University of Minho
School of Engineering

Sara Patrícia Monteiro Martins

A metagenomic approach to identify and characterize wastewater populations

MSc dissertation

Master in Bioinformatics

This work was realized under the supervision of:

Professor Pedro Santos

March 2017

ACKNOWLEDGEMENTS

I would like to thank Professor Pedro Santos from the Centre of Molecular and Environmental Biology from University of Minho for the opportunity to work in this research project and all the support and availability.

To the directors, colleagues and all the staff of the Centre of Biological Engineering for all the support, kindness and friendship, I express my sincere gratitude.

To my master colleagues, for making me feel among friends.

To Tiago Barbosa, for all the help and for always being there for me.

I would like to thank my parents, José Rosas and Fernanda Monteiro for all the support and for never let me give up.

Thank you!

ABSTRACT

Water scarcity and pollution are two main ecological focus nowadays. Knowledge of wastewater composition, regarding microorganisms and pollutants, is of great importance to improve the capacities of the effluent treatment plants (ETP). Advances in Next-generation sequencing (NGS) methodologies allowed for faster, cheaper and more accurate study of microbial communities. Besides being an extremely powerful analysis resource, whole shotgun metagenomic analysis comprises many challenging aspects, regarding the processing and analysis.

In the present work a shotgun metagenomic bioinformatics analysis was performed comprising three samples from common ETPs (CETP) and four samples from a petrochemical complex ETPs (wastewaters with low and high salts collected in two distinct timepoints). The samples were sequenced with Illumina® HiSeq, generating paired-end reads with 2x150bp length. The main goals of this project were to evaluate currently available tools, establish a customized bioinformatics pipeline and to extract relevant biological information from the sequenced datasets.

There were generated simulated datasets representative of the target data, in order to evaluate the performance of the available bioinformatics tools. Datasets were generated with three coverage levels and were used to test pre-processing, assembly and taxonomic tools. The target datasets, both with and without coverage split, were then subjected to processing and analysis using the pre-defined pipeline. A preliminary functional study was also performed using *MG-RAST* and *MGX*.

Results from the evaluation of the performance of the bioinformatics tools showed that different tools behave differently in distinct datasets. The pipeline was defined using *BayesHammer* and *Fastq-mcf* as pre-processing tools, *SPAdes* for assembly and *MetaPhlAn v2.0* for the taxonomical analysis.

The assembly results for the target datasets showed a higher contiguity for high coverage levels and a lower contiguity for low coverage levels, highlighting the differences in microorganisms' abundance and diversity and its impact during analysis.

Taxonomical composition suggests the presence of putative pathogenic and opportunistic microorganisms on two of the CETP datasets (A2 and AKR12). It also suggests a more hostile environment in petrochemical complex ETPs datasets, which is concordant with a higher abundance of defence mechanisms on this datasets.

The present results must be accounted to the effluent treatment processes.

Keywords: whole shotgun metagenomic, next-generation sequencing, effluent-treatment plant

RESUMO

A escassez de água e a poluição são dois dos principais problemas ecológicos atualmente. O conhecimento da composição das águas residuais, referente a microrganismos e poluentes, é de grande importância para melhorar as capacidades das estações de tratamento de águas residuais (ETAR). Os avanços nos métodos de sequenciação de nova geração permitiram o estudo mais rápido, barato e preciso de comunidades microbianas. Apesar de ser um meio de análise altamente poderoso, a análise metagenómica por *whole shotgun* compreende muitos aspetos desafiadores, no que respeita o processamento e a análise.

No presente trabalho, uma análise bioinformática de dados metagenómicos de *shotgun* foi efetuada incluindo três amostras de ETARs comuns e quatro amostras de ETARs de um complexo petroquímico (águas residuais com baixos e altos teores de sais, colhidas em dois momentos distintos). As amostras foram sequenciadas com *Illumina® HiSeq,* gerando *paired-end reads* com comprimento igual a 2x150pb. Os principais objetivos deste projeto foram avaliar ferramentas disponíveis atualmente, estabelecer uma *pipeline* bioinformática personalizada e extrair informação biológica relevante dos *datasets* sequenciados.

Foram gerados *datasets* simulados representativos dos dados a analisar, de forma a avaliar a performance das ferramentas bioinformáticas disponíveis. Os *datasets* foram gerados com três níveis de *coverage* e foram usados para testar ferramentas de pré-processamento, *assembly* e taxonomia. Os *datasets* alvo, com e sem divisão por *coverage*, foram então sujeitos a processamento e análise usando a *pipeline* pré-definida. Um estudo funcional preliminar foi realizado com *MG*-RAST e *MGX*.

Os resultados da avaliação da performance das ferramentas bioinformáticas mostraram que diferentes ferramentas comportam-se de forma diferente em *datasets* distintos. A *pipeline* foi definida usando *BayesHammer* e *Fastq-mcf* como ferramentas de pré-processamento, *SPAdes* para *assembly* e *MetaPhlAn v2.0* para a análise taxonómica.

Os resultados de *assembly* para os *datasets* alvo mostraram uma grande contiguidade para altos níveis de *coverage* e baixa contiguidade para baixos níveis de *coverage*, realçando as diferenças de abundância e diversidade dos microrganismos e o seu impacto durante a análise.

A composição taxonómica sugere a presença de microrganismos potencialmente patogénicos e oportunistas nos dois datasets de ETARs comuns (A2 e AKR12). Sugere também um ambiente mais hostil nos datasets das ETARs do complexo petroquímico, o que é concordante com uma maior abundância de mecanismos de defesa nestes *datasets*.

Os presentes resultados devem ser tidos em conta nos processos de tratamento de águas residuais.

Palavras-chave: metagenómica de *whole shotgun*, sequenciação de nova geração, estação de tratamento de águas residuais

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS, SYMBOLS AND ACRONYMS

CETP	Common effluent treatment plant
COGs	Clusters of orthologous groups
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphate
ETP	Effluent treatment plants
FN	False negatives
FP	False positives
GPUs	Graphics processing units
HMM	Hidden Markov models
KO	KEGG orthology
LSU-rRNA	Large subunit of ribosomal ribonucleic acid
NGS	Next-generation sequencing
NOGs	Non-supervised orthologous groups
ORFs	Open reading frame
OTUs	Operational taxonomic units
PCR	Polymerase chain reaction
PPV	Positive predictive values
rRNA	Ribosomal ribonucleic acid
SEN	Sensitivity
SOLiD	Sequencing by oligo ligation detection
SSU-rRNA	Small subunit of ribosomal ribonucleic acid
TDS	Total dissolved solids
TP	True positives

1. INTRODUCTION

1.1 Wastewater treatment and metagenomic analysis

Besides Earth being largely covered with water, only 3% can be considered freshwater. Problematically, two thirds of the whole freshwater is not suitable for human use nor consumption. Anthropogenic pollution/contamination of freshwater resources tend to further reduce its availability for human consumption (1).

Around 1.1 billion people worldwide do not have ready access to potable water. This water scarcity allied to the massive wastewater production leads to a dire need for its reuse. Consequently, 2.4 billion people are exposed to inadequate sanitation and 842000 people are estimated to die each year from diarrhoea resulting from drinking unsafe and polluted water, bad sanitation and hand hygiene. The proliferation of microorganisms in wastewater derived from poor sanitization processes is of great concern to society as it can potentially increase the number of resistant strains and contaminate even more potable water sources (1,2).

Climacteric changes and human activities have been increasing the water scarcity and the increase of contamination by pollutants; if the actual consumption rate maintains, it is estimated that by 2025 two-thirds of the world's population will experience water scarcity. In order to optimize water decontamination it is paramount to assess exactly which organisms and pollutants are present. This is a very important step as the wastewater treatments can be adjusted for each more prevalent organism and maximizing the efficiency of treatment complexes (1–3).

As aforesaid, the treatment of domestic and industrial wastewaters is an ecological problem of great concern. The wastewater treatments aim to remove or reduce suspended, biodegradable organic compounds, nutrients such as nitrates and phosphates that can lead to high algae concentrations and pathogenic organisms. Industrial wastewaters contain many and diverse toxic compounds which must not be delivered untreated into the environment. Industrial wastewaters are usually treated in effluent treatment plants (ETP) comprising physico-chemical and biological treatments. This is usually effective for larger industries more than in minor industries as they cannot afford their own ETP. Therefore the effluents of small industries are collected in the so called common effluent treatment plant (CETP) which may lack specificity for the treatment. Identifying microorganism strains and principal pollutants can prove a vital step for optimization of CETPs and wastewater sanitization (4). This work will focus on shotgun metagenomic sequencing using next-generation sequencing (NGS) technology aiming to characterize the

microbial communities present in ETPs, and thus, providing the necessary insights for optimization of the treatment processes.

Next-generation sequencing can be described as an unsolved puzzle. In this case, the puzzle pieces are the DNA fragments, called reads. While working with a single genome, there is a unique puzzle (an organism genome) with many pieces (DNA reads) and the objective is to solve the puzzle and find the hidden image (genomic information). This puzzle can have equal or quite similar pieces (repeated sequences), damaged pieces (sequencing errors), and pieces with patterns that are not from the puzzle (adapters and primers contaminants) as well as some missing pieces. This leads to a serious difficulty in solving the puzzle, this is, reconstruction the present genome. Moreover, when working with a metagenome, besides all the referred challenges, there are an undefined number of repeated different puzzles (different organisms) and some puzzles are repeated more times than others (different abundances) (5,6).

Advances in NGS allowed for faster, cheaper and more accurate study of microbial communities. The metagenomic study of these samples is able to give information not only about the microorganisms within but also about their functions and infer the major components of the sample (5). This information is invaluable for the optimization of wastewater treatments and may prove a valid tool for analysis. However, this analysis has to be performed with caution regarding the mentioned complexity.

1.2 Overview/Research goals

The aim of this study is to perform a thorough analysis of shotgun metagenomic data for identification and characterization of different organisms in different wastewaters treatments from India. The samples were sequenced with Illumina® HiSeq, generating paired-end reads with 2x150bp length.

The main goal is to access the microbial samples composition and the main functions presents in the metagenome.

Whole shotgun metagenomic analyses are very challenging and many questions can arise during the process. The first step of the process is the selection of the strategies and tools to be used to perform the analysis. Then, a question arises: “How to choose the best tools?” In order to answer this question the construction of a simulated dataset would permit the comparison between the expected and the given results. Since different tools may work better with different datasets characteristics, the simulated dataset must be representative of the target datasets.

The first objective was, therefore, the construction of a simulated dataset representative of the target datasets in order to evaluate the bioinformatics tools to be used. Then, the objective was the selection of a pipeline to use to pre-process, assemble and perform taxonomic analysis on the target datasets. Once the tool selection is completed, the target data analysis may begin.

Metagenomic datasets have different organismal abundances. Therefore, another problematic was addressed regarding the different coverage of the organisms present. Datasets were split into three coverage levels and results from split and not split datasets were compared for assembly, taxonomic and function composition.

The assembly impact on taxonomic and functional analysis was also addressed.

A functional analysis was also performed on the target data using both datasets with and without coverage split and with and without assembly.

1.3 Dissertation organization

The present document is structured in five chapters. The introductory chapter highlights the water scarcity problematic, the water reuse issues, the wastewater treatment importance and the advantages of the metagenomic analysis of those types of samples in order to optimize the treatment processes.

The second chapter expresses an overview for the NGS concepts, platforms, data analysis steps and main tools, focusing on metagenomic analysis and its characteristics.

A description of the analysed samples and the methodology used can be found in the third chapter.

In the fourth chapter the main results are presented in two parts. The first part comprises the results attained in the tool analysis and the further pipeline assignment are presented. The results obtained with the sample's analysis, discussing their bioinformatics and biological impact, are afterwards thoroughly described in a second part.

Finalizing, in chapter five, a brief general conclusion is provided, focusing on the achievements made, qualitative analysis of the work as well as an insight into possible future opportunities for this thematic.

2. STATE OF ART

2.1 Next-Generation Sequencing

2.1.1 NGS concepts

NGS platforms are able to sequence millions of small fragments of DNA at the same time, leading to a significantly cheaper and faster approach when comparing with Sanger sequencing. In addition it also uses less DNA sample as input and it is more precise and consistent (7). The methodology behind NGS comprises a library preparation where input DNA sample is fragmented and common adapters added to the small fragments. Then, depending on the sequencing technology, clonally clustered amplicons can be generated by different approaches, such as *in situ* polonies, emulsion PCR and bridge PCR. The present work will be focus on Illumina sequencing methodology which uses a bridge PCR approach, as shown in Figure 1. In bridge PCR cluster generation, the DNA fragments are loaded into a flow cell with two types of primers, each one complementary to the adapters in each side of the fragment. The fragment hybridizes with one of the complementary primers and a complement of the hybridized fragment is generated. The molecule is denatured and the original strand is washed away, then the other adapter hybridizes with the second type of primer and the strand is clonally amplified by bridge amplification. This process is repeated over and over amplifying all the fragments. After the cluster generation, a sequencing by synthesis step takes place. The sequencing begins with the extension of the sequencing primer; afterwards dNTPs are then incorporated based on the sequence of the template. Post the addition of each nucleotide the clusters are excited by a light source being then the characteristic signal emitted. The output at the end of the sequencing process is a numerous amount of small fragments sequenced. Bioinformatics tools are needed to reorganize all those fragments into a complete molecule (5,8).

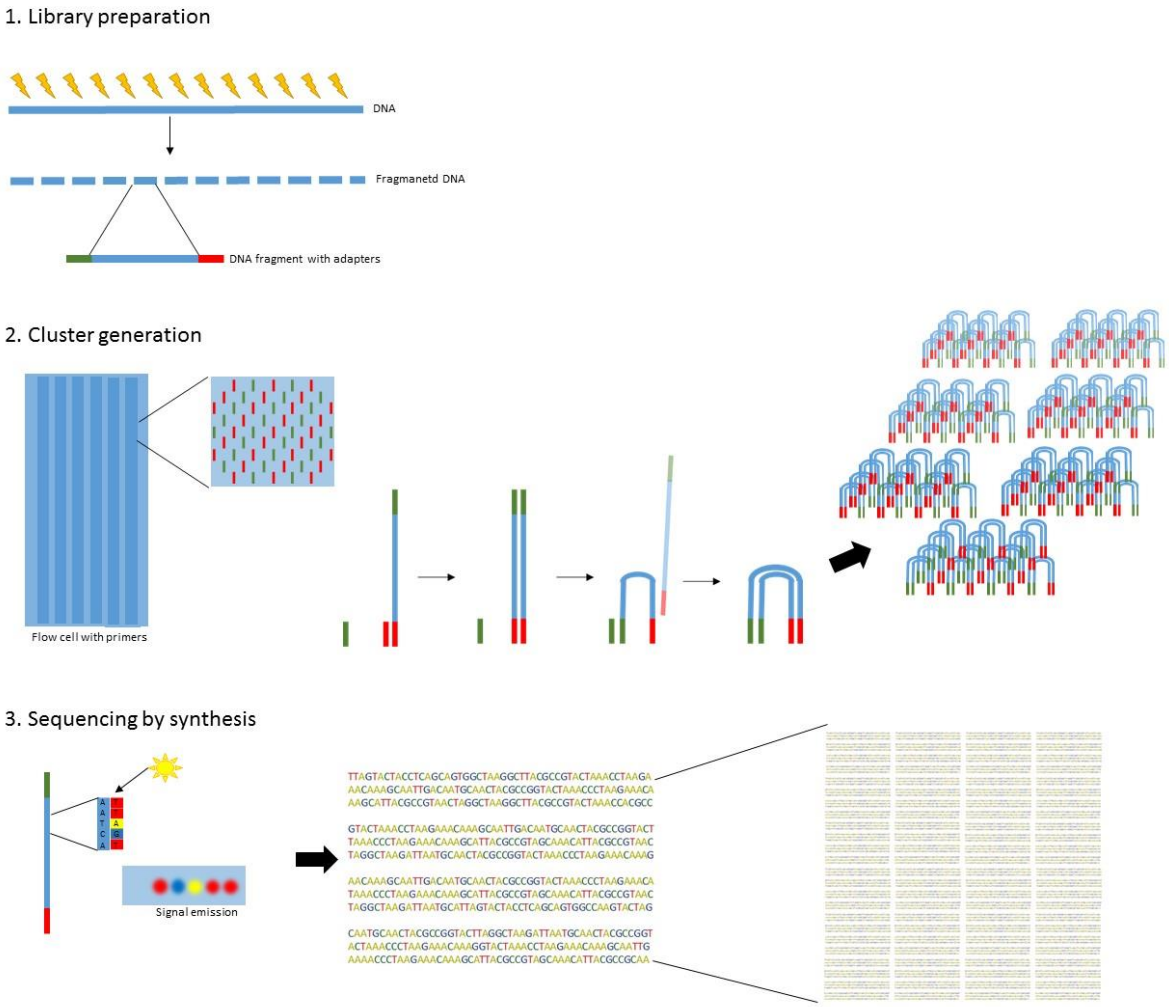


Figure 1 Schematic representation of the NGS Illumina sequencing technology

2.1.2 NGS platforms

The first NGS technology was the 454 pyrosequencing method by Life Sciences commercialized in 2005 (acquired by Roche in 2007). This technology generated an output of 20 Mb, approximately 200 000 reads with 100 to 150bp. In 2006, Solexa released Genome Analyzer, a sequencing by synthesis technology with an output of 1G per run. In the next year, Solexa was purchased by Illumina. Also in 2007 Applied Biosystems, which had acquired Agencourt Personal Genomics in 2006, released the Sequencing by Oligo Ligation Detection (SOLiD) generating approximately 3G with 35bp reads. SOLiD uses a technology of two-base sequencing based on ligation sequencing (5,9).

The mentioned methods have improved and have different characteristics: 454 GSFLX (Roche), which was discontinued in 2016, is able to give an output of 700Mb with up to 1000bp reads with a 1% error regarding insertions and deletions; HiSeq4000 (Illumina) can generate up to 1500Gb of 2x150bp reads with 0.1% substitution error rate; SOLiD5500xl (Applied Biosystems) is capable of generate up to 2 Human Genomes per run with less than 0.1% adenine/thymine bias (10).

Ion Torrent released, in 2010, a system with semiconductor technology developed by the 454 founder, Jonathan Rothberg. This system, named Persona Genome Machine, generate an output up to 2Gb with approximately 200bp reads with 1% insertion/deletion error rate (9,10).

Other systems were developed such as the SBS system MAX-Seq from Intelligent Bio-Systems (purchased by Qiagen in 2012), Polony sequencing, or the single molecule fluorescent sequencing Helicos Genetic Analysis System. In 2010 Pacific Biosystems released the, so called, third generation sequencing PacBio RS generating up-to-several-kilobase-long reads. This system uses a real-time sequencing by synthesis technology (9).

Oxford Nanopore Technologies released MinION in 2015, a nanopore DNA sequencing sequencer that can generate reads with the same length as the input fragment, being theoretically capable of sequencing a whole DNA molecule. The longest reported read had between 230 and 300kbp with a 12% insertion/deletion error rate (9–11)

The methodologies described have different characteristics, involving time, costs, sample type and computer requirements (9,10).

2.2 Data analysis

2.2.1 Pre processing

The reads obtained with NGS may contain errors from misidentified nucleotides, low quality base calls, adapters used on the sequencing process, duplicates and contaminants. These erroneous bases lead to an inefficient interpretation of the generated sequencing data. In order to minimize this issue, different bioinformatics tools have been developed (12).

- Quality control

There is a number of tools allowing a first assessment of the data quality before proceeding with any correction. *FASTQC* is a quality control tool with graphical interface that can be used as an overview of the data quality. Other tools and toolkits are available not only for quality control but also with pre-processing capabilities, as for example *NGS QC Toolkit*, *QC chain*, *FASTX Toolkit* and *SAMtools* (13–16).

- Error correction

Correction of misidentified bases relies in the idea that when laying all the reads containing a specific position, the correct base is the one that appears more often (17). An approach is the k-mer spectrum which works by decomposing the reads in k-mers (read substrings of k size) and relying on their frequency to determine if they are *solid* or *insolid*. The algorithms proceed then to the correction of the *insolid* k-mers (17). Examples of this error correction type are *Reptile*, *Quake*, *BayesHammer*, *Musket* and *RACER*. Other type of error correction uses the multiple sequence alignment of reads sharing the same k-mers seeking for correction relying on the multiple alignments. *Coral* and *Echo* are tools built upon this concept (17–19). On the other hand, *SHREC* and *HiTEC* tools use, respectively, a suffix trie and a suffix array data structure instead of an alignment algorithm (20,21).

- Quality trimming

NGS outputs contain information of each nucleotide quality score (Phred score) (22). Quality trimming tools aim to remove the nucleotides with quality score behind a defined threshold. There are different approaches to accomplish this pre-processing step. Running sum algorithms, such as *Cutadapt*, *SolexaQA*, *Erne-Filter* and *QcReads,* which run from one side of the read to the other calculating a sum formula based on Phred score and a quality threshold. A position is marked to cut when it reaches a minimum. On the other hand, *ConDeTri*, *FASTX toolkit*, *PRINSEQ*, *Sickle*, *SolexaQA*, *Trimmomatic* and *Btrim* use sliding window-based algorithms, where a window size and a mean base quality threshold is defined and the window slide from one side of the read to the other until a passing quality window is encountered. Other approach is used by *UrQT* which comprises a probabilistic unsupervised segmentation with no need for manual set parameters (23–28).

Yun *et* al had also defined masking as an alternative to quality trimming. While trimming leads to the removal of bases, masking is the substitution of low quality bases to "N"s. *FASTQ Masker* available both in Galaxy and *FASTX Toolkit* and *SubN* are examples of masking tools (29).

