and the presence of species with an abundance below the detection limit of previous studies. Therefore, the future experimental dataset might reveal more complex and contain a greater variety of K-mer, leading to an increase in memory consumption during the assembly. However, it will most likely not be an issue, as the optimized version of the pipeline required only 90Gb of memory at the peak of the assembly and the server on which the assembly of the experimental dataset has a capacity of 256Gb, which leaves room for a significant increase in dataset complexity.

B. Optimisation

1. Optimisation process

The choice of a supervised optimisation process is a convenient approach, easier to implement than other methods such as Bayesian optimisation, which are often considered less biased because they require no supervision. Still, the supervised approach required significant human effort in the selection of parameters and values to test and analysing the results to ensure suitable variable space coverage.

The parameter-by-parameter approach yielded the discovery of several optimal parameter values yet did not require testing all combinations of parameters, which would be intractable. The inconvenience of this method is that interaction between parameters could potentially affect the final set of parameters chosen. However, this drawback is mitigated by the validation process, which prevents eventual detrimental interactions from remaining undetected.

The main weakness of the optimisation process used here is that the parameters were tested on a single dataset, which leads to a higher risk of overfitting. This choice was necessary because computation time for each run did not allow the testing of each parameter value combination on multiple 100Gbp datasets. Again, a validation procedure was employed to aid in the detection of overfitting.

An alternative option may have been to optimise the pipeline on multiple smaller datasets with a coverage maintained at 2.7x for *C. scindens*. In such an ensemble approach, the average results of multiple runs could be compared, thus reducing the potential risk of overfitting. Yet, this design also presents two issues: **1)** some parameter values, such as --k-min 1 for the assembly, could be optimal for low-abundance bacteria in small datasets, but resulting in memory errors (see section III.A.3) when employed on the large validation datasets. **2)** It would require increasing the abundance of *C. scindens* to maintain the same coverage (Equation 1). This could potentially bias the result, if the abundance turned out to also impact the assembly performance in ways other than by affecting the coverage.

C. Validation

The assembly results on the validation dataset are almost identical to the best results on the optimisation dataset (No read trimming, k-min=21, k-step=10). The GFA of *C. scindens* is of 59.1% at 0.01% abundance and 96.4% at abundance 0.05% (**Figure 13.A**). This is within the goal that was set based on previously published results (see section II.J). These results also confirm that the coverage is the limiting factor for the assembly of low abundance species and therefore that the sequencing depth must be calculated to offer a coverage of at least 5x at the lowest targeted abundance.

For the binning parameters, however, an important drop in quality for the bin of *C. scindens* at 0.01% abundance was observed compared to the optimisation dataset (contamination 72% vs <5% previously, **Figure 14.A**). Indeed, on this dataset, CONCOCT appeared unable to distinguish between *C. scindens* (coverage 2.7x), *Clostridium sp. D5* (GCF_000190355, coverage 2.9x), *Blautia hydrogenotrophica* (GCF_000157975, coverage 1.7x) and other low abundant species and merged them in a single bin. It is interesting to note that the 3 species most prevalent in the bin have relatively similar coverages and they all belong to the Clostridiales order, the two most prevalent even being in the same genus. This is likely a challenge for CONCOCT as species in the same genus will have relatively similar genome sequence and therefore be more difficult to distinguish.

The initial hypothesis explaining these results was that a detrimental interaction between the length threshold and the clusters parameters might have occurred. However, restarting the binning process with length-threshold = 750 and clusters = 400 lead to almost identical results, leading to this hypothesis being rejected.

The next hypothesis was that the values of these two parameters were overfit for the optimisation dataset: they happened to perform better on the dataset they were trained on but at the cost of their general performance to other datasets. This hypothesis is supported by the fact that running CONCOCT with the default parameters on the validation dataset (**Figure 14.B**) produced results comparable to the ones previously obtained with the default parameters on the optimisation dataset (**Figure 10.A**).