- Adapter trimming

One important step in the pre-processing of NGS data is the removal of adapters, primers or other exogenous contaminant present in the reads. *AdapterRemoval*, *Alien Trimmer*, *leeHom*, *ngsShoRT*, *PEAT* and *QcReads* are examples of tools designed to remove this exogenous contaminants present in the data (30–33). *Fastq-mcf* removes adapters by scanning it on sequences and performs clipping based on a log-scale threshold. It performs also skewing detection and quality trimming (34). *Btrim* and *Trimmomatic* have also an adapter trimming capability.

Besides the quality and adapter trimming itself, most trimming tools allow also the exclusion of small fragments. After the sequencing and trimming processes, if a fragment is considerably small it probably does not have biological significance. Accounting for that, it is better to exclude a small fragment which will probably compromise the assembly process, given that its biological significance will have almost no impact.

2.2.2 Assembly

The short reads generated by the NGS platforms leads to a bioinformatics challenge in the assembly process. The available tools can be classified in three types: the overlap-and-extend approach, which extends two overlapping reads; the string graph assemblers that built a string graph from the data; and the *de Bruijn* graph approach in which the original sequence reads are segmented into smaller fragments of k size (k-mers) and a direct graph is constructed using the k-mers as edges and the k-1 prefixes and suffixes as nodes (6,35). Although some assemblers with option for metagenomic data exists, this tools are based on adaptations and an assembler dedicated at the whole shotgun metagenomic type of data is not yet available.

2.3 Metagenomic analysis

Metagenomic analysis is the study of genomic DNA from an assemblage of communities. It has an important role in the study of organisms that cannot be cultured as well as for the study of microbial communities as a whole (6).

An approach to identify microbes within a sample is the amplicon sequencing. This consists in the amplification and sequencing of a genomic marker common to the majority of the organisms in order to classify the different microbes. The small-subunit of ribosomal RNA (SSU-rRNA) 16S is often used to characterize Bacteria and Archaea. The 16S amplicon sequencing is a useful method to identify different

organisms in a community. However, it has some limitations, including PCR and sequencing errors, the fact that different organisms have different number of 16S copies, the fact that 16S locus can be transferred between distantly related taxa and also because some microorganisms, especially virus, cannot be identified with this approach (6,36).

The development of next generation sequencing techniques provided the possibility to identify and characterize each organism in a sample by shotgun metagenomic sequencing. This strategy is an alternative to 16S amplicon sequencing which avoids the aforesaid limitations and, since it works with the whole sequence, it is possible to infer gene related functions in the metagenome (6,37).

2.4 Whole shotgun metagenomic sequencing

The tools used to treat metagenomic data rely on the same approach used for whole shotgun genome sequencing. Metagenomics; however, comprises a greater challenge regarding not only the large data volume needed to get meaningful results, as also the issues related with indefinite number of organisms, some of them not yet characterized or even known (6). The communities' divergence may lead to genomes that are not completely covered by reads. On the other hand, high proximity between organisms in the metagenome can lead to overlaps between reads of different organisms leading to chimeras (6). The presence of host DNA, such as human, can also be a problem for following assembly. Some tools have been developed to minimize this problem (38).

There are different algorithms and different strategies to treat metagenomic data, which can be roughly summarised in pre-processing, assembly, taxonomic classification, gene prediction and functional analysis techniques (6).

2.4.1 Pre-processing

As mentioned, the tools used for pre-processing a whole shotgun genome sequencing can be also used for metagenomic data. Nevertheless, there are some tools designed for metagenomic data pre-processing, such as *Meta-QC-chain* and *PRINSEQ* which work as both as quality control and trimming tools (24,39).

2.4.2 Assembly

The assembly of reads can be performed comparing to a referenced genome. A high percentage of organisms in metagenomic data are, however, unknown, leading to the need of a de novo assembly. The

tools used on this strategy are usually based on the *de Bruijn* graph approach (6,37). A number of tools have been adapted and developed for metagenome reconstruction

Meta-IDB, a *de Bruijn* graph approach directed to metagenomic assembly, is now out of maintenance, being substituted by *IDBA-UD* (40). The last is an extension of *IDBA* for assembly of low-depth regions by paired end reads and error correction in high depth regions by the usage of progressive depth on contigs. It uses an iteration strategy from a minimum to a maximum k, where in each it creates an accumulated *de Bruijn* graph, increases the value of depth cut-off threshold and corrects the errors in the reads. At the end of the iteration process it constructs and returns the scaffolds and the maximum confident contig (41). This tool performed better than other assemblers regarding the computational speed and contig alignment size; however, it has shown a higher percentage of chimeras (42)

GeneStich uses also a de *Bruijn* graph approach in which each path in the graph is referent to a gene. To infer the paths it uses a reference genome (43).

RayMeta is an extension of the *Ray* genome assembler which also uses a *de Bruijn* graph approach that finds specific sub-sequences and extends each one into a contig. This tools have the particularity of running in many computers simultaneously, with the possibility of running in just one processor core. In addition to the assembly process, *RayMeta* also performs taxonomic profiling using a graph colouring strategy which adds a different colour for each reference genome (44).

PRICE uses paired-read information to iteratively increase the size of existing contigs. Initially, those contigs can be individual reads from a subset of the paired-read dataset, non-paired reads from sequencing technologies that provide non-paired data, or contigs that were output from a prior run of *PRICE* or any other assembler (45). A different methodology is the overlap-and-extend approach used by *Omega* (overlap-graph metagenome assembler) (46).

Metavelvet is an extension of the *de Bruijn* graph based assembler *Velvet* which decomposes the *de Bruijn* graph constructed from mixed short reads into individual sub-graphs. To disconnect two subgraphs, it identifies nodes shared by both sub-graphs (chimeric nodes) and uses these nodes as the break point. *Metavelvet- SL* is an improvement of *Metavelvet* with the use of supervised learning for the classification of the shared nodes (chimeric nodes). This tool first creates the general *de Bruijn* graph, extracts and classifies each chimeric node with use of a learning module, splits the graph into subgraphs and performs the scaffolding procedure. It provides also a pipeline connecting *Metavelvet-SL* and the profiling method *MetaPhlan* in order to generate the training sample for the supervised learning (47).

SPAdes, another assembler based on the *de Bruijn* graph, was primarily created for single cell assembly. However, the last release comprises a metagenomic option. *SPAdes* was created with intuit of reducing

sequencing errors, non-uniform coverage, insert size variation, chimeric reads and bireads. It starts by creating a *de Bruijn* graph using k-mers, and then it operates on graph topology, coverage, and sequence lengths without the use of the k-mers or the sequence. Finally, the consensus DNA sequence is restored (48).

There have also been described combined assembly strategies in order to conduct to more robust and accurate assemblage (42,49).

2.4.3 Taxonomic characterization

One of prior interests in metagenomic data analysis is to know which organisms are present in the community. The taxonomic characterization of metagenomic data can be assessed by the study of marker genes, such as 16S, with the use of binning strategies to assign each read to a taxonomic group or by assembling to a known genome (6,37).

Phymm uses interpolated Markov models to characterize variable-length oligonucleotides into a phylogenetic group. *PhymmB* is a hybrid method gathering information from *Phymm* and *BLAST* (50).

An approach to identify the organisms in the sample is the use of marker genes. *MetaPhyler* relies on 31 phylogenetic marker genes. It uses *BLASTX* to build taxonomic classifiers and classifies the sequences concerning the best reference hit (51). *AMPHORA* also relies on a marker gene database, containing bacterial markers. *AMPHORA2* is an improvement of *AMPHORA* including archaeal marker genes (52). *MetaPhlan2* is an improved version of *Metaphlan* and uses also an marker gene database approach comprising both bacteria and archaea and uses nucleotide *BLAST* to align the reads into the database (53,54).

Instead of marker genes, *PhylOTU* relies on operational taxonomic units (OTUs) identified by means of SSU-rRNA. The algorithm uses the phylogenetic distance acquired from a phylogenetic tree of SSU-rRNA reads to cluster reads into OTUs (55).

To perform an taxonomic characterization, *PhyloSift* call on *LAST* for sequence similarity search; applies the *hmmalign* program from the *HMMER 3.0* software package to perform the alignment to reference multiple alignment; uses *pplaner* to place the sequences into a phylogenetic reference tree; and produces Krona plots for visualization (56). *Metaxa2* (an improvement of *Metaxa*), also uses hidden markov models (HMM) with *HMMER* to align the sequences to conserved regions (SSU and LSU-rRNA) (57). HMM is a stochastic method to create probabilistic model of randomly changing systems assuming that the future states depend only on the present and not on the previous events. A HMM can be presented as a simplest dynamic Baysian network (58).

Parallel-meta 2.0, the improved version of *Parallel-META 1.0*, includes functional analysis based on Gene Ontology term and SEED annotation. The taxonomy is also assigned based on the HMM algorithm and uses both 16 and 18S rRNA markers (59).

SeMeta is a taxonomic characterization tool using a semi-supervised learning algorithm, which groups reads into clusters and then uses the *Lowest Common Ancestor algorithm* to assigns the lowest common taxon to each cluster (60).

2.4.4 Functional annotation

Perhaps the greatest advantage of the shotgun metagenomic analysis is the possibility to perform a functional analysis and annotation, which can be independent of the identified taxonomy. Thus, the functional annotation of coding sequences within the reads permits characterize the organisms present and the metagenome. The first step is to predict the coding genes present in the sample. This can be performed in assembled or unassembled data with binning processes, protein classification and de novo gene prediction techniques (37).

Using a stochastic approach for predicting bacterial and archaea genes, *MetaGene* functions in two main steps: first it extracts and scores all possible ORFs by their base compositions and lengths; then it calculates an optimal combination of ORFs considering both the ORFs' scores and the scores of orientations and distances of neighbouring ORFs (61). *MetaGeneAnnotator* is an improvement of *MetaGene* including an prophage model, an ribosomal binding site model and a self-training model in order to predict both typical and atypical genes (62).

Ab initio tools have the advantage to identify genes not referenced in the databases. Some of these tools uses a HMM approach. *MetaGeneMark* plugin developed by *GeneProbe* is an example of a HMM based approach which relies on the nucleotide composition (63).

FragGeneScan is also based on HMM and can be applied both at complete genomes and metagenomic fragments. The algorithm relies on codon usage, sequence patterns for start/stop codons and sequencing error models (64).

GlimmerMG uses a interpolated Markov model approach, it uses *Phymm* to do a phylogenic classification of the sequences and to make initial gene predictions; then the sequences are clustered with *Scimm* achieving a final gene prediction (65).

Some functional annotation tools use a machine learning approach to an *ab initio* gene prediction. *Orphelia* identifies, extracts and scores all ORFs by its GC content using a machine learning model. Then, a combination of highly probable genes is selected by a greedy method with a maximal overlap constraint

(66). *MetaGUN* is an improvement of *MetaTISA* using a machine learning approach. First it groups the fragments by phylogeny by means of a k-mer based naïve Bayesian sequence binning method. Then, extracts and scores all possible ORFs using support vector machine classifiers. To conclude, it adjust the translation initial sites of all predicted genes (67). *MGC* is an improvement of *Orphelia* which computes separate learning models for different GC ranges (68).

The study of the gene's functions within a metagenomic sample can be achieved by the use of similarity search algorithms such as *BLAST*. However, the high computational time required by *BLAST* turns this task impossible. Some faster versions of *BLAST* such as *BLAT*, *LAST*, *LASTP*, *UBLAST* and *USEARCH* were developed, but their lack of sensitivity turn them insufficient for metagenomic analysis (6,69). An approach to improve the speed of the similarity search maintaining sufficient sensitivity is the use of graphics processing units (GPUs). *GHOSTMO* and *CLAST* are examples of search tools for metagenomic analysis implemented as GPU systems (70,71) .

FR-*HIT* constructs a k-mer hash table for the reference genome sequences; identifies fragments of reference sequences capable of aligning with the query; removes the fragments that do not enclose qualified alignments and performs banded alignment (72).

RAPSearch2 follows the seed-extension approach used in *BLAST*, with a reduced amino acid alphabet using 10 symbols to represent amino acids groups. Instead of using a suffix array as in *RAPSearch*, *RAPSearch2* uses a collision-free hash table (data structure that assigns each key to a unique value) to index a protein database (73) .

- Functional annotation databases

After the gene prediction, it is necessary to get a meaningful function for each gene and correspondent protein. To accomplish this, there have been created several protein databases with sequence, structure and function information. Pfam, a database focused on the protein domain level and TIGRFAM, a database containing whole protein chains, comprise a large number of family proteins, based on UniProt database, and are represented by multiple sequence alignments and HMMs (74,75).

The gene ontology database uses both ontologies and annotations of the genes and proteins producing a scheme to describe function at different levels (76).

The clusters of orthologous groups (COGs) allows the assignment of orthologues and paralogs for most genes. Orthologues are genes from different organisms with a common ancestral. An orthologous family is formed by a group of three or more proteins from distant genomes having a higher similarity between them than to any other proteins from the same genomes. After a major update in 2014, the COG database

comprises 4631 COGs divided in 26 functional categories (77,78). *EggNOG* database is an extension of the COG database, including more genomes and non-supervised orthologous groups (NOGs) which lack of manual supervision and annotation (79).

The KEGG database includes fifteen manually curated databases and a computationally generated database, which are divided in four categories (system, genomic, chemical, and health information). KEGG orthology (KO) database comprises sequence functional groups and functional orthologs. Unlikely COGs, KOs may consist of a single gene or may contain multiple sequence similarity groups. KOs are defined taking into consideration pathways, genes clusters and phylogeny (80).

SEED is another functional database that relies on a subsystem based annotation. A subsystem is defined as a collection of functional roles that implement a specific biological process or a structural complex (81).

Many tools have been developed to provide a functional analysis and annotation based in one or more databases. Some tools have been developed to perform annotation based only on the SEED system, such as *RAST* and *SUPERFOCUS* (82,83). *RAMMCAP* performs clustering and ORF finder followed by annotation with Pfam, Tigrfam and COG (84). *COGNIZER* framework also provides functional analysis with KEGG, Pfam, GO and SEED databases (85).

MEGAN retrieves taxonomical and functional information on previous aligned data. Regarding the functional analysis *MEGAN* can use InterPro2GO, eggNOG, SEED and KEGG databases (86,87). *MG-RAST*, a web-based software can also perform quality control, statistical, clustering, taxonomical and functional analysis on assembled or unassembled data (88). Another web-based tool comprising metagenomic tools, including quality control, taxonomy, ORF calling, clustering and functional analysis tools is the *WebMGA* server (89).

2.4.5 Pipelines and Workflows

The analysis of whole shotgun data requires some knowledge in the bioinformatics field. In order to conduct this kind of analyses, there were developed software packages, workflows or simple pipelines to enable the biological or medical researcher to analyse the data without advanced knowledge in bioinformatics and command line tools. These pipelines may include more or less analysis steps and parameter selections. *QIIME* is a software that performs microbial community analysis (90). *Mothur* software also performs microbial community analysis and allows also trimming, alignment and taxonomic analysis based on 16S rRNAs (91). *MetAMOS* pipeline permits not only the assembly process as also the taxonomical and functional analysis. *InteMAP* is a pipeline for assembly using *ABySS*, *IDBA-UD* and

CABOG assembler regarding the coverage sequence depth (49). This and other software can be much easily used by the operator and are being developed in a large scale. However, different datasets need different processing procedures and this type of analysis needs user sensitivity and knowledge to choose the best approach for each analysis.

2.4.6 Coverage

An important factor to consider when analysing whole shotgun metagenomic data is the fraction of the metagenome represented in the dataset. A metagenome is formed from different organisms in different abundances and therefore it is important to note that some of these organisms are highly represented in the sample and others are represented in small amounts. This different coverage of the organisms in the sample dataset is of great importance. After the sequencing process, the DNA's of the different organisms are fragmented and mixed together (92). This and the putative presence of errors can lead to a misinterpretation of low coverage reads as erroneous sequences, due to its low prevalence in the dataset. This is critic during the assembler steps, as most assemblers are based on *de Bruin* graphs: the less common reads can be assigned to leaves and further discharged. Considering that it is important to treat the data in order to ease the assembly process, one way of doing it is a random subsampling. However, while subsampling high coverage reads will give a good amount of data to be assembled; subsampling low coverage reads could lead to the exclusion of those reads.

Another strategy is to use a normalization approach, where the highly present reads will be reduced, and low coverage ones will be maintained. Also a split of the dataset taking into account the coverage rate can be also performed. An example of tool that was designed to access this kind of coverage based normalization is *bbnorm* from *BBtools* package (93).

3. MATERIAL AND METHODS

3.1 Datasets

3.1.1 Target datasets

This works aims to analyse and characterize the organisms present in seven Illumina® MiSeq datasets from wastewater treatment procedures from India. The sequenced samples generated paired-end reads with 2x150bp length.

The metagenomes were obtained from activated sludge collected from industrial ETPs. Dataset AKR06 was collected from Jeedimetla Effluent Treatment Limited (JETL), a CETP at Jeedimetla, which is a popular industry area located at the Hyderabad city. Datasets A2 and AKR12 were collected from two CETP at Ankleshwar city, a place with lots of dyes and textile industries. Datasets L1, L2, H1 and H2, were collected from a petrochemical complex in Western India that generates two types of wastewater, one with low total dissolved solids (TDS) –L1 and L2 - and the other with high TDS - H1 and H2. Both streams are treated separately in different ETPs.

For each metagenome dataset, there were collected 9 samples that were then pooled to make a homogenous sample.

The summary description of each dataset can be found in Table 1.

Table 1 Metagenome datasets description

Dataset	*Description*	*Number of paired-end reads*
AKR06	JETL CETP activated sludge	5 217 070
A2	Ankleshwar CETP	4 374 324
AKR12	Ankleshwar CETP	6 135 013
L1	Low TDS activated sludge (wastewater from petrochemical complex) timepoint 1	4 421 598
L2	Low TDS activated sludge (wastewater from petrochemical complex) timepoint 2	5 017 797
H1	High TDS activated sludge (wastewater from petrochemical complex) timepoint 1	5 279 434
H2	High TDS activated sludge (wastewater from petrochemical complex) timepoint 2	4 316 467

3.1.2 Simulated datasets

In order to analyse the tools performance and to set a pipeline to subject the target datasets, two simulated datasets were generated (Figure 2).

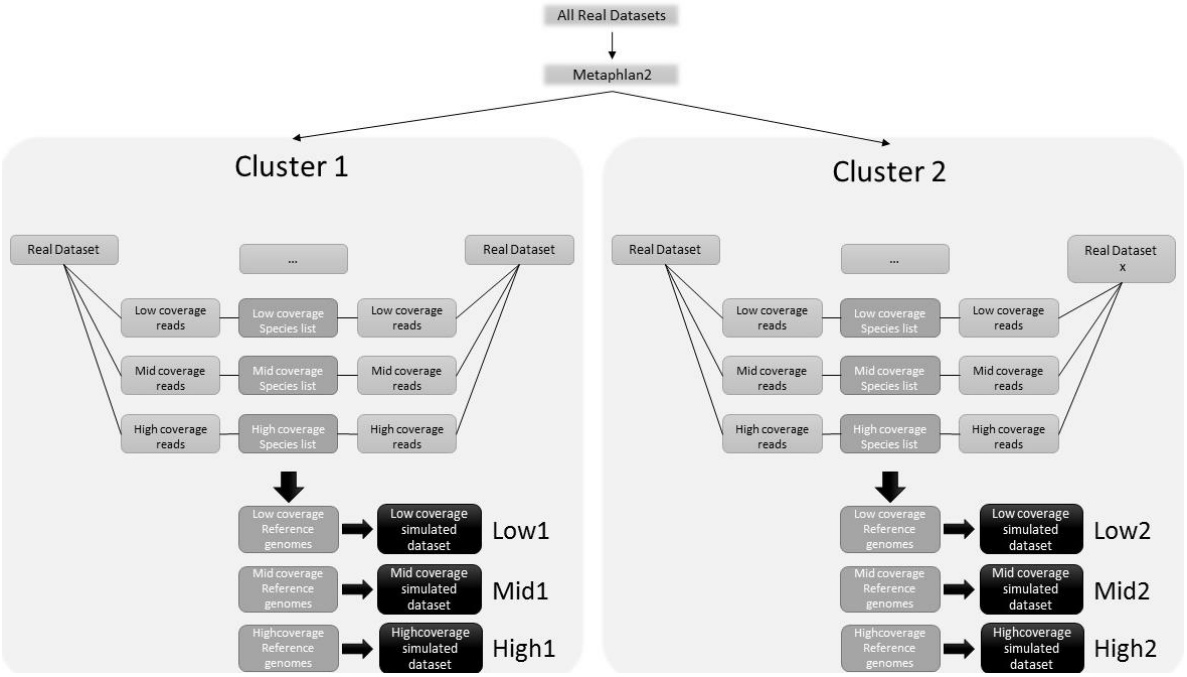


Figure 2 Steps overview to create the simulated datasets.

Two data clusters were generated using *MetaPhlAn v2.0.* Then, a coverage split was performed on the target datasets using *bbnorm* from *BBTools* package. Each dataset was separated in low (less than 3 times coverage), mid (between 3 and 10 times coverage) and high (higher than 10 times coverage). Afterwards, for each level of coverage, the datasets were again submitted to *MetaPhlAn v2.0* extracting a species list (script1). The reference genome for each specie was extracted from NCBI creating a file with each reference genome. These steps are described on S-commands1.

In order to generate the simulated datasets, the *grinder* simulator was used with the reference genome created for each cluster and coverage level. As the target datasets are paired-end reads with 2x150bp length, the average length was set to 150bp. An insert size of 584bp for cluster 1 and 597bp for cluster 2 was used in order to generate paired end reads. A median of the target datasets insert sizes were used as insert size. To model Illumina errors a 4th degree polynome was used as described by Korbel *et al* (94) :

$$3 \times 10^{-3} + 3.3 \times 10^{-8} \times i^4$$

Regarding the abundance distribution, it was used a logarithmic 1.8 distribution. The dataset size was set based on the median dataset size for each coverage level.

3.2 Tools

Table 2 Tools tested in order to generate a pipeline for metagenomic data analysis

Analysis stage		*Tools*
Pre-Processing	Error correction	*BayesHammer, Coral, Musket*
	Quality and adapter trimming	*Fastqmcf; Flexbar; Sickle; Trimmomatic*
Assembly		*Ray (v2.3.1), SPAdes 3.9*
Taxonomic composition		*MetaPhlAn v2.0, Metaxa2 (version 2.1.3)* and *Parallel-META 3.3.2*

3.2.1 Error correction

Simulated datasets were subjected to correction tools: *BayesHammer, Coral* and *Musket*. *BayesHammer* was used as part the *SPAdes* genome assembler flagged for metagenomic data. *Coral* and *Musket* were tested with the default parameters.

3.2.2 Trimming

After being subjected to the defined error correction tool, the datasets were trimmed using four different tools*: Fastq-mcf, Flexbar*, *Sickle* and *Trimmomatic*. The following adapter sequences were used with *Fastq-mcf, Flexbar* and *Trimmomatic*:

Table 3 Adapters used with the trimming tools

Adapter description	*Adapter sequence*
Nextera_circularized_duplicate_junction_adapter	*CTGTCTCTTATACACATCTAGATGTGTATAAGAGACAG*
Nextera_circularized_single_junction_adapter	*CTGTCTCTTATACACATCT*
Nextera_circularized_single_junction_adapter_reverse_complement	*AGATGTGTATAAGAGACAG*
Nextera_read_1_external_adapter	*ATCGGAAGAGCACACGTCTGAACTCCAGTCAC*
Nextera_read_2_external_adapter	*GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT*

The minimum read length was set in every tool to two thirds of the main length (100bps). Besides that*, Fasq-mcf, Flexbar* and *Sickle* were used with default parameters and *Trimmomatic* was used with the quick start as suggested by the authors.

3.2.3 Assembly

The pre-processed samples were subjected to two different assemblers: *Ray (v2.3.1)* and *SPAdes 3.9*. The k-mer values were selected for each software and dataset regarding the highest reference genomes coverage. The remaining parameters were used as default.

3.2.4 Taxonomic composition

For taxonomic analysis software selection, the pre-processed simulated datasets were subjected to *MetaPhlAn v2.0*, *Metaxa2 (version 2.1.3)* and *Parallel-META 3.3.2*.

3.2.5 Functional analysis

The functional analysis were performed in the target datasets with and without coverage split using assembled contigs. The analysis was performed with *MG-RAST* version 4.0 which gives an overview on the COG, NOG, KO and SEED databases. COG results were normalized with *Musicc* for further analysis. To assesses the assembly impact on functional analysis, both unassembled and assembled data were analysed with *MGX*, and the COG results were considered to analysis.

3.3 Evaluation strategies

3.3.1 Statistical metrics

In order to compare the contigs generated with different error correction approaches, a Python script was created to calculate the *Total number of sequences,* the *smaller* and the *larger sequence length*, the *mean* and *median sequence length*, the *N25*, *N50*, *N75* and the *GC content* (script 2).

3.3.2 Pattern matching

A pattern matching strategy was used to analyse the results achieved with the simulated datasets. The pattern matching package *MUMmer* v3.23 was used as an evaluating tool for the resultant contigs. The alignment of the contigs with the reference was performed with *NUCmer*. The resultant file was filtered with delta-filter to retain only the alignments scoring 95% minimum identity. The *dnadiff* wrapper was afterwards used to generate a report quantifying the differences between reference and query.

3.3.3 Taxonomy analysis evaluation strategy

A specific evaluation strategy was defined for the selection of the taxonomic analysis software, considering the software's different references and outputs. As the genomes used to generate the simulated datasets were selected with *MetaPhlAn v2.0*, the phylum and genus outputs from the two other tools were first checked against *MetaPhlAn v2.0* database to find any ambiguous result (script 3). With the construction of the simulated dataset, grinder generates a relative abundance percentage file (S-tables 1-6). Using this information there were created reference abundance files for phylum and genus (S-tables 7-18). Therefore, after the first triage, the outputs were compared with the reference abundance files. Two different approaches were used. The first one comprises the real taxonomical abundance and had the objective of determining the following metrics (script 4 and 4a):

- True positives (TP) – number of reads mapped to existing taxa
- False positives (FP) – number of reads mapped to non-existing taxa
- False negatives (FN) - number of reads that did not map
- Sensitivity – TP/(TP+FN)
- Positive predictive values – TP/(TP+FP)

The second analysis comprises only the positive results (TP and FP) and calculate the relative abundance log-odds scores (script 5 and 5a).

Grinder ranks are output as relative abundance percentage. In order to use this reference data in the metrics approach, the real abundance was calculated by multiplying the relative values to the number or total reads. *MetaPhlAn v2.0* also outputs the taxonomic results as relative abundance percentage. To achieve the real abundance, the relative abundance was multiplied by the number of reads mapped with *Bowtie2*, which is part of the *MetaPhlAn v2.0* analysis.

Metaxa2 and *Parallel-META* both outputs the integer number of sequences detected. Then, to use this data to calculate the relative abundance log-odds scores, it was needed to calculate this value, by diving the number of hits for a given taxa by the total number of hits.

3.3.4 Alignment and insert size calculation

In order to access the insert size to use in the simulated datasets, the real datasets were pre-processed using *BayesHammer* and *Sickle* and assembled with *SPAdes.* Then, *Bowtie2* was used with the aim to align reads against contigs, and the insert size values were achieved with *CollectInsertSizeMetrics* command from *picardtools*.

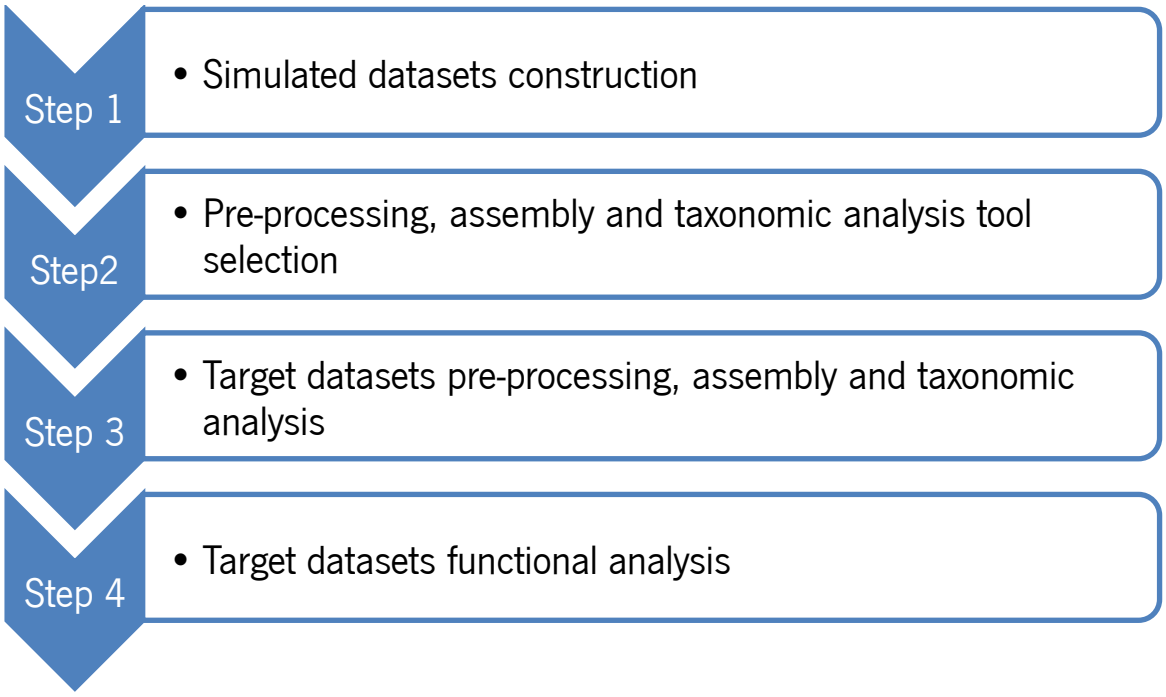
3.4 Python scripts

Scripts	*Description*	*Input*	*Output*
Script 1	Saves all species found in a list of *MetaPhlAn v2.0* output files into a txt file	LIst with *MetaPhlAn v2.0* output filenames	txt file with all species found
Script 2	Calculate the the following statistical metrics in a fasta or fastq file: number of sequences, smaller sequence length, larger sequence length, sequence length mean, sequence length median, N25, N50, N75 and GC percentage.	fasta or fastq file file type	csv with the statistical metric
Script 3	Check if the taxa found with the software tested has the same nomenclature as the reference	software name file to check	csv file with the unmatched nomenclature
Script 4	Compares the total abundance found on the given file with the abundance on the reference file and calculates true positives (TP), false-positives (FP) , false negatives (FN), sensitivity (SEN) and Positive predictive values PPV).	software name taxa (phylum or genus) software filename reference file (created using ranks from grinder) output filename	csv with metrics (TP, FP, FN, SEN, PPV)
Script 5	Compares the relative abundance found on the given file with the abundance on the reference file and calculates the log-odds scores.	software name taxa (phylum or genus) software filename reference file (created using ranks from grinder) output filename	csv with software relative abundances, reference relative abundances and log-odds score

4. Results and Discussion

4.1 Overall strategy

In order to access biological important findings regarding the different sampled data, a pipeline was defined based on simulated datasets representative of the target data. Then, the selected tools were used to analyse the target datasets, comparing both different coverage levels and assembled and unassembled data. A summary of the used strategy can be found on the following flowchart:



4.2 Step 1 – Preparation of simulated datasets

4.2.1 Preliminary analysis of datasets to be analyzed

Since the simulated datasets are intended to mimic the target data, they were generated based on the information extracted from the targets. Primarily, all the target datasets were subjected to *MetaPhlAn v2.0* to extract their taxonomic distribution. The summary of this approach is highlighted in the heat-map graph of Figure 3.

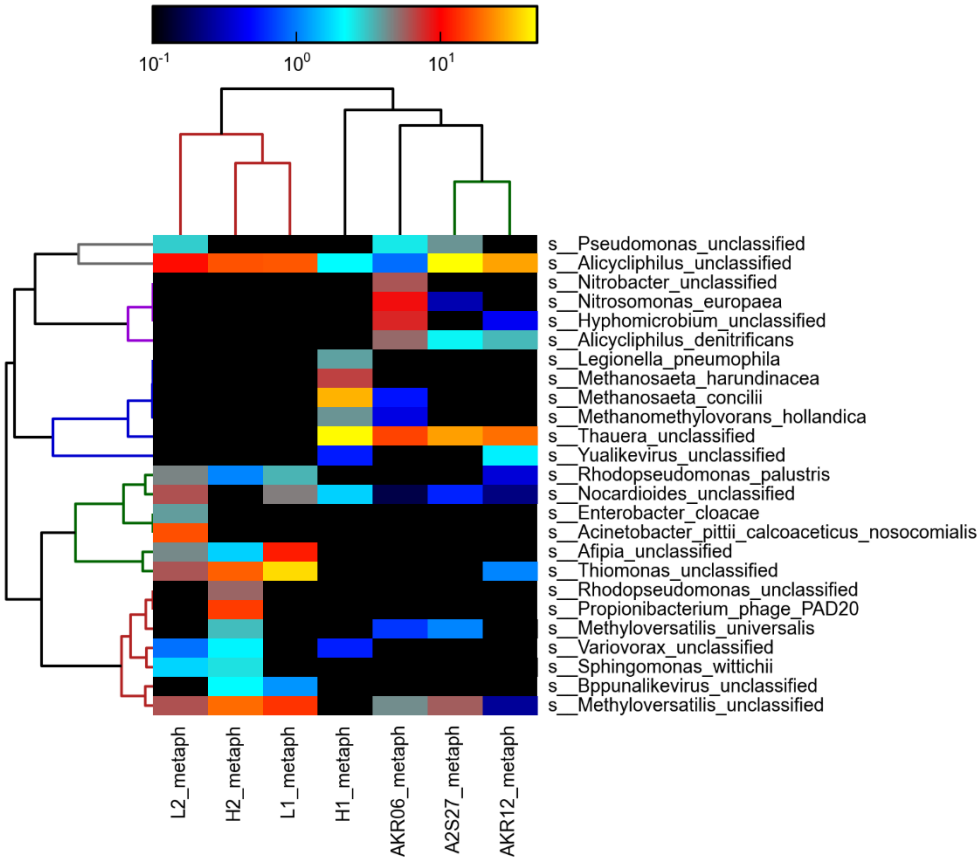


Figure 3 Abundance heat-map acquired with MetaPhlAn v2.0

The heat-map was generated based on *MetaPhlAn2* tutorial (95) using *hclust2* to get a preliminary visualization at the species level and to group the target datasets based on the species content. First, it was generated a species abundance table for every dataset. Then, a heat-map was generated with *hclust2* using the parameters suggested in the tutorial. The graph suggests two clusters, one with datasets L1, L2 and H2 and a second one formed by datasets A2, AKR06, AKR12 and H1. The first cluster comprises the petrochemical complex ETP datasets L1, L2, and H2, excepting H1 which are grouped with the CETP datasets A2, AKR06, AKR12. The CETPs from Ankleshwar (A2 and AKR12) show more similarity together than with AKR06.

The target datasets were split in three coverage levels and the species lists were accessed with *MetaPhlAn v2.0*. The genome sequences were downloaded from NCBI and grouped regarding the two mentioned clusters and coverage levels, resulting in six reference genomes files. This reference files where further used to generate the simulated datasets. In order to accomplish that, some parameters had to be defined, such as dataset size and insert size. To calculate the insert size, the target datasets were assembled with *SPAdes* and the resulting contigs aligned with reads. The insert size used on the simulated dataset is the median of the median insert sizes of the cluster datasets (as described in Table 4).

Table 4 Insert size results obtained with CollectInsertSizeMetrics command from picardtools under the alignemnet of the dataset reads against the contigs obtained with SPAdes

		MEDIAN INSERTSIZE	*MEDIAN ABSOLUTE DEVIATION*	*MEAN INSERT SIZE*	*STANDARD DEVIATION*
Cluster 1	*L1*	584	100	586.569189	168.636971
	L2	554	132	535.18377	201.494441
	H2	588	112	579.488604	185.67821
	Median	**584**			
Cluster 2	*A2*	528	121	520.620818	181.409958
	AKR06	606	111	600.380093	188.313901
	AKR12	605	108	603.180549	179.309514
	H1	589	106	583.986822	175.249703
	Median	**597**			

After splitting the target datasets, a median of the number of reads for each cluster and coverage level was calculated and further used to determine the dataset size of the simulated dataset (Table 5)

Table 5 Datasets size

		LOW reads	*MID reads*	*HIGH reads*
Cluster 1	*L1*	4 665 238	2 650 972	1 526 986
	L2	6 346 528	2 997 148	691 918
	H2	5 985 508	2 359 164	288 262
	Median	**5985508**	**2 650 972**	**691918**
Cluster 2	*A2*	3 159 936	2 079 360	3 509 352
	AKR06	7.388.670	2 387 090	658 380
	AKR12	3 228 530	3 915 734	5 125 762
	H1	3 949 704	3 094 100	3 515 064
	Median	**3 589 117**	**2 740 595**	**3 512 208**

The target datasets show different coverage distribution, except H1 which has not significant differences between the three coverage levels. Contrasting, dataset H2, corresponding to the second timepoint of H1, shows significant discrepancies in coverage distribution. This suggests a more homogenous abundance distribution in H1 comparing with H2. Nevertheless, a strange behaviour of H1 was observed on further analysis, suggesting that this dataset may have surfer some erroneous treatment between sampling and sequencing. H2 LOW reads have more than 20 times the reads of H2 HIGH reads, suggesting that H2 comprises a large number of low represented microorganisms and few predominant organisms in the sample. However, high coverage reads could also be related with conserved reads throughout the organisms, being all grouped together, instead of one or few prevalent organisms.

4.3 Step 2 – Pre-processing, assembly and taxonomic analysis tool selection

4.3.1 Error correction

- Pattern matching results

To evaluate the performance of error correction tools, the simulated datasets were treated with three different tools: *BayesHammer*, *Coral* and *Musket*. Therefore, the corrected reads were assembled with *SPAdes* and the resulting contigs were aligned with the reference genomes using *nucmer* (Table 6).

Table 6 Simulated datasets subjected to different error correction approaches followed by assembled with SPAdes. The treated reads were then aligned against the reference genomes.

		None	*Bayes hammer*	*Coral*	*Musket*
LOW1	Aligned bases	79822203 (91.16%)	78924749 (90.14%)	79900802 (91.25%)	78125117 (89.23%)
	Total SNPs	375100	310829	373674	422900
	Total Indels	82205	72903	23970	85056
MID1	Aligned bases	7177418 (99.51%)	7178535 (99.53%)	7175977 (99.49%)	7178640 (99.53%)
	Total SNPs	227	258	1664	347
	Total Indels	11	6	110	13
HIGH1	Aligned bases	4172038 (99.22%)	4172971 (99.25%)	4172414 (99.23%)	4172969 (99.25%)
	Total SNPs	135	133	271	188
	Total Indels	3	2	6	6
LOW2	Aligned bases	103279138 (43.35%)	104277628 (43.77%)	103701472 (43.53%)	103228102 (43.33%)
	Total SNPs	1263320	1167508	1266348	1264819
	Total Indels	297702	293478	298355	297634
MID2	Aligned bases	41880429 (94.22%)	41628399 (93.65%)	41843238 (94.14%)	41841606 (94.13%)
	Total SNPs	166588	131276	165676	168294
	Total Indels	39598	35254	39513	39681
HIGH2	Aligned bases	7270580 (99.30%)	7272301 (99.32%)	7272641 (99.33%)	7267666 (99.26%)
	Total SNPs	273	337	408	340
	Total Indels	45	40	38	40

Regarding the percentage of the reference aligned bases with the contigs, there is no clear difference between the different tools. However a slight improvement can be observed when comparing with the untreated data. This may suggest that the error correction tools may be dismissed for some datasets or when using assemblers which are less affected by the sequencing errors.

- Statistical Analysis

With the purpose of evaluating the impact of error correction in the data assembly, statistical metrics were calculated for the contigs which have been subjected to the mentioned error correction tools (S-table 19). The N50 is a statistical metric that indicates an average contig length such that 50% of the sum of the total length of contigs are achieved in the contig of this size or larger. It is commonly used to analyse the data contiguity and the results can be found in Figure 4.

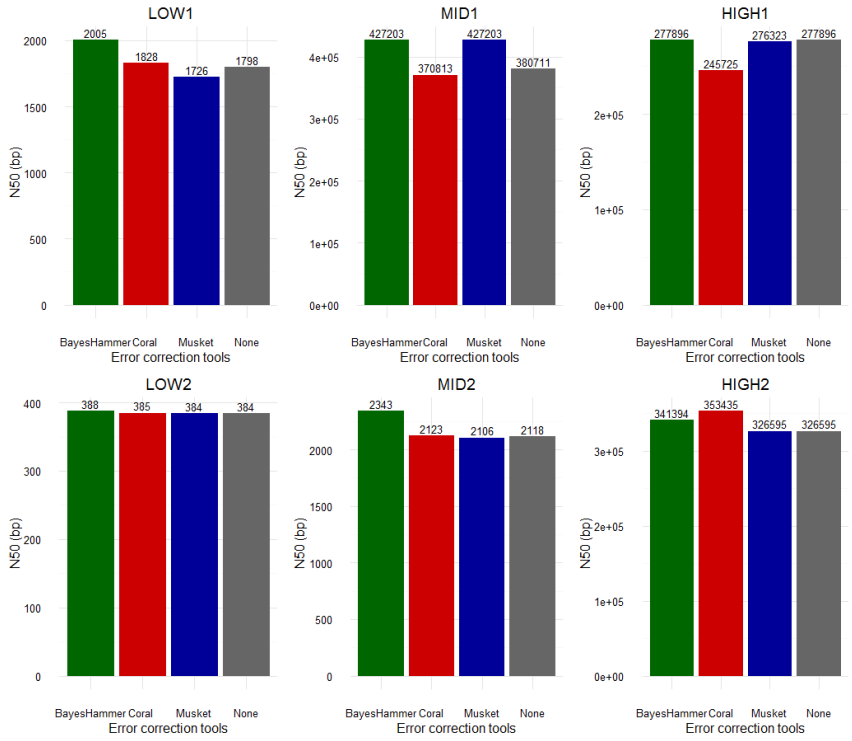


Figure 4 N50 value for contigs obtained by subjecting the simulated datasets to different error correction approaches followed by assembled with SPAdes.

Taking into consideration the described results, *BayesHammer* seems to have stronger results by reference genomes coverage, where it achieved good alignment results both regarding the number of aligned bases and the reduced snps and indels. In terms of contiguity, a high N50 value were also achieved on data corrected with *BayesHammer.* Therefore, *BayesHammer* was chosen to be used in the following analysis steps.

4.3.2 Trimming

- Pattern matching results

In order to assess the trimming tool's performance, the reads were first corrected with *BayesHammer* and further subjected to four different tools: *Fastq-mcf*, *Flexbar*, *Sickle* and *Trimmomatic*. Then, the error corrected and trimmed reads were assembled with *SPAdes* and the resulting contigs were aligned with the reference genomes using *nucmer*. Results regarding the number of aligned bases, total SNPs and total indels can be found in Table 7.