Taken together, these results indicate that default binning parameters are more robust than the optimised parameter. They could distinguish between relatively closely taxonomically related genomes at similarly low abundance. The optimized parameters, however, were not able to do it on the validation dataset and therefore appear to be less reliable. Consequently, the final suggestion is in fact to use default binning parameters. This important discrepancy between results highlights the importance of validating optimized pipelines on newly generated datasets.

In the end, the main contribution of the optimisation was showing that a low-quality threshold for the read trimming was the best to maximize the GFA of low-abundance species, especially since no default parameter is suggested by the Trimmomatic manual. On the other hand, the default parameters for Megahit and CONCOCT proved to perform the most reliably in this study.

The final pipeline with the validated parameters produces MAG with 20% completeness for *C. scindens* at 0.01% abundance, the lowest end of the range expected for bile acid degrading bacteria. At higher abundances, namely 0.05%, the completeness reaches 80% (**Figure 14.B**). Based on the information provided by bacteria of other species with coverages between 5x and 10x in the validation dataset, one would expect to reliably reach >70% completeness at 0.02% already as this would coincide with a coverage of 5.4x for *C. scindens*.

For the bacteria in the 0.01% to 0.02% range, other approaches may be needed to identify the presence of bile acid degrading gene despite the absence of medium to high quality MAG. As a GFA of almost 60% can be attained even for bacteria at an abundance of 0.01%, it could be possible to select the contigs that were binned in low-quality bin or not binned at all and directly annotate the ORF they harbour. Afterwards, use BLAST to search for bai gene variants in these unbinned contigs. This approach will not allow for studying of the whole genome of the bacteria containing these genes but might still provide valuable information for the design of degenerate PCR primer or on the diversity of bai genes within each condition tested. Otherwise, further increasing the completeness of the bin obtained will likely have to be done by increasing the sequencing depth, as the coverage appears to remain the major limiting factor in the assembly and binning of these bacteria.

V. Conclusion

The first aim of this project, creating a synthetic dataset to mimic the gut microbiota community of mice and test different metagenomic assembly pipelines was completed. The artificial datasets were created using high-quality reference genomes and empirically-based abundance values. As most of the process was automated, the scripts could easily be reused to create other artificial datasets with different levels of abundance for *C. scindens*. As for the second aim, creating a pipeline for discovery of low-abundance species in WMGS datasets, most of the criteria were met. The pipeline developed enables the performing of a WMGS analysis in a single command line. The optimisation process facilitated the identification of optimal read trimming parameters and, for the other tools, showed that the default parameter selected by their author were ideal for this study.

Further, the validation step confirmed that the current tools can accommodate artificial datasets as large as 100Gbp. It also showed that the goal: reaching a GFA of between 50-70% for *C. scindens* at 0.01% abundance, was attained: with our results demonstrating a GFA of 59.1%. Our results also suggest that bins with medium to high completeness and low contamination will be created across a large range at which the bacteria of interest are expected to be found (0.01% to 0.4%), with the potential exception of the extreme low-end (0.01% to 0.02%). To discover bacteria of even lower abundance, the sequencing depth will likely have to be further increased. Another approach could be to try increasing the relative abundance of the bacteria of interest in the community, for example by changing the host diet¹³.

The main limitation of this study is that artificial datasets are inherently limited by currently available resources. Experimental datasets might reveal more complexity and thus result in higher computational requirements, both in memory and time. Still, as assembling the artificial dataset required 90Gb of memory, the server used here, with a capacity of 256 Gb, should be able to perform the task even if an increase in complexity is observed. To conclude, the pipeline developed here is a user-friendly software which is both portable and freely available for the scientific community (see section VIII). It is anticipated to save significant effort of future researchers by facilitating the discovery and analysis of low-abundance bacteria in WMGS.

VI. Glossary

7-DH	7-dehydroxylation
BA	Bile acid
CAG	Co-abundance gene groups
DCA	Deoxycholic acid
GFA	Genome fraction assembled
GMM	Gaussian Mixture Model
LCA	lithocholic acid
MAG	Metagenome-assembled genome
MGS	MetaGenomic Species
ORF	Open reading frame
OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
QC	Quality Control
sDBG	Succinct De Bruijn Graph
WMGS	Whole MetaGenome shotgun Sequencing

VII. Supplementary tables

<i>Parameters</i>	<i>Value</i>	<i>Number of bins</i>	<i>Medium quality or higher</i>
<i>--kmer-length</i>	3	54	29
	4	64	39
	6	12	6
<i>--length-threshold</i>	250	-	-
	375	371	8
	500	83	27
	750	84	38
	1000	64	39
	2000	63	42
<i>--clusters</i>	200	68	38
	400	64	39
	800	70	39
	1200	70	43

Table 3: Describes for each binning optimisation parameter, the resulting number of bins obtained and the number of them with a MIMAG quality label of medium quality or higher. The parameters in bold were selected for validation.

VIII. Scripts

All the scripts described in this document are available on Github at:
<https://github.com/leojequier/Metagenomic-pipeline>

IX. List of figures:

Figure 1: A) Example of 7-DH reaction. CA is transformed in DCA in a multistep process by gut microbiota bacteria harbouring Bai genes. B) The position and size of different Bai genes present in *C. scindens* ATCC 35705. Determined by BLASTing the nucleotide sequence of Bai genes sequences from Heinken et al. (2019) against the genome of *C. scindens*.

Figure 2: Schematic procedure of a 16S sequencing experiment to measure the composition of a bacterial community. The 16S rRNA-encoding gene is amplified using specific PCR primers and the resulting fragments are sequenced. Then the overlapping read pairs are merged. Very similar sequences are clustered in Operational Taxonomic Units (OTU) to prevent the added complexity of sequencing errors. The relative abundance of each OTU is estimated with the number of reads mapping to it. Finally, the resulting OTU are taxonomically annotated by finding their closest match in a database.

Figure 3: Schematic procedure of a WMGS experiment followed by an assembly-based analysis. The DNA is fragmented and sequenced without any amplification step. Then the short reads are assembled in contigs with a length ranging from hundreds to millions of base-pairs. Finally, the binning step clusters the contigs according to their predicated species of origin. This process produces MAGs, which can be further analysed in a similar way to typical draft genomes.

Figure 4: Order- and family-level taxonomic composition of the artificial datasets based on taxonomically assigned MGS (n=541) in Xiao et al, 2015. Relative abundance measures of the MGS with identical species taxonomy were merged, resulting in 96 unique taxa. For each of them, the average relative abundance in five mice of the same strain and diet (C57_BL, HFD) was calculated, resulting in 66>0 MGS.

Figure 6: Example of base quality-score profiles of reads in the artificial dataset. Left panel is R1 and right panel is R2. Base 1, on the left of the x-axis, is the 5' end of the reads and base 150, on the right, is the 3' end. Plots created using FastQC. As expected with Illumina sequencer, a drop in base quality scores can be observed at the 3' end of the read, especially in R2.

Figure 5: Flowchart presenting an overview of the programs and parameters used in the pipeline. The black dotted rectangles above each step highlights for each processing step, the input (green), the output (red) and the parameters tested during the optimisation process. The blue dotted rectangles below present the quality control step and the metrics computed.

Figure 7: **A)** Estimation of the computing time needed for read deduplication. The computing time was measured with datasets of 125K to 32M reads. In R, a polynomial model of degree two was fit to the data and extrapolated to the final size (100Gbp or 333M reads). This resulted in an estimated computing time of 50.8 hours.

Figure 8: **A):** Genome fraction assembled for *C. scindens* at abundance 0.01% as a function of quality threshold. A lower quality threshold appears to be associated with an increased GFA. The GFA was measured with Metaquast by comparing the contigs to the reference genomes. **B):** Mismatches per 100kbp for *C. scindens* in function of the quality threshold. Mismatches were measured by Metaquast.