Table 7 Simulated datasets subjected to BayesHammer error correction followed by different trimming approaches and assembled with SPAdes. The treated reads were then aligned against the reference genomes.

		None	*Fastq-mcf*	*Flexbar*	*Sickle*	*Trimmomatic*
LOW1	Aligned Bases	78924749 (90.14%)	77933403 (89.01%)	77450098 (88.46%)	77933137 (89.01%)	77941102 (89.02%)
	Total SNPs	310829	381591	365472	381781	380944
	Total Indels	72903	95676	92029	95681	95512
MID1	Aligned Bases	7178535 (99.53%)	7174085 (99.47%)	7174123 (99.47%)	7174085 (99.47%)	7174126 (99.47%)
	Total SNPs	258	198	200	198	199
	Total Indels	6	7	10	7	7
HIGH1	Aligned Bases	4172971 (99.25%)	4171007 (99.20%)	4170912 (99.20%)	4171007 (99.20%)	4171007 (99.20%)
	Total SNPs	133	129	131	129	129
	Total Indels	2	1	5	1	1
LOW2	Aligned Bases	104277628 (43.77%)	77239246 (32.42%)	75250926 (31.59%)	77239306 (32.42%)	77378103 (32.48%)
	Total SNPs	1167508	1036219	952099	1036188	1036021
	Total Indels	293478	257052	236525	257061	257000
MID2	Aligned Bases	41628399 (93.65%)	40719842 (91.61%)	40520218 (91.16%)	40720622 (91.61%)	40729435 (91.63%)
	Total SNPs	131276	169176	161833	169132	168622
	Total Indels	35254	45729	44076	45727	45658
HIGH2	Aligned Bases	7272301 (99.32%)	7269284 (99.28%)	7268845 (99.27%)	7269284 (99.28%)	7269284 (99.28%)
	Total SNPs	337	452	450	452	427
	Total Indels	40	44	51	44	44

The results obtained suggested that the trimming tools have actually reduced assembly quality, as the number of aligned bases decreased with the trimming treatment. However, it has to be taken into account that the simulated datasets have only generated two levels of quality score, a high (30) and a low (10) score. Sequencing quality scores are given by the sequencer as the base call accuracy in Phred format which uses a set of ASCII characters. The present datasets uses Phred+33 with a specific ASCII encoding,

comprising scores from 0 to 41. Then, the quality score generated for the simulated datasets, comprising only 2 values, instead of the real data quality spectrum may create a bias in the trimming results.

- Statistical Analysis

Statistical metrics were also calculated in order to evaluate the impact of the different trimming tools (S-table 20). Again, the trimming tools don't seem to improve the results obtained with the simulated datasets, being once again worse than the untreated data. Taking into account that the simulated datasets do not perfectly mimic the sequencing quality scores of a real dataset, a statistical analysis was performed on the (not split) target datasets which were first corrected with *BayesHammer*. The statistical analysis results can be found as supplementary data (S-table 21) and N50 values are plotted in Figure 5.

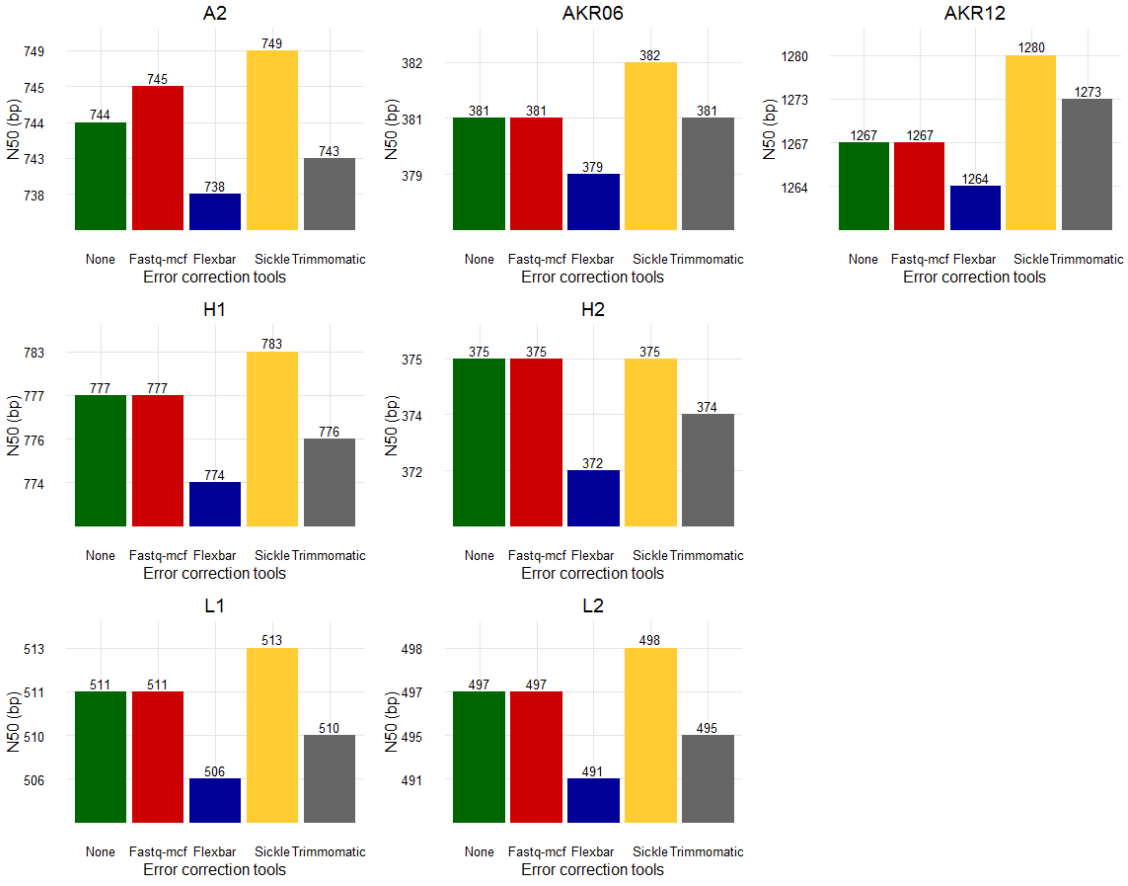


Figure 5 N50 value for contigs obtained by subjecting the target datasets to Bayes Hammer error correction followed by different trimming approaches and assembled with SPAdes

The analysis of the trimming tools was not conclusive, since the simulated datasets results do not express a notorious data improvement with any of the tested tools. However, in the target datasets *Sickle*, followed by *Fastq-mcf*, seems to have a slightly improvement into the contiguity of the assembled reads.

Since *Sickle* do not trim adapters, which could be present in the target datasets, *Fastq-mcf* was chosen as the tool to be used in the following analysis steps. *Fastq-mcf* had a consistent performance both in the pattern alignment analysis and in contiguity.

4.3.3 Assembly

- Pattern matching results

To select an assembly tool and to define the k-mer values, *SPAdes* and *Ray* were tested in every pre-processed simulated datasets with different k-mer values. The tool and k-mer value was selected regarding the percentage of the reference dataset genomes aligned with the contigs, acquired with *nucmer* (S-tables 22 and 23). The best results obtained with both tools are shown in Table 8.

Table 8 Pre-processed simulated datasets subjected to two different assemblers. The resulted contigs were then aligned against the reference genomes.

	Assembly tool	*k-mer size*	*Aligned Bases*
LOW1	spades	default	89.01%
	ray	25	73.42%
MID1	spades	default	99.47%
	ray	33	99.02%
HIGH1	spades	default	99.20%
	ray	29	99.52%
LOW2	spades	25	42.95%
	ray	25	16.87%
MID2	spades	default	91.61%
	ray	25	75.53%
HIGH2	spades	default	99.28%
	ray	31,33,39-47	98.74%

Concerning *SPAdes*, there were selected the second bests results for MID1 and HIGH1, since the difference between the two better results were only 0.01% and the second best result was acquired with the default k-mer length which suggest a stronger k-mer size selection when applied to the target datasets.

- Statistical Analysis

The assemblies were also evaluated with statistical metrics (S-table 24). The N50 values suggest that a higher contiguity is achieved with *SPAdes* (Figure 6).

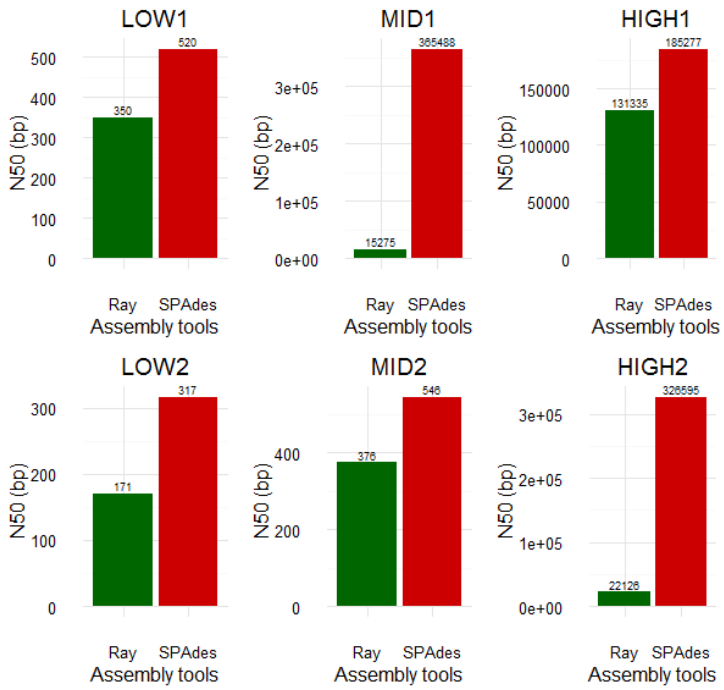


Figure 6 N50 value for contigs obtained using Ray and SPAdes assemblers

Besides *Ray* having a slightest better result for HIGH1 dataset, *SPAdes* shown to be a stronger tool and therefore it will be used on the target datasets with default k-mer size for every datasets except for the ones corresponding to LOW2, where it will be used a k-mer size equal to 25. The k-mer size selection has a great impact in assembly, since a smaller value may lead to collapse more repeated areas together. However, when using larger k-mer on low coverage regions may lead to prevent the detection of overlaps between reads. When using the *SPAdes* with default k-mer size, the tool performs the assembly with 21, 33 and 55 k-mer and combines the best results.

4.3.4 Taxonomy

Concerning the evaluation of the taxonomic tools, the simulated datasets were pre-processed with *BayesHammer* and *Fastq-mcf* and analysed with *MetaPhlAn v2.0, Metaxa2* and *Parallel-meta*. The number of true positives (TP), false positives (FP) and false negatives (FN) were assessed being the sensitivity and positive predictive values calculated (S-tables 25-26). Genus level performance metrics can be found in Figure 7. *MetaPhlAn v2.0 retrieved* the higher number of true positives, with the exception of the HIGH1 dataset. It did not retrieve any false positives or false negatives on the MID and HIGH datasets being the best tool regarding this metrics on LOW2 dataset. *MetaPhlAn v2.0 had* a slight higher value than *Metaxa2* regarding false positives in the LOW1 dataset. *Parallel-meta* found more *genus* when

compared with the other tools which can be understood as a better performance in a superficial analysis. This tool; however, retrieves a large number of false positives which should be accounted.

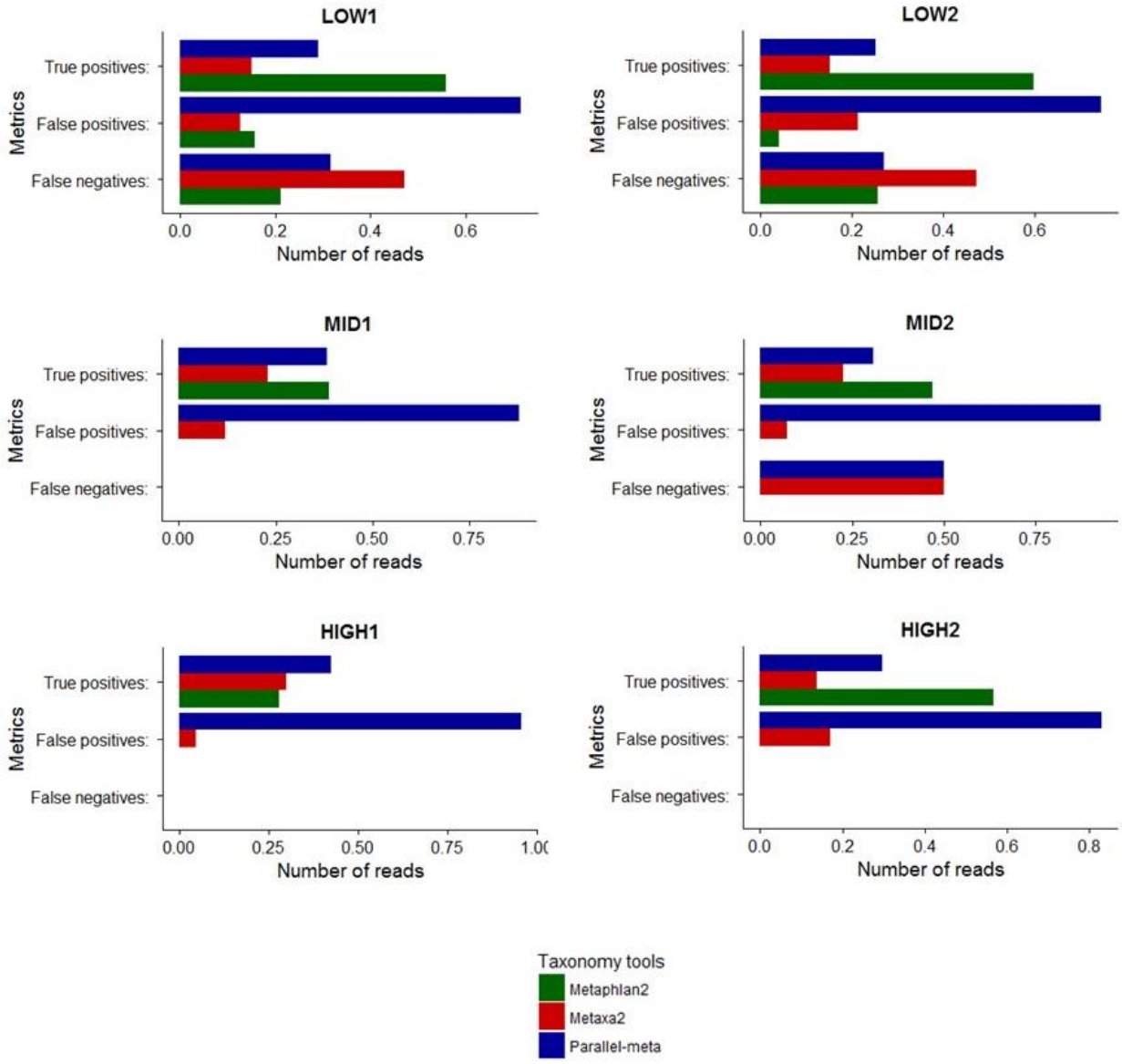


Figure 7 Genus level metrics' relative frequencies for the taxonomic analysis tools.

Expected relative abundances (generated by grinder) and the relative abundances given by the tools were also compared and can be found as supplementary data (S-table 27-63).

Regarding this results, *MetaPhlAn v2.0* was chosen as the taxonomic tool to be used. The assembled datasets were also subjected to the analysis but no taxa were identified. For that reason, the target datasets taxonomy analysis were only performed on unassembled data.

4.4 Step 3 - Target datasets pre-processing, assembly and taxonomic analysis

4.4.1 Pre-processing and Assembly

The present work comprises the metagenomic analysis of 7 datasets, 3 coming from CETPs and 4 from a petrochemical complex ETP. All the datasets were analysed with and without coverage split (as described for the simulated datasets). The data was previous pre-processed with *BayesHammer* and *Fastq-mc*f and assembled with *SPAdes*. All the datasets (with and without coverage split) where assembled with *SPAdes* default k-mer size. Exception was made for the low coverage fraction of A2, AKR06, AKR12 and H1 datasets which correspond to the simulated LOW2, where it was used a k-mer size equal to 25. Statistical metrics comparing the assembled datasets with and without coverage split can be found in Table 9.

The highest sequencing depth of coverage showed a higher continuity in the majority of the datasets. In contrast, the low coverage datasets fraction exhibited a lower continuity. This may be due to an erroneous sequencing depth split. When grouping the reads with higher coverage it is possible that instead of grouping organisms with high representability in the sample, there were grouped the conservative and repetitive genomic areas of several organisms. This sequences will easily assemble due to its similarities and to repetitive areas. Taking this into account, the high contiguity may be due to the chimeras' formation between similar sequences and not to a real genome assembly.

This further highlights the fragility of evaluating an assembly regarding the contiguity. Contiguity is here described using N50 metric and this and other metrics are often used to access the assembly quality. However, a high contiguity may be related with the formation of chimeras and the exclusion of important specific genomic areas.

The low coverage datasets here present can be constituted of specific genomic areas from different organisms, lacking the more conservative genomic areas and thus hindering the assembly.

Table 9 Statistical metrics comparing target datasets with and without coverage split.

		Number of sequences:	*Smaller sequence length:*	*Larger sequence length:*	*Sequence length mean:*	*Sequence length median:*	*N25:*	*N50:*	*N75:*	*GC percentage:*
AKR06	without coverage split	27273	1001	175509	2712.922	1608	8079	3480	1699	60.15%
	high coverage	1642	1003	66625	2758.313	2001	5582	3206	1950	66.89%
	mid coverage	13868	1001	64137	2767.372	1961	6027	3311	1946	60.50%
	low coverage	4734	1001	3972	1282.218	1188	1476	1233	1096	59.98%
A2	without coverage split	20543	1001	902269	3127.553	1626	17667	4383	1863	59.64%
	high coverage	2044	1001	566836	4538.727	1679	119034	15229	2550	54.82%
	mid coverage	10970	1001	60653	2875.108	1809	7941	3643	1895	61.32%
	low coverage	3750	1001	3223	1267.182	1182	1449	1226	1096	59.37%
AKR12	without coverage split	27661	1001	828132	3969.401	1683	34536	8226	2401	64.10%
	high coverage	2782	1001	398734	6384.213	1797	65775	25220	6077	62.72%
	mid coverage	16066	1001	103178	3799.422	2119	12655	5946	2618	65.31%
	low coverage	4341	1001	3993	1272.481	1186	1457	1234	1098	62.11%
L1	without coverage split	31010	1001	821816	2767.616	1555	16466	3239	1653	60.00%
	high coverage	1194	1001	188621	7364.967	2328.5	69438	19083	7214	64.39%
	mid coverage	12506	1001	197627	3049.36	1639.5	17465	4013	1802	58.56%
	low coverage	4369	1001	3627	1249.721	1169	1422	1206	1084	60.47%
L2	without coverage split	34255	1001	919053	2596.173	1542	8713	2995	1591	63.02%
	high coverage	2492	1001	16443	1804.557	1533.5	2628	1838	1377	67.19%
	mid coverage	14496	1001	124198	2994.274	1723	11671	3860	1860	63.25%
	low coverage	4939	1001	3458	1247.945	1172	1413	1209	1087	60.67%
H1	without coverage split	22899	1001	623752	3671.049	1622	25893	7170	2164	58.65%
	high coverage	2741	1001	352085	5263.494	2322	56684	9434	3850	62.71%
	mid coverage	12445	1001	79582	3893.632	2144	14499	6098	2657	58.21%
	low coverage	4247	1001	4148	1290.954	1192	1506	1245	1100	56.92%
H2	without coverage split	18100	1001	104373	2410.085	1503	7182	2707	1503	64.75%
	high coverage	757	1001	36395	1750.651	1450	2450	1709	1293	65.25%
	mid coverage	10047	1001	109432	3633.44	1815	18697	6086	2241	65.24%
	low coverage	3598	1001	3432	1231.968	1154	1384	1189	1077	61.70%

The coverage split is of great importance, since, as also shown by this results, the presence of low coverage reads difficult the assembly and may lead to the exclusion of less abundant features and organisms. This means that a less abundant organism or a specific function with lower representability can be excluded due to the supra representation of other species and more common genomic areas. Nevertheless, the coverage split approach has to be better studied in order to accurately split different organismal abundances, instead of splitting a unique genome in two different datasets.

4.4.2 Taxonomy

The taxonomy was assessed with *MetaPhlAn v2.0,* being the reads pre-processed with *BayesHammer* and *Fastq-mcf*. The datasets were analysed with and without coverage split. Since the main objective of the coverage split was the assessment of low coverage species information, it was expected that the split data could detect the lowest abundant organisms, translating in a higher diversity. However, this was not observed in every dataset. AKR06 and H2 retrieved lower number of found genus in the data split by coverage. Taking this into account, the genus appearing in both split and not split datasets were considered more robust.

AKR06 dataset was collected from an industrial area CETP at Jeedimetla. The most abundant genus found in this dataset was *Thauera. Thauera* is the most, or the second most abundant genera in CETP datasets, as well as in timepoint 1 of the High TDS activated sludge dataset (H1). *Thauera* strains are found on activated sludge systems that are used for the treatment of wastewater and have a denitrifying function. *Thauera* species were found to be capable of selenite, ammonium and humus reduction under anaerobic conditions (96,97). Other denitrifying populations such as *Methyloversatilis*, *Hyphomicrobium*, *Azoarcus* and *Paracoccus* strains were also found (96,98). *Pseudomonas* strains are also associated with denitrification; however, this genus is also associated with opportunistic pathogens (99).

Alicycliphilus, another abundantly found genus, is also found on activated sludge systems (100,101). These organisms grow under aerobic or anoxic conditions having a strictly oxidative metabolism (102). *Alicycliphilus* species have been shown to be capable of degrading high-strength chemical compounds, such as cyclohexanol, benzene, and acetone (103–107).

Nitrosomonas and Nitrobacter species are nitrifying bacteria, the first convert's ammonia to nitrite and the second convert's nitrite to nitrate. These two genera are abundant in AKR06 dataset. Nitrospira genus, also found on AKR06 dataset, comprises bacteria capable of oxidize completely ammonia to nitrate, named complete ammonia oxidizers (Comammox) (108). The high abundance of *Nitrosomonas* and *Nitrobacter* species and the presence of *Nitrospira* genus, together with the mentioned denitrifying population may suggest a higher concentration of ammonia and other nitrogen cycle related compounds on this dataset.

The genus *Methylocystis* was also found on AKR06 dataset. *Methylocystis* species are methanotrophic bacteria with capability to use different nitrogen sources (109). The presence of this microorganism may be related with a high level of methane in this dataset.

Leucobacter can be found in a variety of environments and most members of Leucobacter genus are chromate-resistant suggesting a higher concentration of this salt on AKR06 (110,111)

Nocardioides species were found to be capable of p-nitrophenol, crude oil and 2,4 dinitroanisole degradation (112). Moreover, a denitrifying specie was found on sludge in a sewage-disposal plant (113). This genus is present in every dataset, except in the high TDS activated sludge (H1, H2).

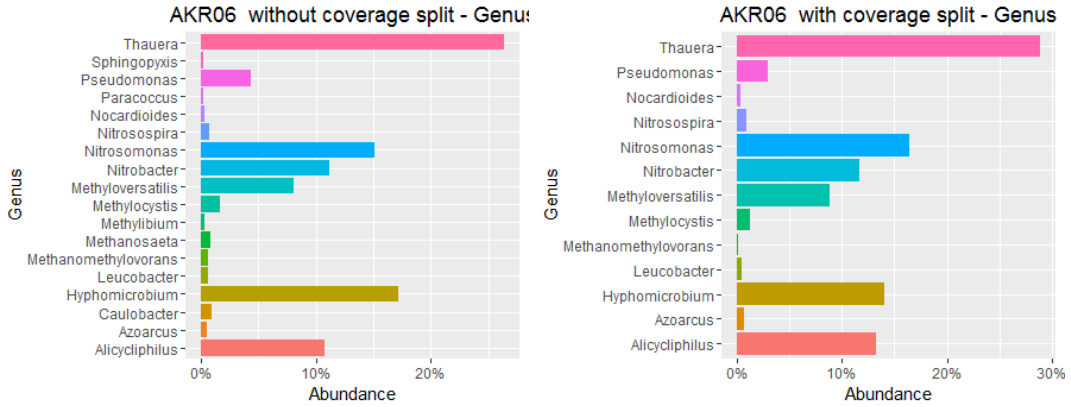


Figure 8 AKR06 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without coverage split

Datasets A2 and AKR12 are from two CETP at Ankleshwar city. This CETPs receive mainly the residues dyes and textile industries. *Thauera* and *Alicycliphilus* are the two most abundant genus found in both A2 and AKR12 datasets.

Arcobacter genus was found on A2. *Arcobacter* species have been found in different environments. In humans, *Arcobacter* genus can act as a pathogen, originating diarrhea and other pathology outbreaks. The infection is usually originated by contaminated food or water (114).

The genus *Rhodococcus* was identified on A2 dataset. This genus is frequently associated with bioremediation and biocatalytic processes. *Rhodococcus* species can degrade a large number of compounds, such as steroids, nitriles, lignins, and organosulfur. Moreover, *Rhodococcus* spp. were associated with human, animal and plant infection (115). *Acidovorax* was also found on A2 dataset, its species can be found in soil and water habitats and some are phytopathogenic. A specie with denitrifying properties was also identified (116).

Bordetella genus, a strictly aerobic organism, that cannot ferment carbohydrates such as glucose, was found on AKR12 dataset. Most members are primary or opportunistic pathogens. *Bordetella, as well as Pseudomonas, Mesorhizobium, Pusillimonas*, were found to be capable of degrading crude oil sludge (117,118). *Mesorhizobium* and *Pusillimonas* were also found on AKR12 dataset, which may suggest a significant amount of crude oil present in this CETP in Ankleshwar, coming from dyes and textile industries.

The genus Sphingopyxis was found on AKR12 dataset and comprises strictly aerobic, chemoheterotrophic, propane-oxidizing bacteria that have been identified in different environments, including activated sludge (119,120).

Oligotropha genus was also detected on AKR12 dataset. This genus only comprises one specie, the *O. carboxidovorans,* a chemolithoautotrophic bacteria able to use CO, CO_2, and H_2 that was found on wastewater (121). It was also found, although in low abundance, the genus *Caulobacter,* which are strictly respiratory and aerobic and comprises generally aquatic species (122). A few pathogenic cases have been identified related with meningitis (123).

Thiomonas strains are capable of oxidizing arsenite and are found ubiquitously in acid mine drainage which have extreme conditions due to the many lethal elements, low levels of organic matter and low pH (124). This genus is highly present in the samples collected from the petrochemical complex, with exception of H1, suggesting a hostile environment, particularly on L1 where it is the predominant genus. *Thiomonas* strains were also found on AKR12 dataset, albeit in lower abundance.

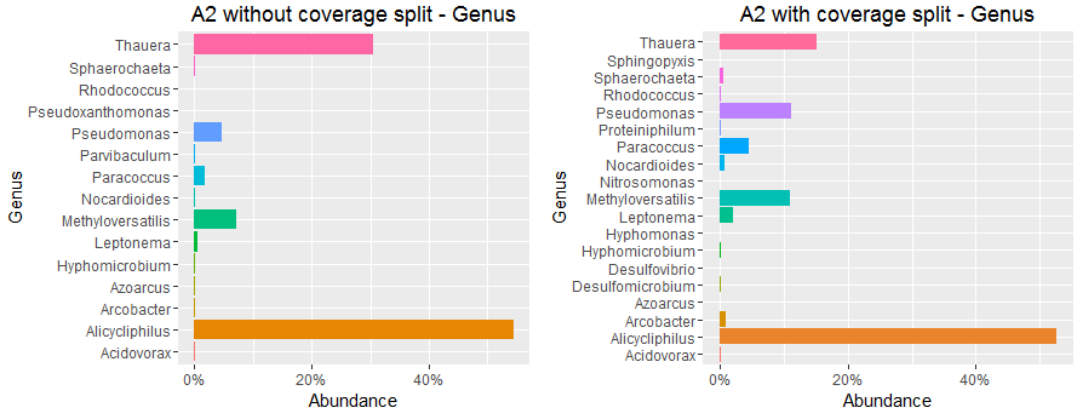


Figure 9 A2 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without split by coverage

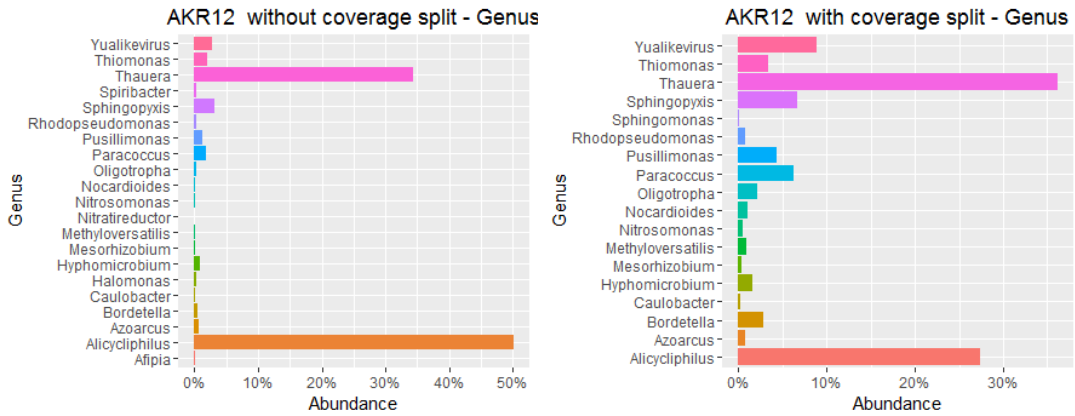


Figure 10 AKR12 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without split by coverage

Regarding low TDS activated sludge datasets, an increase in biodiversity can be observed between timepoint 1 and 2, suggesting a transition to a least selective environment. The higher toxicity of the L1 dataset sample can be also observed, as referred, by the predominance of *Thiomonas* strains. *Bppunalikevirus,* also named *Bpp1virus*, was found on L1 dataset but not on L2 dataset. *Bpp1virus* genus

comprises two species, one is a *Bordetella* phage found to display a marked tropism, and the other is a *Burkholderia* phage (125). None of this bacterium was yet found on this dataset.

The most abundant genus found on the timepoint 2 dataset (L2), *Acinectobacter*, was not found on L1. This genus comprises a variety of species, many of them being opportunistic pathogens in humans, which may be accounted for the effluent treatment process (126).

Although in small percentage, the genus *Variovorax* was also detected on both low and high TDS activated sludge datasets. *Variovorax* species were shown to be capable of diuron and linuron mineralization, two phenylurea herbicides, be involved on benzene degradation, and on denitrification (127–129).

Sphingomonas, a chemoheterotrophic, strictly aerobic bacteria that has been identified in distinct environments, was also found on L2 (130).

Afipia genus was found on both L1 and L2 and also on H2. *Afipia* species are capable of degrading haloacetic acids which are disinfection byproducts formed during the chlorination and chloramination of drinking water. Since the consumption of haloacetic acids has been linked to human health risks, these bacteria can have a significant impact in reducing the concentrations of these compounds in drinking water (131). However, *Afipia* species have also been described to be opportunistic pathogens and to be *Legionella* like amoebae pathogens related with pneumonia cases (132,133). Therefore, the presence of *Afipia* genus has to be considered for the effluent treatment process.

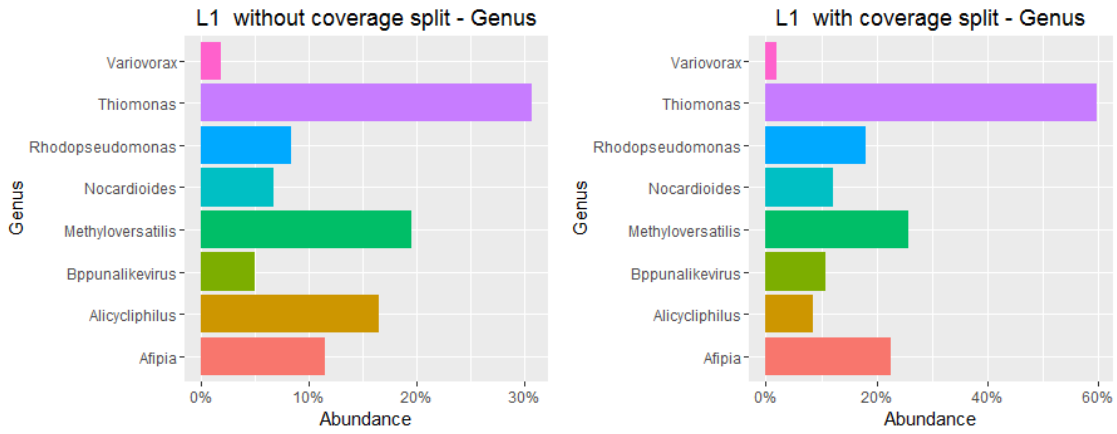


Figure 11 L1 dataset taxonomy analysis (131)*performed by MetaPhlAn v2.0 with and without split by coverage*

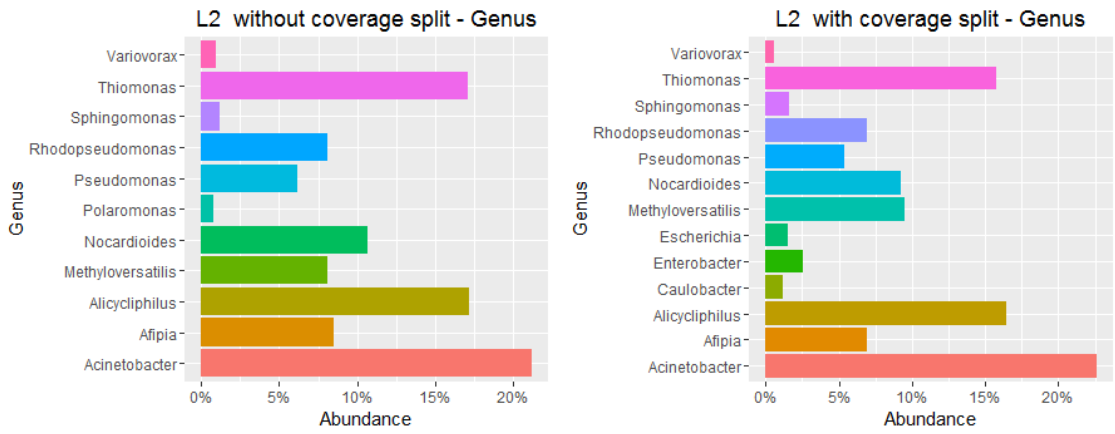


Figure 12 L2 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without split by coverage

Concerning the high TDS activated sludge datasets, a notable difference can be observed between timepoint 1 (H1) and timepoint 2 (H2). *Thiomonas* genus, which are related with extreme conditions was expected to be present on timepoint 1; however, this genus was not found on H1. The two most abundant genus found on H1, *Thauera* and *Methanosaeta*, were not found on H2. This unexpected behaviour of H1 dataset should be confirmed with replicate analyses. *Methanosaeta* strains are acetoclastic methanogens which can suggest a higher acetate concentration on timepoint 1 compared with timepoint 2 (134). There were also found two different phages with low abundance: *Yulikevirus* on H1 and *Bppunalikevirus* on H2.

Legionella was found on H1 dataset. These bacteria are opportunistic pathogens that can occur on tap water. Fortunately, this genus did not appear on the timepoint 2 (H2). However, afipia genus, which comprises *Legionella* like amoebae pathogens was found on dataset H2.

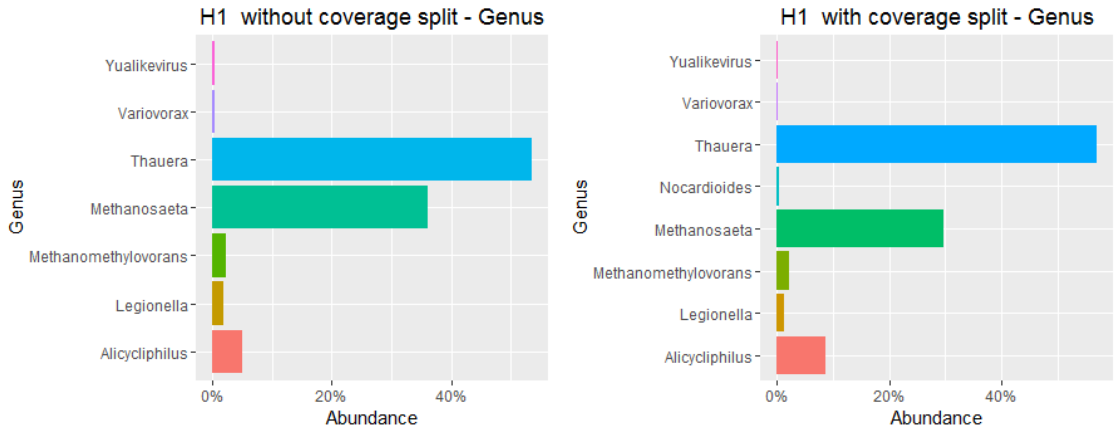


Figure 13 H1 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without split by coverage

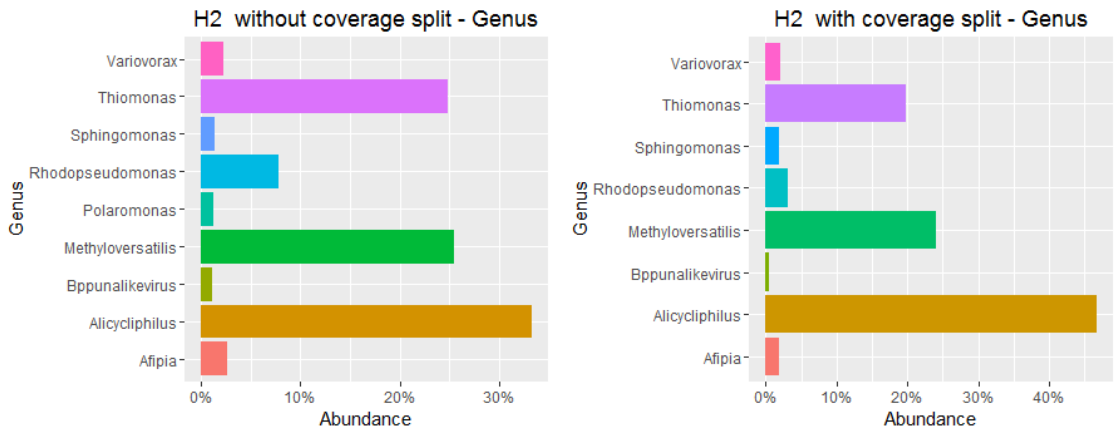


Figure 14 H2 dataset taxonomy analysis performed by MetaPhlAn v2.0 with and without split by coverage

A rarefaction curve was constructed in order to assess if there were enough observations to get a reasonable estimate of the quantity measured in each sample. As shown on Figure 15, the quantity of genus identified in each sample, with exception of A2, had converged on a good estimate of the correct value. This is, A2 curve increases as more sequences are added, indicating that a more extensive sampling should be performed for this dataset. The remaining datasets' curves reached a horizontal asymptote, suggesting that the quantity of genus found is a good estimate of the real value of genus existing in the sample.

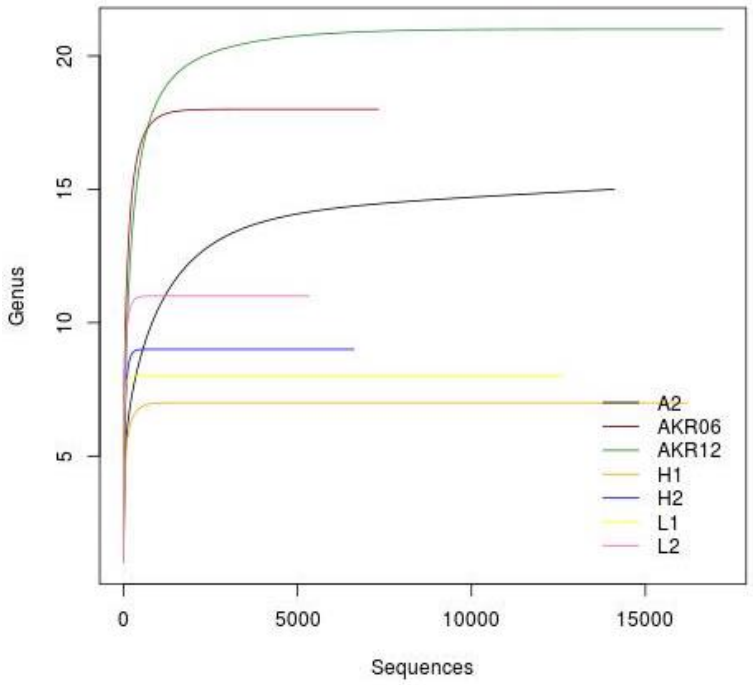


Figure 15 Rarefaction curve constructed with results without coverage split.

The previous results showed a higher biodiversity in the CETPs (AKR06, A2 and AKR12) compared with the petrochemical complex ETPs (L1, L2, H1, H2). Also, a prevalence of the *Thauera* genus in the CETPs datasets against the *Thiomonas* genus prevalence in the ETPs, suggests a more hostile environment on the petrochemical complex datasets.

A hierarchical cluster was also performed using the results without coverage split. The graph has a similar behaviour as the heat-map created to separate the data in two clusters in order to generate the simulated datasets (Figure 3).

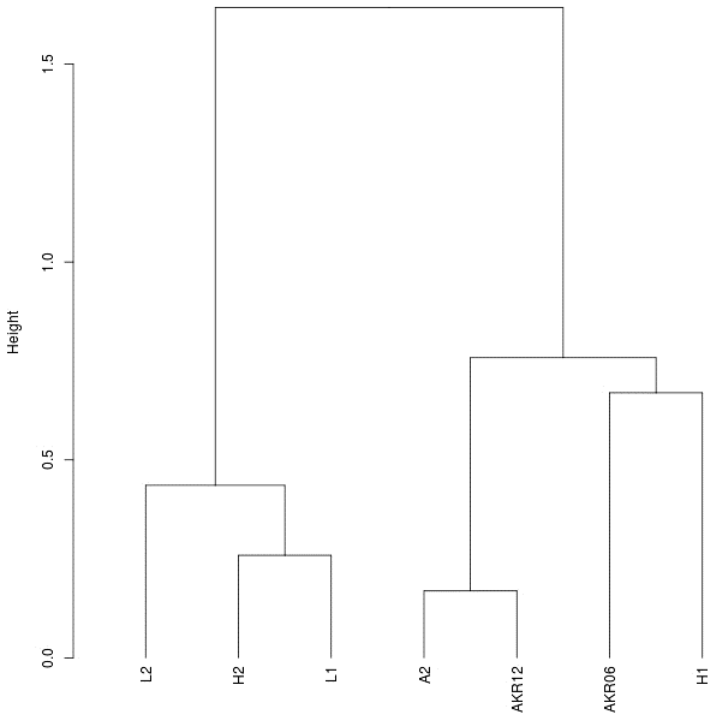


Figure 16 Hierarchical clustering for taxonomic results without coverage split

There were observed two clusters, the first one including the datasets from petrochemical complex ETPs and the second including the datasets from CETPs. Exception made for H1, which shows again an unexpected behaviour, being clustered with AKR06. Once more, CETPs from Ankleshwar (A2 and AKR12) show more similarity between each other than with AKR06. The second timepoint of high TDS (H2) shows more similarity with the first timepoint of low TDS (L1), suggesting a distinct microbial content regarding the level of dissolved solids.