Figure 9: **A)** Effect of "k-step" on the *C. scindens* genome fraction assembled. 3 values were compared with otherwise identical parameters. The genome fraction assembled was measured by Metaquast. **B)** Same as A but showcasing the effect of "k-min" on *C. scindens* genome fraction assembled. The GFA peaks at 48% with a k-min size of 21.

Figure 10: **A)** Completeness of each bin in function of the coverage of the main genome in the bin. The colour scale represents the contamination. *C. sp D5* and *C. scindens* are labelled for comparison with figure 14. The dotted lines indicate the low and high ends of the coverage range expected for 7-DH bacteria. Completeness and contamination values were calculated by MetaQuast by comparison with the reference genomes. The coverage was measured with samtools. Binning parameters: length threshold 750; clusters 400. **B)** Presents the same data than A, but the colour represents the MIMAG standard quality label for completeness and contamination. These labelled were assigned based on the result of Checkm, which does not require reference genomes.

Figure 11: Only bins containing contigs from *C. scindens* are represented here. Bins with the higher quality will be found at the top left of each of these plots. **A)** Effect of different “--kmer_length” values on the completeness and contamination of all bins containing contigs mapping to *C. scindens*. The completeness represents the proportion of the *C. scindens* genome contained in the bin. The contamination represents the proportion of the bin that does not map to *C. scindens*. **B)** Same as A, but for the length threshold **C)** Same as A, but for the number of initial clusters.

Figure 12: **A)** Highlights the high number of very low completeness bins created by using a length threshold of 750 and their efficient filtering by removing bins with CheckM completeness values <10 (in red). **B)** Same figure for the parameter “--cluster” = 1200.

Figure 13: **A)** GFA in function of the average coverage for each genome in the dataset with *C. scindens* at an abundance 0.01%. The red points represent *C. scindens*, with a square and a triangle for abundances 0.01% and 0.05%, respectively. To allow a better readability, other genomes in the 0.05% datasets are not represented as the results are almost identical. **B)** Same as A) but for NA50 values.

Figure 14: **A)** Results of the validation with the optimized parameters. The square and the **triangle** represent *C. scindens* at an abundance of 0.01 and 0.05%, respectively. The colour scale represents the contamination. Bin with higher quality will be at the top left and coloured blue. The dotted lines indicate the low and high ends of the coverage range expected for 7-DH bacteria. Completeness and contamination values were calculated by MetaQuast by comparison with the reference genomes. The coverage was measured with samtools **B)** Same as A) but for the second validation, with the default binning parameters

X. List of tables

Table 1: MIMAG standard quality label and associated criteria.

Table 2: Parameters used for the different tools during the validation runs. The first run was performed with the parameters selected based on the optimisation process (validation 1). A second run was completed after the proposed parameters appeared to be overfit to the training set. This second time (validation 2) the parameters for the binning tool, CONCOCT, were reverted to the defaults.

Table 3: Describes for each binning optimisation parameter, the resulting number of bins obtained and the number of them with a MIMAG quality label of medium quality or higher. The parameters in bold were selected for validation.

XI. Bibliography

1. Chatelier, E. L. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
2. Bäckhed, F. *et al.* The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci.* **101**, 15718–15723 (2004).
3. Ridaura, V. K. *et al.* Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice. *Science* **341**, 1241214 (2013).
4. Ma, H. & Patti, M. E. Bile acids, obesity, and the metabolic syndrome. *Best Pract. Res. Clin. Gastroenterol.* **28**, 573–583 (2014).
5. Vitek, L. & Haluzík, M. The role of bile acids in metabolic regulation. *J. Endocrinol.* **228**, R85–R96 (2016).
6. Thomas, C. *et al.* TGR5-Mediated Bile Acid Sensing Controls Glucose Homeostasis. *Cell Metab.* **10**, 167–177 (2009).
7. García-Cañaveras, J. C., Donato, M. T., Castell, J. V. & Lahoz, A. Targeted profiling of circulating and hepatic bile acids in human, mouse, and rat using a UPLC-MRM-MS-validated method. *J. Lipid Res.* **53**, 2231–2241 (2012).
8. Ridlon, J. M., Kang, D.-J. & Hylemon, P. B. Isolation and characterization of a bile acid inducible 7 α -dehydroxylating operon in *Clostridium hylemonae* TN271. *Anaerobe* **16**, 137–146 (2010).
9. Ridlon, J. M., Kang, D. J., Hylemon, P. B. & Bajaj, J. S. Bile Acids and the Gut Microbiome. *Curr. Opin. Gastroenterol.* **30**, 332–338 (2014).
10. Heinken, A. *et al.* Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome* **7**, 75 (2019).
11. Ridlon, J. M., Alves, J. M., Hylemon, P. B. & Bajaj, J. S. Cirrhosis, bile acids and gut microbiota. *Gut Microbes* **4**, 382–387 (2013).