4.5 Step 4 - Target datasets functionality analysis

The assessment to the genetic functions present in the metagenomes was performed with *MG-RAST* using the assembled datasets. An error occurred and the results regarding the AKR12 dataset weren't available for analysis.

As performed in taxonomy, there were analysed datasets with and without coverage split. The summary functions of COG, NOG, KEEG and SEED from dataset AKR06, showing the differences between coverage split and not split data, are shown as example (Figure 17-Figure 20). The remaining data can be found on supplementary data (S-figures 1-20). The split data results were summed in order to allow an easier visualization.

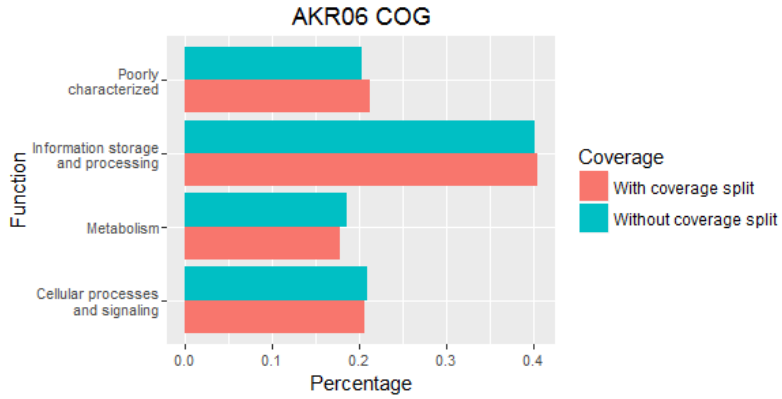


Figure 17 COG summary functions for AKR06 dataset with and without coverage split.

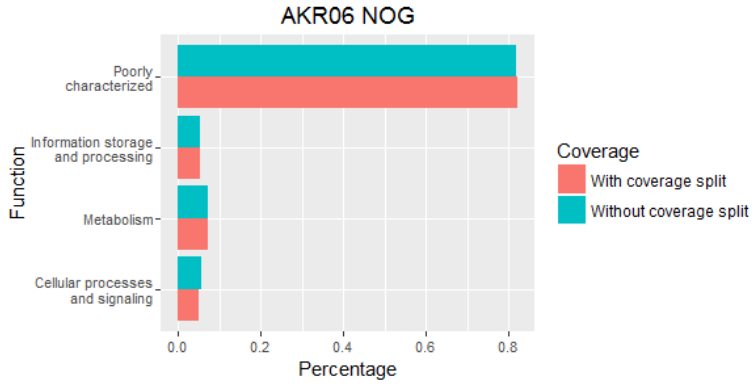


Figure 18 NOG summary functions for AKR06 dataset with and without coverage split.

COG and NOG were summarised by MG-RAST in four different categories: "poorly characterized"; "information storage and processing", which was the most predominant function in COG results; "metabolism" and "cellular processes and signalling". NOG results were not taken into consideration to further analysis since there were mainly poorly characterized results and also because it is not a supervised nor annotated database. The generic function results don't show a difference between split and not split data for both COG and NOG.

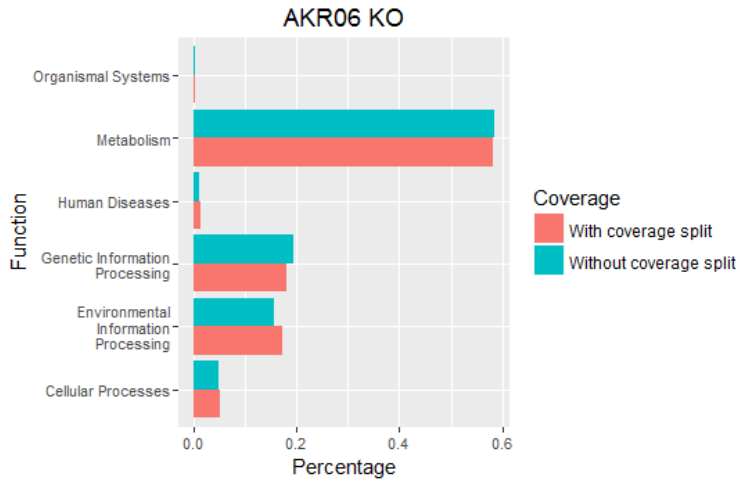


Figure 19 KO summary functions for AKR06 dataset with and without coverage split.

KO functions were split into six different categories, being the "metabolism" the most common function group found. SEED classification level 1 is used by MG-RAST summary analysis. The most common function was the "clustering based subsystems" followed by "carbohydrates". Again, a difference between split and not split data was not noticeable.

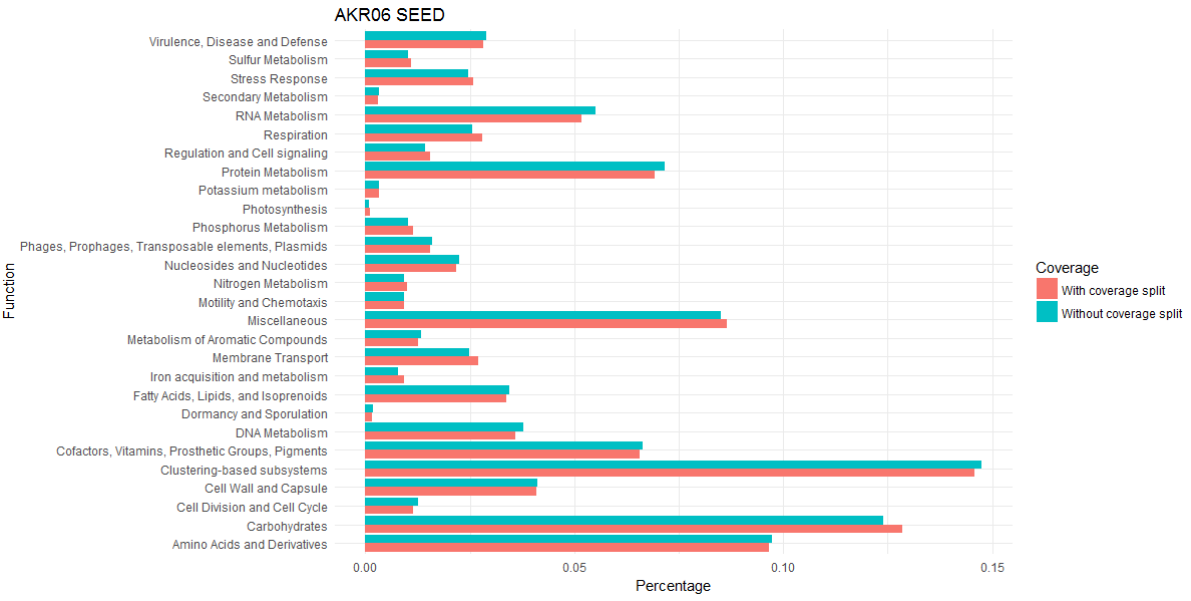


Figure 20 SEED functions for AKR06 dataset with and without coverage split.

After the summary overview displayed by *MG-RAST*, a more accurate analysis was performed with COG database. Before analysis, an inter-sample normalization was performed with *Musicc* software in order to get the abundance of each gene in the microbiome. Results from COG normalization can be found as supplementary data (S-table 64).

The 10 more common COGs found on each dataset can be found in Tables 10 and 11. The most abundant COG in every dataset, except H2, (COG1028) is described as a "NAD(P)-dependent

dehydrogenase, short-chain alcohol dehydrogenase family". Short-chain alcohol dehydrogenase are enzymes of great functional diversity, in addition, this COG is assigned to four different functions ("lipid transport and metabolism", "secondary metabolites biosynthesis", "transport and catabolism" and "general function prediction only"). Besides the general ambit of this COG and the presence in every dataset, it was noted a slight increase in this function was observed in the second timepoint of both low and high TDS datasets from petrochemical complex (L1/L2 and H1/H2).

Table 10 Top 10 most abundant COGs found on CETP datasets

Dataset	*COG*	*Function*	*COG description*	*Abundance*
AKR16	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	15,87
	COG0642	T	Signal transduction histidine kinase	11,68
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	10,52
	COG0477	GEPR	MFS family permease	9,36
	COG0500	QR	SAM-dependent methyltransferase	8,59
	COG0841	V	Multidrug efflux pump subunit AcrB	8,05
	COG1012	C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	8,01
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	7,82
	COG1132	V	ABC-type multidrug transport system, ATPase and permease component	7,28
	COG1136	M	ABC-type lipoprotein export system, ATPase component	7,12
A2	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	17,68
	COG0477	GEPR	MFS family permease	15,78
	COG0642	T	Signal transduction histidine kinase	15,02
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	14,43
	COG0583	K	DNA-binding transcriptional regulator, LysR family	12,57
	COG0841	V	Multidrug efflux pump subunit AcrB	10,82
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	8,93
	COG1024	I	Enoyl-CoA hydratase/carnithine racemase	8,77
	COG1012	C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	8,62
	COG2204	T	DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains	8,48

Table 11 Top 10 most abundant COGs found on Petrochemical ETP datasets

Dataset	*COG*	*Function*	*COG description*	*Abundance*
L1	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	16,58
	COG0642	T	Signal transduction histidine kinase	15,02
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	12,69
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	11,62
	COG0477	GEPR	MFS family permease	11,03
	COG0583	K	DNA-binding transcriptional regulator, LysR family	9,86
	COG1012	C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	8,78
	COG1024	I	Enoyl-CoA hydratase/carnithine racemase	8,70
	COG0463	M	Glycosyltransferase involved in cell wall bisynthesis	8,38
	COG2204	T	DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains	8,21
L2	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	19,19
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	16,23
	COG0642	T	Signal transduction histidine kinase	14,50
	COG0477	GEPR	MFS family permease	13,62
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	11,16
	COG1024	I	Enoyl-CoA hydratase/carnithine racemase	10,76
	COG0318	IQ	Acyl-CoA synthetase (AMP-forming)/AMP-acid ligase II	9,37
	COG0583	K	DNA-binding transcriptional regulator, LysR family	9,21
	COG0438	M	Glycosyltransferase involved in cell wall bisynthesis	9,02
	COG1012	C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	8,93
H1	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	18,75
	COG0642	T	Signal transduction histidine kinase	17,50
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	14,46
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	14,37
	COG0583	K	DNA-binding transcriptional regulator, LysR family	11,84
	COG0477	GEPR	MFS family permease	10,56
	COG1012	C	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	10,22
	COG0318	IQ	Acyl-CoA synthetase (AMP-forming)/AMP-acid ligase II	9,37
	COG0183	I	Acetyl-CoA acetyltransferase	9,28
	COG1024	I	Enoyl-CoA hydratase/carnithine racemase	9,21
H2	COG0642	T	Signal transduction histidine kinase	24,25
	COG1028	IQR	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	19,99
	COG1960	I	Acyl-CoA dehydrogenase related to the alkylation response protein AidB	16,86
	COG0477	GEPR	MFS family permease	16,16
	COG0583	K	DNA-binding transcriptional regulator, LysR family	15,15
	COG0745	TK	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	12,62
	COG2204	T	DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains	11,59
	COG0840	NT	Methyl-accepting chemotaxis protein	10,73
	COG0784	T	CheY chemotaxis protein or a CheY-like REC (receiver) domain	10,42
	COG1024	I	Enoyl-CoA hydratase/carnithine racemase	10,30

In the top 10, there were also found two COGs related with signal transduction mechanisms: COG0642 a signal transduction histidine kinase, and, COG0745 an OmpR family DNA-binding response regulator. Proteins from OmpR family are involved in different processes such as osmoregulation, oxidative and acid stress response, motility, virulence and outer membrane biogenesis (135).

As already notice on the taxonomical analysis, H1 appears to have a deviant behaviour in many of the analysed functions and therefore it analysis as to be taken carefully. This may be due to an erroneous sampling or a problem in sample preparation and sequencing. Therefore, an impact of its results should be verified with a replicate.

COG1960, also found on top 10, is described as an "acyl-CoA dehydrogenase related to the alkylation response protein AidB". Alkaline agents can be present in the cell and in the environment and can produce cytotoxic and mutagenic lesions. Therefore they are used as chemotherapy drugs binding to DNA and thus preventing proper DNA replication. They are also used in petrochemical industry for the production of important intermediates such as ethyl benzene and cumene (136). Concerning the CETP datasets, a higher abundance of COG1960 was found on dataset A2. This may be related with the dyes and the chemicals from textile production present in this king of CETP. Regarding the petrochemical complex datasets, an increase can be observed on the second timepoints (L2 and H2), which may be related with the microorganisms defence mechanisms against the alkylating agents related with petrochemical processes. This suggests a high abundance of this chemicals which should to be considered in the treatment process.

Besides the top 10, there were found discrepancies in some COG functions that can have an important impact. These COGs abundances can be found in Table 12.

Table 12 Relevant COGs found

COG	*Functions*	*Description*	AKR06	A2	L1	L2	H1	H2
COG0841	V	Multidrug efflux pump subunit AcrB	8.05	10.82	**7.13**	8.19	6.29	9.86
COG0845	MV	Multidrug efflux pump subunit AcrA (membrane-fusion protein)	3.30	6.70	5.34	**3.97**	3.29	8.42
COG1131	V	ABC-type multidrug transport system, ATPase component	5.49	6.37	7.04	8.80	8.27	7.51
COG1132	V	ABC-type multidrug transport system, ATPase and permease component	7.28	6.38	6.44	5.13	7.06	6.77
COG0842	V	ABC-type multidrug transport system, permease component	2.04	2.84	2.94	2.48	1.35	3.97
COG0577	V	ABC-type antimicrobial peptide transport system, permease component	2.66	2.87	4.89	3.15	2.96	3.29
COG2274	V	ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase domain	0.27	0.70	1.67	0.71	0.75	1.16
COG0610	V	Type I site-specific restriction-modification system, R (restriction) subunit and related helicases ...	1.55	2.58	**2.05**	1.64	1.80	1.64
COG4096	V	Type I site-specific restriction endonuclease, part of a restriction-modification system	0.78	0.98	**1.08**	0.36	0.65	0.34
COG1002	V	Type II restriction/modification system, DNA methylase subunit YeeA	1.12	1.75	**2.27**	1.85	1.91	1.60
COG0286	V	Type I restriction-modification system, DNA methylase subunit	1.38	4.21	**1.91**	1.87	3.44	1.51
COG2124	QV	Cytochrome P450	1.12	1.47	1.70	3.01	1.09	1.68
COG1518	V	CRISPR/Cas system-associated endonuclease Cas1	0.19	0.45	**1.31**	0.41	1.64	0.96
COG3649	V	CRISPR/Cas system type I-B associated protein Csh2, Cas7 group, RAMP superfamily	0.00	0.19	**0.39**	0.06	0.10	0.15
COG1353	V	CRISPR/Cas system-associated protein Cas10, large subunit of type III CRISPR-Cas systems, contains HD superfamily nuclease domain	0.00	0.10	**0.29**	0.15	0.17	0.45
COG2141	HR	Flavin-dependent oxidoreductase, luciferase family (includes alkanesulfonate monooxygenase SsuD and methylene tetrahydromethanopterin reductase)	2.88	3.39	1.47	3.47	1.91	2.55
COG0543	HC	NAD(P)H-flavin reductase	0.841201	1.31174	1.828178	0.818181	0.9375	2.465753
COG2931	Q	Ca2+-binding protein, RTX toxin-related	4.94	1.22	2.64	2.71	1.21	4.33

Regarding the functions related with defence mechanisms, some variations can be observed between the datasets. COG0841 is described as a multidrug efflux pump subunit AcrB, and is highly abundant in all datasets. The highest abundance is found on A2. *E.coli* AcrB works as a proton/drug antiporter being part of a tripartite flux system AcrA/AcrB/TolC, related with the efflux of antibiotics, dyes, bile salts and detergents (137). Since A2 data refers to a textile and dye industry area CETP, this value may be due to the capability of the microorganisms in the sample to pump this type of chemicals. Also, an increase in COG0841 is observed between L1 and L2. However, a decrease between this two datasets is observed in COG0845, the pump subunit AcrA.

COG1131, COG1132, COG0842, COG0577 and COG2274 are described as ATP-binding cassette (ABC) transporters related with drug transport. The abundance of this transporters decreased between L1 and L2, with exception of COG1131 which value increased from L1 to L2. This can suggest an less hostile environment in the second timepoint Also, H2 seems to have higher values of this type of transporters when comparing with L2, which may be related with a less efficient treatment process regarding the total dissolved solids concentration.

Type I and type II restriction modification systems are abundant in A2 dataset. Furthermore, this systems appear to be less abundant on dataset L2 comparing with L1. Restriction modification systems are related with defence against bacteriophages, which may be suggestive of the phage abundance in the samples (138) .Additionally, the same behaviour regarding the low salts petrochemical ETP, was observed with CRISPR/Cas system related COGs, which are also related with defence against bacteriophages (139).

Cytochrome P450 (COG2124) was highly found on L2 dataset and may be related with the oxidation of exogenous and endogenous chemicals (140).

COG2141, a "flavin-dependent oxidoreductase, luciferase family (includes alkanesulfonate monooxygenase SsuD and methylene tetrahydromethanopterin reductase)", is a member of the Flavin-utilizing monoxigenase superfamily and it description is associated at a *Thiomonas* bhuboneswarensis protein (UniProtKB). The increase between H1 and H2 can be associated with the taxonomy analysis where there was observed an increase in the *Thiomonas* genus. However, the percentage found on L1 is not compatible with the abundance of *Thiomonas* genus, since it was the prevalent genus found on this dataset. COG0543, a NAD(P)H-flavin reductase, behaviors differently, having a decrease from L1 to L2, keeping high values on H2. The luciferase family comprise oxidative enzymes that produce bioluminescence. Flavin-dependent proteins are important in both aerobic and anaerobic pathways and are necessary to maintain basic metabolic functions (141).

Ca2+-binding protein, RTX toxin-related, is more abundant on datasets AKR06 and H2, suggesting a more cytotoxic capability of the organisms found on those datasets.

Another analysis was performed with the *MGX* software using both the pre-processed reads and the assembled contigs (S-table 65). The top 10 COG results comparing the results from MG-RAST with assembled reads and the results from *MGX* can be found in Table 13 and 14.

Table 13 Comparision between COGs percentages for assembled data, using MG-RAST, and unassembled and assembled data using MGX on CETP datasets.

Datasets	*MG-RAST assembled data*		*MGX* unassembled data		*MGX* assembled data	
	COG	Percentage	COG	Percentage	COG	Percentage
AKR06	COG1028	0.89%	COG0642	0.95%	COG0642	1,09%
	COG0642	0.66%	COG1028	0.75%	COG0438	0,79%
	COG0745	0.59%	COG0841	0.67%	COG0477	0,61%
	COG0477	0.53%	COG1960	0.62%	COG0841	0,58%
	COG0500	0.48%	COG0477	0.61%	COG0515	0,52%
	COG0841	0.45%	COG1012	0.52%	COG1012	0,51%
	COG1012	0.45%	COG2217	0.51%	COG1132	0,48%
	COG1960	0.44%	COG1132	0.48%	COG1028	0,48%
	COG1132	0.41%	COG0318	0.45%	COG0463	0,47%
	COG1136	0.40%	COG2931	0.43%	COG2217	0,44%
A2	COG1028	0.86%	COG0642	0.90%	COG0642	1,40%
	COG0477	0.77%	COG0841	0.88%	COG0515	0,84%
	COG0642	0.73%	COG0477	0.61%	COG0841	0,74%
	COG1960	0.70%	COG2217	0.57%	COG0477	0,62%
	COG0583	0.61%	COG1629	0.55%	COG1960	0,61%
	COG0841	0.53%	COG1028	0.53%	COG1028	0,60%
	COG0745	0.44%	COG1960	0.53%	COG2801	0,51%
	COG1024	0.43%	COG1012	0.47%	COG1012	0,50%
	COG1012	0.42%	COG3696	0.44%	COG0438	0,50%
	COG2204	0.41%	COG0845	0.44%	COG3696	0,45%
AKR12	-	-	COG0642	0,95%	COG0642	1,19%
	-	-	COG0841	0,88%	COG1629	1,11%
	-	-	COG1960	0,75%	COG0841	0,80%
	-	-	COG1028	0,70%	COG1960	0,78%
	-	-	COG0477	0,67%	COG0477	0,66%
	-	-	COG0583	0,58%	COG2217	0,63%
	-	-	COG1012	0,56%	COG1012	0,60%
	-	-	COG2217	0,54%	COG0318	0,60%
	-	-	COG1629	0,51%	COG1028	0,60%
	-	-	COG0318	0,50%	COG0582	0,58%

Table 14 Comparision between COGs percentages for assembled data, using MG-RAST, and unassembled and assembled data using MGX on petrochemical ETP datasets.