12. Ridlon, J. M., Kang, D.-J. & Hylemon, P. B. Bile salt biotransformations by human intestinal bacteria. *J. Lipid Res.* **47**, 241–259 (2006).
13. Xiao, L. *et al.* High-fat feeding rather than obesity drives taxonomical and functional changes in the gut microbiota in mice. *Microbiome* **5**, 43 (2017).
14. Baker, G. C., Smith, J. J. & Cowan, D. A. Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods* **55**, 541–555 (2003).
15. Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **13**, 47–58 (2012).
16. Rappé, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
17. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1–11 (2019).
18. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* **18**, 1403–1414 (2016).
19. Edgar, R. C. Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**, e4652 (2018).
20. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *bioRxiv* 192211 (2017) doi:10.1101/192211.
21. Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* **69**, 330–339 (2007).
22. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
23. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

24. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).
25. Itoh, N., Kazama, M., Takeuchi, N., Isotani, K. & Kurokawa, J. Gene-specific amplicons from metagenomes as an alternative to directed evolution for enzyme screening: a case study using phenylacetaldehyde reductases. *FEBS Open Bio* **6**, 566–575 (2016).
26. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
27. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32**, 1088–1090 (2016).
28. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
29. Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist’s Perspective - Not Only Size Matters! *PLOS ONE* **12**, e0169662 (2017).
30. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
31. Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc. Natl. Acad. Sci.* **112**, 14024–14029 (2015).
32. Li, Z. *et al.* Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief. Funct. Genomics* **11**, 25–37 (2012).
33. Chatterji, S., Yamazaki, I., Bai, Z. & Eisen, J. A. CompostBin: A DNA Composition-Based Algorithm for Binning Environmental Shotgun Reads. in *Research in Computational Molecular Biology* (eds. Vingron, M. & Wong, L.) 17–28 (Springer Berlin Heidelberg, 2008).
34. Mitchell, A. L. *et al.* EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**, D726–D735 (2018).

35. Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* **14**, R2 (2013).
36. Kultima, J. R. *et al.* MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLOS ONE* **7**, e47656 (2012).
37. Lai, B., Wang, F., Wang, X., Duan, L. & Zhu, H. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. *BMC Bioinformatics* **16**, 244 (2015).
38. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
39. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
40. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6**, 898–901 (2012).
41. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
42. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
43. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
44. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
45. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
46. Andrews, S. *et al.* *FastQC*. (2012).

47. 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).
48. Tange, O. *GNU Parallel 2018*. (Ole Tange, 2018). doi:10.5281/zenodo.1146014.
49. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
50. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
51. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
52. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).
53. Petersen, K. R., Streett, D. A., Gerritsen, A. T., Hunter, S. S. & Settles, M. L. Super Deduper, Fast PCR Duplicate Detection in Fastq Files. in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* 491–492 (ACM, 2015). doi:10.1145/2808719.2811568.
54. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
55. Haider, B. *et al.* Omega: an Overlap-graph de novo Assembler for Metagenomics. *Bioinformatics* **30**, 2717–2722 (2014).
56. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29**, 987–991 (2011).