Datasets	*MG-RAST assembled data*		*MGX* unassembled data		*MGX* assembled data	
L1	COG1028	0.88%	COG0642	1.52%	COG0642	1,41%
	COG0642	0.80%	COG1028	0.66%	COG0438	0,79%
	COG1960	0.67%	COG1960	0.64%	COG1960	0,59%
	COG0745	0.62%	COG0841	0.63%	COG1028	0,57%
	COG0477	0.59%	COG0477	0.61%	COG0477	0,52%
	COG0583	0.52%	COG2204	0.47%	COG0841	0,52%
	COG1012	0.47%	COG3696	0.47%	COG0463	0,51%
	COG1024	0.46%	COG0515	0.46%	COG0515	0,49%
	COG0463	0.45%	COG1012	0.44%	COG0457	0,46%
	COG2204	0.44%	COG0438	0.43%	COG2931	0,42%
L2	COG1028	0.95%	COG0642	1.13%	COG0642	1,43%
	COG1960	0.80%	COG1028	0.79%	COG0515	0,78%
	COG0642	0.72%	COG1960	0.76%	COG0438	0,78%
	COG0477	0.67%	COG0477	0.65%	COG1028	0,75%
	COG0745	0.55%	COG0515	0.63%	COG1960	0,75%
	COG1024	0.53%	COG0841	0.62%	COG0477	0,61%
	COG0318	0.46%	COG0318	0.52%	COG0841	0,48%
	COG0583	0.46%	COG1012	0.50%	COG2217	0,47%
	COG0438	0.45%	COG2217	0.44%	COG0463	0,45%
	COG1012	0.44%	COG0438	0.43%	COG0745	0,45%
H1	COG1028	0.88%	COG0642	1.13%	COG0642	1,70%
	COG0642	0.82%	COG1960	0.94%	COG0438	0,64%
	COG0745	0.68%	COG1028	0.73%	COG1960	0,60%
	COG1960	0.67%	COG1012	0.63%	COG0477	0,54%
	COG0583	0.55%	COG0841	0.58%	COG1132	0,54%
	COG0477	0.49%	COG0318	0.55%	COG1028	0,53%
	COG1012	0.48%	COG0477	0.51%	COG2217	0,51%
	COG0318	0.44%	COG3181	0.49%	COG0318	0,46%
	COG0183	0.43%	COG2217	0.48%	COG0463	0,45%
	COG1024	0.43%	COG0183	0.44%	COG1012	0,44%
H2	COG0642	1.02%	COG0642	1.51%	COG0642	1,63%
	COG1028	0.84%	COG1028	0.71%	COG0438	0,75%
	COG1960	0.71%	COG1960	0.70%	COG0477	0,70%
	COG0477	0.68%	COG0515	0.69%	COG1028	0,69%
	COG0583	0.64%	COG0841	0.66%	COG1960	0,67%
	COG0745	0.53%	COG0477	0.61%	COG0515	0,62%
	COG2204	0.49%	COG2204	0.53%	COG0841	0,51%
	COG0840	0.45%	COG3696	0.49%	COG0318	0,51%
	COG0784	0.44%	COG0318	0.48%	COG2204	0,47%
	COG1024	0.44%	COG1012	0.46%	COG0463	0,46%

The main COGs appear to be the same between the different assembled and unassembled data and between the two softwares. However, some discrepancies were found comparing the strategies. The discrepancies between the different approaches are higher in some COGs than in others. However, even minor differences may lead to change the top position, the prevalence of that function comparing with other and further implicate a different biological interpretation. Since the *MGX* is intended to use with unassembled data, there were also created pseudo-reads using the assembled data (data not shown). A stronger similarity is observed between the *MGX* assembled data and pseudo-reads which may suggest that the discrepancies may be induced not only by the software performance regarding the sequences length but also by the assembly step.

To highlight the assembly impact on the functional analysis, a hierarchical clustering analysis was performed using both assembled and unassembled data analysed with *MGX* (Figure 21). There can be observed three main clusters, the first cluster contains datasets A2 and AKR12 where assembled data and reads are grouped together. The second clusters comprises only reads (datasets AKR16, L1 L2, H1 and H2) and the third cluster contains the remaining assembled data.

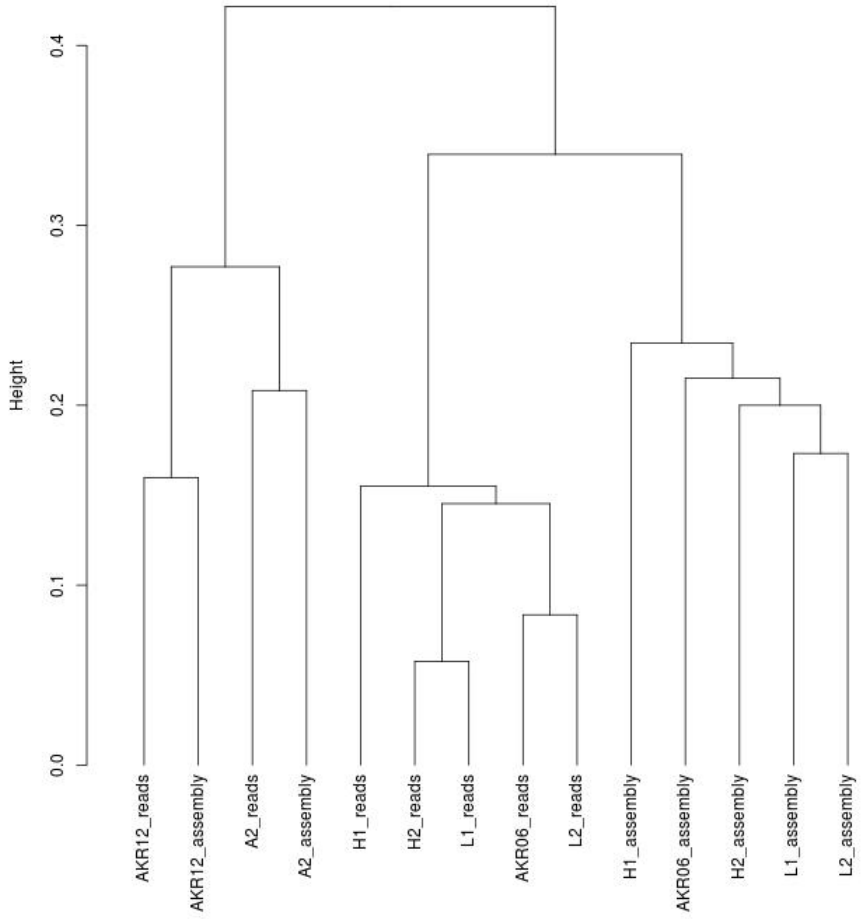


Figure 21 Hierarchical clustering analysis for COG results using MGX for both assembled and unassembled data

It is also important to notice that the clustering inside the second and third main clusters is different, which highlights the discrepancies between analysing reads and assembled contigs. Also, assembly appears to have different impacts in different datasets, since A2 and AKR12 behave similar with assembled and unassembled data. Considering this, a further study must be performed on this subject
The discrepancies found in the datasets can be helpfully to fully understand the content of this ETPs and CETPs and to access the treatment improvements between timepoints. A higher abundance of defence mechanisms on L1 comparing with L2 suggests a transition to a less hostile environment. Nevertheless, the putative presence of alkaline agents in A2 and on the second timepoints of the petrochemical complex ETPs has to be further analysed to effectively treat this effluents. Also, the higher abundance of COGs related with drug transport on H2 when comparing with L2 may has also to be considered for further analysis.

5. CONCLUSION

Whole shotgun metagenomic analysis is a very challenging research field. There was designed a strategy to process and analyse this type of data and there were performed analysis of seven target datasets comprising wastewater treatment processes.

Two representative simulated datasets were generated from the target data to analyse. This strategy may be important to design and perform a further analysis. However, some improvements must be implemented regarding the sequencing quality thresholds, coverage depth and inclusion of unknown organisms.

Differences in coverage depth were assessed by splitting the data in three coverage levels. Comparing the results from data with and without coverage split highlighted the importance of the different abundances in metagenomic data and the need to account for this differences so that low coverage data will not be excluded. Further studies must be performed to study different ways of coverage assessment in order to split different organismal abundances instead of grouping conservative and repetitive genomic areas in one dataset and specific genomic sequences in the opposite dataset. Also, normalization techniques may be tested before and after assembly.

Different assembly tools and k-mers were tested showing the importance of tool and k-mer selection regarding the datasets. Unfortunately, the impact of assembly was not extensively assessed, since the taxonomic analysis with the assembled data did not retrieved valuable results. This may be explored using other taxonomic tools to see the impact on taxonomy analysis. The assembly impact on functional analysis was assessed by using unassembled and assembled data on *MGX.* The results show discrepancies between the two datatypes, which were highlighted in a hierarchical clustering analysis. It is possible to distinguish three main clusters, a first one containing two datasets where reads and assembled contigs are clustered together and two other main clusters, one containing only reads and the other containing only assembled data.

CETP datasets showed to have a higher taxonomical diversity compared with petrochemical complex ETP datasets. Important finding genus suggest a higher presence of ammonia, methane and chromate in AKR06, comparing with the remaining datasets, which may be accounted on the effluent treatment. Both A2 and AKR12 datasets showed the presence of possible pathogenic or opportunistic microorganisms, *Arcobacter* and *Rhodococcus* genus on A2 and *Bordetella* genus on AKR12. Also, AKR12 taxonomical composition suggests a higher crude oil concentration on this dataset. Moreover, petrochemical complex ETP datasets showed to have a predominant genus related with a more hostile environment, *Thiomonas*,

which suggest a more challenging treatment process in this ETPs rather than in the CETPs. Regarding the petrochemical complex datasets, H1 showed a divergent behaviour both in taxonomy and functional analysis which alerts for some erroneous step in sample collection, preparation or sequencing. This leads to questioning the need of at least a replicate for each dataset in order to assure the findings. Comparing the low TDS activated sludge datasets, an increase in biodiversity between L1 and L2, together with the decrease of *Thiomonas* genus abundance, suggests a less hostile environment at the second timepoint. However, a putative opportunistic organism was identified on L2, *Acinectobacter* genus, which should be considered for the effluent treatment. Also *Afipia* genus which comprises opportunistic species was also found on L1, L2 and H2.

Analysing the COG functions, the hostile environment in L1 suggested by the identified microorganisms, is also supported by the abundance of defence mechanisms found on this dataset.

The higher abundance of ATP-binding cassette (ABC) transporters related with drug transport on L1 comparing with L2 suggest a decrease in this compounds prevalence with the effluent treatment. However, a higher value of this putative functions on H2 may suggest the need for a more challenging treatment for the high TDS petrochemical effluents.

AKR12 *MG-RAST* functional results weren't available for analysis; however, some insights from a CETP at Ankleshwar city were achieve with A2 dataset that suggests the presence of alkaline agents.

In summary, there were tested different approaches and raised questions regarding the metagenomic analysis. This data encourages a further analysis of questions such as normalization, coverage, datasets discrepancies, assembly parametrization and taxonomic and functional analysis. The biological findings may be important to direct the effluent treatments regarding the principal organisms and compounds identified in the samples.

6. FUTURE PERSPECTIVES

As already mentioned, metagenomics comprises a really challenging analysis and different aspects have to be considered, such as the dataset characteristics, pre-processing, importance of assemble data to further analysis, sequence coverage depth and data normalization.

The use of simulated datasets appears to be a good strategy to select the bioinformatics tools to use on the target data; however, some adjustments to the present approach can be made. To access the importance and accuracy of coverage split, simulated datasets with and without coverage split may be generated. Also, strategies to include and mimic unknown microorganism must be studied. The sequencing quality simulation may also be addressed in order to be more representative of the real data. Different coverage assessment and split should be tested in order to address the problem of grouping conserved and repetitive datasets instead of high abundant organisms. Also, the normalization issue may be also addressed both prior and after processing in order to assure a reliable comparative analysis.

This improvements of the simulated datasets creation will be further useful in understanding not only the impact of different processing and analysing tools, as also to visualize the impact of normalization and coverage split.

A more extensive study on taxonomical analysis tools should be performed, including more tools and assessing the impact of assembly in the taxonomical analysis.

In terms of functionality, different tools should be tested, including gene prediction tools and the alignment with the different available databases. A further analysis on the impact of the use of unassembled or assembled data should be performed. It is important to create a strategy to make use of simulated datasets to assure the functional information given. Finally, a parallelism between the different functional annotation databases information must be assessed.

7. REFERENCES

1. Water Scarcity | Threats | WWF [Internet]. [cited 2016 Feb 7]. Available from: http://www.worldwildlife.org/threats/water-scarcity
2. WHO | Drinking-water. World Health Organization; [cited 2016 Feb 7]; Available from: http://www.who.int/mediacentre/factsheets/fs391/en/
3. Wiener MJ, Jafvert CT, Nies LF. The assessment of water use and reuse through reported data: A US case study. Sci Total Environ [Internet]. Elsevier B.V.; 2016;539:70–7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0048969715306161
4. Pathe PP, Suresh Kumar M, Kharwade MR, Kaul SN. Common Effluent Treatment Plant (CEPT) for Wastewater Management from a Cluster of Small Scale Tanneries. Environ Technol [Internet]. 2004 May [cited 2016 Dec 19];25(5):555–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15242231
5. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135–45.
6. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci [Internet]. 2014 Jan [cited 2014 Jul 9];5:209. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4059276&tool=pmcentrez&rendertype=abstract
7. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol [Internet]. 2008;26(10):1135–45. Available from: http://www.nature.com/doifinder/10.1038/nbt1486
8. An Introduction to Next-Generation Sequencing Technology [Internet]. [cited 2016 Feb 7]. Available from: http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf
9. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012.
10. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet [Internet]. 2016 May 17 [cited 2017 Jan 29];17(6):333–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27184599
11. Oxford. Specifications - Community - Oxford Nanopore Technologies [Internet]. [cited 2016 Feb 12]. Available from: https://nanoporetech.com/community/specifications
12. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One [Internet]. 2013 Jan [cited 2014 Jul

10];8(12):e85024. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871669&tool=pmcentrez&rendertype=abstract

13. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A. Manipulation of FASTQ data with Galaxy. Bioinformatics [Internet]. 2010 Jul 15 [cited 2015 Dec 16];26(14):1783–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2894519&tool=pmcentrez&rendertype=abstract
14. Patel RK, Jain M. NGS QC toolkit: A toolkit for quality control of next generation sequencing data. PLoS One. 2012;7(2).
15. Zhou Q, Su X, Wang A, Xu J, Ning K. QC-Chain: fast and holistic quality control method for next-generation sequencing data. PLoS One [Internet]. 2013 Jan [cited 2015 Dec 1];8(4):e60234. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3615005&tool=pmcentrez&rendertype=abstract
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 15 [cited 2014 Jul 9];25(16):2078–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract
17. Yang X, Chockalingam SP, Aluru S. A survey of error-correction methods for next-generation sequencing. Brief Bioinform [Internet]. 2013;14(1):56–66. Available from: http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbs015
18. Salmela L, Schroder J. Correcting errors in short reads by multiple alignments. Bioinformatics [Internet]. 2011;27(11):1455–61. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr170
19. Kao W, Chan AH, Song YS. ECHO : A reference-free short-read error correction algorithm. 2011;1181–92.
20. Salmela L. Correction of sequencing errors in a mixed set of reads. Bioinformatics. 2010;26(10):1284–90.
21. Ilie L, Fazayeli F, Ilie S. HiTEC: Accurate error correction in high-throughput sequencing data. Bioinformatics. 2011;27(3):295–302.

22. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res [Internet]. 1998 Mar 1 [cited 2015 Aug 26];8(3):186–94. Available from: http://genome.cshlp.org/content/8/3/186.long
23. Smeds L, Künstner A. ConDeTri–a content dependent read trimmer for Illumina data. PLoS One [Internet]. 2011 Jan [cited 2016 Feb 7];6(10):e26314. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3198461&tool=pmcentrez&rendertype=abstract
24. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics [Internet]. 2011 Mar 15 [cited 2014 Jul 11];27(6):863–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3051327&tool=pmcentrez&rendertype=abstract
25. Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics [Internet]. 2010 Jan [cited 2015 May 19];11:485. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2956736&tool=pmcentrez&rendertype=abstract
26. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med [Internet]. 2014;9(1):8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4064128&tool=pmcentrez&rendertype=abstract
27. Kong Y. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics [Internet]. 2011 Aug [cited 2016 Feb 7];98(2):152–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21651976
28. Modolo L, Lerat E. UrQt: an efficient software for the Unsupervised Quality trimming of NGS data. BMC Bioinformatics [Internet]. BioMed Central; 2015 Jan 29 [cited 2016 Jan 28];16(1):137. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0546-8
29. Yun S, Yun S. Masking as an effective quality control method for next-generation sequencing data analysis. BMC Bioinformatics [Internet]. 2014 Jan [cited 2016 Feb 7];15:382. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4268903&tool=pmcentrez&rendertype=abstract
30. Criscuolo A, Brisse S. AlienTrimmer: a tool to quickly and accurately trim off multiple short

contaminant sequences from high-throughput sequencing reads. Genomics [Internet]. Jan [cited 2016 Jan 18];102(5-6):500–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23912058

31. Li Y-L, Weng J-C, Hsiao C-C, Chou M-T, Tseng C-W, Hung J-H. PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. BMC Bioinformatics [Internet]. 2015 Jan [cited 2016 Feb 7];16 Suppl 1:S2. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4331701&tool=pmcentrez&rendertype=abstract
32. Ma Y, Xie H, Han X, Irwin DM, Zhang Y-P. QcReads: an adapter and quality trimming tool for next-generation sequencing reads. J Genet Genomics [Internet]. 2013 Dec 20 [cited 2016 Feb 7];40(12):639–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24377870
33. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med. 2014;9(1):8.
34. Aronesty E. InsideDNA: fastq-mcf - Scan sequence file for adapters and perform clipping [Internet]. [cited 2016 May 10]. Available from: https://insidedna.me/tools/page/fastq_mcf
35. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights [Internet]. 2015 Jan [cited 2015 Nov 23];9:75–88. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4426941&tool=pmcentrez&rendertype=abstract
36. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol [Internet]. 2013 Aug 18 [cited 2014 Jul 14];16(9):2659–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24102695
37. Bragg L, Tyson GW. Metagenomics using next-generation sequencing. Methods Mol Biol [Internet]. 2014 Jan [cited 2015 Nov 10];1096:183–201. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24515370
38. Haque MM, Bose T, Dutta A, Reddy CVSK, Mande SS. CS-SCORE: Rapid identification and removal of human genome contaminants from metagenomic datasets. Genomics [Internet]. Elsevier Inc.; 2015;106(2):116–21. Available from: http://dx.doi.org/10.1016/j.ygeno.2015.04.005
39. Zhou Q, Su X, Jing G, Ning K. Meta-QC-Chain: comprehensive and fast quality control method for

metagenomic data. Genomics Proteomics Bioinformatics [Internet]. 2014 Feb [cited 2015 Dec 17];12(1):52–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4411374&tool=pmcentrez&rendertype=abstract

40. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: A de Novo assembler for metagenomic data. Bioinformatics. 2011;27(13):94–101.

41. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics [Internet]. 2012 Jun 1 [cited 2016 Jan 7];28(11):1420–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22495754

42. Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, et al. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res [Internet]. 2015 Jan 13 [cited 2015 Jan 14];43(7):e46. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4402509&tool=pmcentrez&rendertype=abstract

43. Wu YW, Rho M, Doak TG, Ye Y. Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics. Bioinformatics. 2012;28(18):363–9.

44. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol [Internet]. BioMed Central Ltd; 2012;13(12):R122. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4056372&tool=pmcentrez&rendertype=abstract

45. Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) [Internet]. 2013;3(5):865–80. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3656733&tool=pmcentrez&rendertype=abstract

46. Haider B, Ahn TH, Bushnell B, Chai J, Copeland A, Pan C. Omega: An Overlap-graph de novo Assembler for Metagenomics. Bioinformatics. 2014;30(19):2717–22.

47. Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res [Internet]. 2014;22(1):69–77. Available from: http://dnaresearch.oxfordjournals.org/content/22/1/69.full

48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol [Internet]. 2012 May [cited 2016 Aug 12];19(5):455–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22506599

49. Lai B, Wang F, Wang X, Duan L, Zhu H. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics [Internet]. 2015 Jan [cited 2016 Jan 27];16:244. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4545859&tool=pmcentrez&rendertype=abstract

50. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods [Internet]. 2009 Sep [cited 2016 Jan 21];6(9):673–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2762791&tool=pmcentrez&rendertype=abstract

51. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics [Internet]. BioMed Central Ltd; 2011;12(Suppl 2):S4. Available from: http://www.biomedcentral.com/1471-2164/12/S2/S4

52. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics [Internet]. 2012 Apr 1 [cited 2016 Jan 14];28(7):1033–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22332237

53. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods [Internet]. 2012 Aug [cited 2015 Mar 6];9(8):811–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3443552&tool=pmcentrez&rendertype=abstract

54. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods [Internet]. 2015 Oct 29 [cited 2016 Dec 19];12(10):902–3. Available from: http://www.nature.com/doifinder/10.1038/nmeth.3589

55. Sharpton TJ, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, et al. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput Biol [Internet]. 2011 Jan [cited 2016 Feb 7];7(1):e1001061.

Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3024254&tool=pmcentrez&rendertype=abstract

56. Darling AE, Jospin G, Lowe E, Matsen FA, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ [Internet]. 2014;2:e243. Available from: https://peerj.com/articles/243

57. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, et al. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Mol Ecol Resour [Internet]. 2015 Nov [cited 2017 Jan 11];15(6):1403–14. Available from: http://doi.wiley.com/10.1111/1755-0998.12399

58. Eddy SR. What is a hidden Markov model? Nat Biotechnol [Internet]. 2004 Oct [cited 2015 Jan 2];22(10):1315–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15470472

59. Su X, Pan W, Song B, Xu J, Ning K. Parallel-META 2.0: Enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization. Zhang Z, editor. PLoS One [Internet]. 2014 Mar 3 [cited 2017 Jan 11];9(3):e89323. Available from: http://dx.plos.org/10.1371/journal.pone.0089323

60. Le V Van, Tran L Van, Tran H Van. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. BMC Bioinformatics [Internet]. 2016;17:22. Available from: http://dx.doi.org/10.1186/s12859-015-0872-x\nhttp://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0872-x\nhttp://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/s12859-015-0872-x?site=bmcbioinformatics.biomedcentral.com

61. Noguchi H, Park J, Takagi T. MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res. 2006;34(19):5623–30.

62. Noguchi H, Taniguchi T, Itoh T. Meta gene annotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res. 2008;15(6):387–96.

63. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38(12):1–15.

64. Rho M, Tang H, Ye Y. FragGeneScan: Predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38(20):1–12.

65. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic

sequences augmented by classification and clustering. Nucleic Acids Res. 2012;40(1):1–12.

66. Hoff KJ, Lingner T, Meinicke P, Tech M. Orphelia: Predicting genes in metagenomic sequencing reads. Nucleic Acids Res. 2009;37(SUPPL. 2):101–5.

67. Liu Y, Guo J, Zhu H. Gene prediction in metagenomic fragments based on the SVM algorithm. Proc - 2011 4th Int Conf Biomed Eng Informatics, BMEI 2011 [Internet]. BioMed Central Ltd; 2011;3(Suppl 5):1738–42. Available from: http://www.biomedcentral.com/1471-2105/14/S5/S12

68. El Allali A, Rose JR. MGC: a metagenomic gene caller. BMC Bioinformatics [Internet]. 2013 Jan [cited 2016 Feb 5];14 Suppl 9:S6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3698006&tool=pmcentrez&rendertype=abstract

69. Ward N, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One [Internet]. 2014 Jan [cited 2015 Dec 23];9(7):e101850. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4094424&tool=pmcentrez&rendertype=abstract

70. Suzuki S, Ishida T, Kurokawa K, Akiyama Y. GHOSTM: a GPU-accelerated homology search tool for metagenomics. PLoS One [Internet]. 2012 Jan [cited 2016 Feb 6];7(5):e36060. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3344842&tool=pmcentrez&rendertype=abstract

71. Yano M, Mori H, Akiyama Y, Yamada T, Kurokawa K. CLAST: CUDA implemented large-scale alignment search tool. BMC Bioinformatics [Internet]. 2014 Jan [cited 2016 Feb 6];15:406. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4271471&tool=pmcentrez&rendertype=abstract

72. Niu B, Zhu Z, Fu L, Wu S, Li W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. Bioinformatics [Internet]. 2011 Jun 15 [cited 2016 Feb 6];27(12):1704–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106194&tool=pmcentrez&rendertype=abstract

73. Zhao Y, Tang H, Ye Y. RAPSearch2 : a fast and memory-efficient protein similarity search tool for

next-generation sequencing data. 2012;28(1):125–6.

74. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res [Internet]. 2004 Jan 1 [cited 2017 Jan 7];32(90001):138D – 141. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14681378

75. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res [Internet]. 2012 Jan 1 [cited 2017 Jan 7];40(D1):D290–301. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22127870

76. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, et al. The UniProt-GO Annotation database in 2011. Nucleic Acids Res [Internet]. 2012 Jan 1 [cited 2017 Jan 7];40(D1):D565–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22123736

77. Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2017 Jan 7];28(1):33–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10592175

78. Galperin MY, Makarova KS, Wolf YI, Koonin E V. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res [Internet]. 2015 Jan 28 [cited 2016 Dec 10];43(D1):D261–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25428365

79. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, et al. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res [Internet]. 2014 Jan [cited 2017 Jan 7];42(Database issue):D231–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1253

80. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res [Internet]. 2017 Jan 4 [cited 2017 Jan 8];45(D1):D353–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27899662

81. Overbeek R, Begley T, Butler RM, Choudhuri J V, Chuang H-Y, Cohoon M, et al. The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res [Internet]. 2005 Sep 25 [cited 2017 Jan 8];33(17):5691–702. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16214803

82. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res [Internet]. Oxford University Press; 2014 Jan [cited 2017 Jan 8];42(Database issue):D206–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24293654

83. Silva GGZ, Green KT, Dutilh BE, Edwards RA. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics [Internet]. 2016 Feb 1 [cited 2016 Dec 10];32(3):354–61. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv584

84. Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics [Internet]. 2009 Jan [cited 2016 Jan 21];10:359. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2774329&tool=pmcentrez&rendertype=abstract

85. Bose T, Haque MM, Reddy C, Mande SS. COGNIZER: A Framework for Functional Annotation of Metagenomic Datasets. PLoS One [Internet]. 2015 Jan [cited 2016 Feb 6];10(11):e0142102. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4641738&tool=pmcentrez&rendertype=abstract

86. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res [Internet]. 2007 Mar [cited 2014 Jul 9];17(3):377–86. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1800929&tool=pmcentrez&rendertype=abstract

87. Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, Meyer F, et al. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics [Internet]. 2011 Feb 15 [cited 2017 Jan 8];12(Suppl 1):S21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21342551

88. Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol. 2016 Jan;1399:207–33.

89. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics [Internet]. 2011 Dec 7 [cited 2017 Jan 8];12(1):444. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21899761

90. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335–6.

91. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009 Dec;75(23):7537–41.

92. Rodriguez-R LM, Konstantinidis KT. Estimating coverage in metagenomic data sets and why it matters. ISME J [Internet]. 2014 Nov 13 [cited 2017 Jan 11];8(11):2349–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24824669

93. Joiny Genome Institute. BBNorm Guide - DOE Joint Genome Institute [Internet]. [cited 2016 Jun 11]. Available from: http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbnorm-guide/

94. Korbel JO, Abyzov A, Mu X, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol [Internet]. BioMed Central; 2009 [cited 2016 Aug 22];10(2):R23. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-2-r23

95. Bitbucket A. MetaPhlAn2 tutorial [Internet]. 2016 [cited 2017 Jan 29]. Available from: https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2#rst-header-create-a-heatmap-with-hclust2

96. Zielińska M, Rusanowska P, Jarząbek J, Nielsen JL. Community dynamics of denitrifying bacteria in full-scale wastewater treatment plants. Environ Technol [Internet]. 2016 Sep 16 [cited 2016 Dec 27];37(18):2358–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26932371

97. Liu B, Frostegård A, Shapleigh JP. Draft genome sequences of five strains in the genus thauera. Genome Announc [Internet]. 2013 Jan 28 [cited 2016 Dec 27];1(1):e00052–12 – e00052–12. Available from: http://genomea.asm.org/cgi/doi/10.1128/genomeA.00052-12

98. Baytshtok V, Kim S, Yu R, Park H, Chandran K. Molecular and biokinetic characterization of methylotrophic denitrification using nitrate and nitrite as terminal electron acceptors. Water Sci Technol [Internet]. 2008 Aug [cited 2016 Dec 28];58(2):359. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18701786

99. de Bentzmann S, Plésiat P. The Pseudomonas aeruginosa opportunistic pathogen and human infections. Environ Microbiol [Internet]. 2011 Jul [cited 2017 Mar 28];13(7):1655–65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21450006

100. Morohoshi T, Okutsu N, Xie X, Ikeda T. Identification of Quorum-Sensing Signal Molecules and a Biosynthetic Gene in Alicycliphilus sp. Isolated from Activated Sludge. Sensors [Internet]. 2016 Aug 2 [cited 2016 Dec 27];16(8):1218. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27490553

101. Weelink SAB, Tan NCG, ten Broeke H, van den Kieboom C, van Doesburg W, Langenhoff AAM, et

al. Isolation and Characterization of Alicycliphilus denitrificans Strain BC, Which Grows on Benzene with Chlorate as the Electron Acceptor. Appl Environ Microbiol [Internet]. 2008 Nov 1 [cited 2016 Dec 27];74(21):6672–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18791031

102. Mechichi T, Stackebrandt E, Fuchs G. Alicycliphilus denitrificans gen. nov., sp. nov., a cyclohexanol-degrading, nitrate-reducing beta-proteobacterium. Int J Syst Evol Microbiol [Internet]. 2003 Jan 1 [cited 2016 Dec 27];53(1):147–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12661531

103. Dullius CH, Chen C-Y, Schink B. Nitrate-Dependent Degradation of Acetone by Alicycliphilus and Paracoccus Strains and Comparison of Acetone Carboxylase Enzymes. Appl Environ Microbiol [Internet]. 2011 Oct 1 [cited 2016 Dec 27];77(19):6821–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21841031

104. Niu B, Zhu Z, Fu L, Wu S, Li W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. Bioinformatics. 2011 Jun;27(12):1704–5.

105. Oosterkamp MJ, Veuskens T, Talarico Saia F, Weelink SAB, Goodwin LA, Daligault HE, et al. Genome Analysis and Physiological Comparison of Alicycliphilus denitrificans Strains BC and K601T. Driessen A, editor. PLoS One [Internet]. 2013 Jun 25 [cited 2016 Dec 27];8(6):e66971. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23825601

106. Morohoshi T, Okutsu N, Xie X, Ikeda T. Identification of Quorum-Sensing Signal Molecules and a Biosynthetic Gene in Alicycliphilus sp. Isolated from Activated Sludge. Sensors. 2016 Aug;16(8):1218.

107. Weelink SAB, Tan NCG, ten Broeke H, van den Kieboom C, van Doesburg W, Langenhoff AAM, et al. Isolation and Characterization of Alicycliphilus denitrificans Strain BC, Which Grows on Benzene with Chlorate as the Electron Acceptor. Appl Environ Microbiol. 2008 Nov;74(21):6672–81.

108. Daims H, Lebedeva E V., Pjevac P, Han P, Herbold C, Albertsen M, et al. Complete nitrification by Nitrospira bacteria. Nature [Internet]. 2015 Nov 26 [cited 2016 Dec 28];528(7583):504–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26610024

109. Dam B, Dam S, Kube M, Reinhardt R, Liesack W. Complete Genome Sequence of Methylocystis sp. Strain SC2, an Aerobic Methanotroph with High-Affinity Methane Oxidation Potential. J Bacteriol [Internet]. 2012 Nov 1 [cited 2016 Dec 28];194(21):6008–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23045511

110. Holland-Moritz HE, Bevans DR, Lang JM, Darling AE, Eisen JA, Coil DA. Draft Genome Sequence of Leucobacter sp. Strain UCD-THU (Phylum Actinobacteria). Genome Announc [Internet]. 2013

Jun 27 [cited 2016 Dec 29];1(3):e00325–13 – e00325–13. Available from: http://genomea.asm.org/cgi/doi/10.1128/genomeA.00325-13

111. Yun J-H, Cho Y-J, Chun J, Hyun D-W, Bae J-W. Genome sequence of the chromate-resistant bacterium Leucobacter salsicius type strain M1-8T. Stand Genomic Sci [Internet]. 2013 Dec 31 [cited 2016 Dec 29];9(3):495–504. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25197435

112. Bagade A V, Bachate SP, Dholakia BB, Giri AP, Kodam KM. Characterization of Roseomonas and Nocardioides spp. for arsenic transformation. J Hazard Mater [Internet]. 2016 Nov 15 [cited 2016 Dec 31];318:742–50. Available from: http://linkinghub.elsevier.com/retrieve/pii/S030438941630694X

113. Woo S-G, Srinivasan S, Yang J, Jung Y-A, Kim MK, Lee M. Nocardioides daejeonensis sp. nov., a denitrifying bacterium isolated from sludge in a sewage-disposal plant. Int J Syst Evol Microbiol [Internet]. 2012 May 1 [cited 2016 Dec 31];62(Pt 5):1199–203. Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.033308-0

114. Gölz G, Alter T, Bereswill S, Heimesaat MM. The Immunopathogenic Potential of Arcobacter butzleri – Lessons from a Meta-Analysis of Murine Infection Studies. Grivennikov S, editor. PLoS One [Internet]. 2016 Jul 20 [cited 2016 Dec 30];11(7):e0159685. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27438014

115. Bell, Philp, Aw, Christofi. The genus Rhodococcus. J Appl Microbiol [Internet]. Blackwell Science Ltd; 1998 Aug [cited 2016 Dec 31];85(2):195–210. Available from: http://doi.wiley.com/10.1046/j.1365-2672.1998.00525.x

116. Heylen K, Lebbe L, De Vos P. Acidovorax caeni sp. nov., a denitrifying species with genetically diverse isolates from activated sludge. Int J Syst Evol Microbiol [Internet]. 2008 Jan 1 [cited 2016 Dec 31];58(1):73–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18175686

117. Weiss A. The Genus Bordetella. In: The Prokaryotes [Internet]. New York, NY: Springer New York; 2006 [cited 2017 Jan 2]. p. 648–74. Available from: http://link.springer.com/10.1007/0-387-30745-1_27

118. Obi LU, Atagana HI, Adeleke RA. Isolation and characterisation of crude oil sludge degrading bacteria. Springerplus [Internet]. 2016 Dec 9 [cited 2017 Jan 2];5(1):1946. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27933233

119. Wang B, Chu K-H. Cometabolic biodegradation of 1,2,3-trichloropropane by propane-oxidizing bacteria. Chemosphere. 2017.

120. Gomez-Alvarez V, Pfaller S, Revetta RP. Draft Genome Sequence of Two Sphingopyxis sp. Strains, Dominant Members of the Bacterial Community Associated with a Drinking Water Distribution System Simulator. Genome Announc [Internet]. American Society for Microbiology (ASM); 2016 Mar 31 [cited 2017 Jan 2];4(2). Available from: http://www.ncbi.nlm.nih.gov/pubmed/27034493

121. Paul D, Bridges S, Burgess SC, Dandass Y, Lawrence ML. Genome sequence of the chemolithoautotrophic bacterium Oligotropha carboxidovorans OM5T. J Bacteriol [Internet]. American Society for Microbiology; 2008 Aug [cited 2017 Jan 3];190(15):5531–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18539730

122. Poindexter JS, Poindexter, S. J. *Caulobacter*. In: Bergey's Manual of Systematics of Archaea and Bacteria [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2015 [cited 2017 Jan 3]. p. 1–25. Available from: http://doi.wiley.com/10.1002/9781118960608.gbm00792

123. Penner F, Brossa S, Barbui AM, Ducati A, Cavallo R, Zenga F. Caulobacter spp: A Rare Pathogen Responsible for Paucisintomatic Persisitant Meningitis in a Glioblastoma Patient. World Neurosurg [Internet]. 2016 Dec [cited 2017 Jan 3];96:611.e11–611.e13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27650802

124. Arsène-Ploetze F, Koechler S, Marchal M, Coppée J-Y, Chandler M, Bonnefoy V, et al. Structure, Function, and Evolution of the Thiomonas spp. Genome. Moran NA, editor. PLoS Genet [Internet]. 2010 Feb 26 [cited 2017 Jan 2];6(2):e1000859. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20195515

125. Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, Baker S, et al. Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol [Internet]. 2004 Mar [cited 2017 Jan 3];186(5):1503–17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14973019

126. Visca P, Seifert H, Towner KJ. Acinetobacter infection–an emerging threat to human health. IUBMB Life [Internet]. 2011 Dec [cited 2017 Jan 3];63(12):1048–54. Available from: http://doi.wiley.com/10.1002/iub.534

127. Villaverde J, Rubio-Bellido M, Merchán F, Morillo E. Bioremediation of diuron contaminated soils by a novel degrading microbial consortium. J Environ Manage [Internet]. 2017 Mar 1 [cited 2017 Jan 3];188:379–86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28011373

128. Posman KM, DeRito CM, Madsen EL. Benzene Degradation by *Variovorax* sp within a Coal-tar Contaminated Groundwater Microbial Community. Appl Environ Microbiol [Internet]. 2016 Dec 2

[cited 2017 Jan 3];AEM.02658–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27913419

129. Im W-T, Liu Q-M, Lee K-J, Kim S-Y, Lee S-T, Yi T-H. Variovorax ginsengisoli sp. nov., a denitrifying bacterium isolated from soil of a ginseng field. Int J Syst Evol Microbiol [Internet]. 2010 Jul 1 [cited 2017 Jan 3];60(7):1565–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19684323

130. Balkwill DL, Fredrickson JK, Romine MF. Sphingomonas and Related Genera. In: The Prokaryotes [Internet]. New York, NY: Springer New York; 2006 [cited 2017 Jan 4]. p. 605–29. Available from: http://link.springer.com/10.1007/0-387-30747-8_23

131. Zhang P, Hozalski RM, Leach LH, Camper AK, Goslan EH, Parsons SA, et al. Isolation and characterization of haloacetic acid-degrading Afipia spp. from drinking water. FEMS Microbiol Lett [Internet]. 2009 Aug [cited 2017 Jan 4];297(2):203–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19634207

132. Marrie T, Raoult D, La Scola B, Birtles RJ, de Carolis E, Canadian Community-Acquired Pneumonia Study Group. Legionella-Like and Other Amoebal Pathogens as Agents of Community-Acquired Pneumonia. Emerg Infect Dis [Internet]. 2001 Dec [cited 2017 Jan 4];7(6):1026–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11747734

133. Lamoth F, Greub G. Amoebal pathogens as emerging causal agents of pneumonia. FEMS Microbiol Rev [Internet]. 2010 May [cited 2017 Jan 4];34(3):260–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20113355

134. Chen S, Cheng H, Liu J, Hazen TC, Huang V, He Q. Unexpected competitiveness of Methanosaeta populations at elevated acetate concentrations in methanogenic treatment of animal wastewater. Appl Microbiol Biotechnol [Internet]. 2016 Nov 17 [cited 2017 Jan 4]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/27858134

135. Itou H, Tanaka I. The OmpR-family of proteins: insight into the tertiary structure and functions of two-component regulator proteins. J Biochem [Internet]. 2001 Mar [cited 2017 Jan 23];129(3):343–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11226872

136. Carlo P. Aromatic intermediates in petrochemical industry. In: Beccari M, Romano U, editors. Encyclopedia of Hydrocarbons. Istituto della enciclopedia Italiana Giovanni Treccani; 2007. p. 605–14.

137. Seeger MA, Diederichs K, Eicher T, Brandstätter L, Schiefner A, Verrey F, et al. The AcrB efflux pump: conformational cycling and peristalsis lead to multidrug resistance. Curr Drug Targets

[Internet]. 2008 Sep [cited 2017 Jan 24];9(9):729–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18781920

138. Wilson GG. Organization of restriction-modification systems. Nucleic Acids Res [Internet]. Oxford University Press; 1991 May 25 [cited 2017 Jan 25];19(10):2539–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2041731

139. Sorek R, Lawrence CM, Wiedenheft B. CRISPR-Mediated Adaptive Immune Systems in Bacteria and Archaea. Annu Rev Biochem [Internet]. 2013 Jun 2 [cited 2017 Jan 30];82(1):237–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23495939

140. Lewis DFV, Wiseman A. A selective review of bacterial forms of cytochrome P450 enzymes. Enzyme Microb Technol. 2005;36(4):377–84.

141. Macheroux P, Kappes B, Ealick SE. Flavogenomics - a genomic and structural view of flavin-dependent proteins. FEBS J [Internet]. 2011 Aug [cited 2017 Jan 12];278(15):2625–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21635694

8. SUPPLEMENTARY DATA DESCRIPTION

S-commands1 Commands used to create a reference genome file for each cluster and coverage level

S-table 1_LOW1_grinder-ranks

S-table 2_MID1_grinder-ranks

S-table 3_HIGH1_grinder-ranks

S-table 4_LOW2_grinder-ranks

S-table 5_MID2_grinder-ranks

S-table 6_HIGH2_grinder-ranks

S-table 7_LOW1_phylum_ranks

S-table 8_LOW1_genus_ranks

S-table 9_MID1_phylum_ranks

S-table 10_MID1_genus_ranks

S-table 11_HIGH1_phylum_ranks

S-table 12_HIGH1_genus_ranks

S-table 13_LOW2_phylum_ranks

S-table 14_LOW2_genus_ranks

S-table 15_MID2_phylum_ranks

S-table 16_MID2_genus_ranks

S-table 17_HIGH2_phylum_ranks

S-table 18_HIGH2_genus_ranks

S-table 19 Statistical analysis of the simulated datasets assembled contigs resulted from different error correction strategies

S-table 20 Statistical analysis of the simulated datasets assembled contigs resulted from different trimming tools

S-table 21 Statistical analysis of the target (not split) datasets assembled contigs resulted from different trimming tools

S-table 22 Simulated datasets subjected pre-processed and assembled with Ray with different k-mer values. The treated reads were then aligned against the reference genomes.

S-table 23 Simulated datasets subjected pre-processed and assembled with SPAdes with different k-mer values. The treated reads were then aligned against the reference genomes.

S-table 24 Statistical analysis of the contigs obtained with the best k-mer size of SPAdes and Ray.

S-table 25 Simulated datasets phylum absolute abundance metric analysis for the three analysed tools (*MetaPhlAn v2.0, Metaxa2 (version 2.1.3)* and *Parallel-META 3.3.2)*

S-table 26 Simulated datasets genus absolute abundance metric analysis for the three analysed tools (*MetaPhlAn v2.0, Metaxa2 (version 2.1.3)* and *Parallel-META 3.3.2)*

S-tables 27-62 Simulated datasets phylum and genus relative abundance analysis for the three analysed tools (*MetaPhlAn v2.0, Metaxa2 (version 2.1.3)* and *Parallel-META 3.3.2).*

S-figures 1-20. COG, NOG, KO and SEED summary functions for A2, L1, L2, H1 and H2 datasets with and without coverage split.

S-table 63 COG results for each dataset acquired with MG-RAST and normalized with musicc software.

S-table 64 COG results for each unassembled and assembled dataset acquired with MGX for CETP datasets

S-table 65 COG results for each unassembled and assembled dataset acquired with MGX for petrochemical complex ETP datasets